Interaction partition. Structure of interacting coordinates for a multivariate stochastic process

Jesús E. García[1] and V. A. González-López[2]

[1] University of Campinas, Brazil
(e-mail: jg@ime.unicamp.br)
[2] University of Campinas, Brazil
(e-mail: veronica@ime.unicamp.br)

Abstract. In this paper we show a methodology to infer the dependence structure between the coordinates of a k-variate Markovian source. The methodology is based on the Bayesian information criterion (BIC). It is consistent in the sense that if the source is Markovian and the dataset is large enough, the exact dependence structure will be retrieved. Consider a set of k sources, for each realization at time t each source produces a letter in the alphabet $A = \{0, 1\}$. The sources interact between them depending on the past states of the set of k sources. In this paper it is proposed a methodology which obtains in a consistent way, a partition of the past such that two possible pasts are in the same part of the partition if, and only if, the set of interacting coordinates given any of this two pasts, is the same. We also obtain, for each possible past, the set of sources which interact between them.
Keywords: Multivariate Markov chain, Dependence structure, Partition Markov model.

1 Introduction

Parameter estimation in multiple interacting processes is a difficult task, even if the joint multivariate process is Markovian. Since the number of parameters grows exponentially not only with the dimension of the underlying alphabet but also with the length of the joint process memory. To mitigate this problem we will use the family of partition Markov models (PMM) (see [3] and [4]) which are generalizations of variable length Markov chain models (VLMC) (see [5], [7], [1] and [2]). The PMM family is more economic than the VLMC family, in relation to the number of parameters required to describe a given Markov process. Also the PMM family is especially convenient to develop an estimation strategy of the interaction structure, which will be exposed in this paper. The strategy is based on the Bayesian Information Criterion (BIC), which allows a consistent estimation of the interaction structure and also a consistent estimation of the PMM model.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

2 Stochastic equivalences

Let X_t be the state of the set of k sources at time t, $X_t = (X(1)_t, ..., X(k)_t)$, where $X(i)_t \in \{0,1\}$ and it is the state of the i-source at time t for $i = 1, ..., k$. $X_t \in A = \{0,1\}^k$. We will assume that X_t is an order M Markov chain, with $M < \infty$. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$, this means that a_m^n is the concatenation of elements from A.
Given the state space of strings of size M that is $\mathcal{S} = A^M$, for each $s \in \mathcal{S}$, $a \in A$, $b \in \{0,1\}$, we denote the conditional joint probability of the process by $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$ and the conditional marginal probability of the source $i \in \{1, ..., k\}$ by $P_i(b|s) = \text{Prob}(X(i)_t = b | X_{t-M}^{t-1} = s)$.

On the next definition it is introduced the notion of equivalence between strings from the state space $\mathcal{S}$, induced by the joint probability of the process. Also following that concept, will be introduced the notion of equivalence between strings of the state space based on the marginals probabilities of the stochastic process.

Definition 1. Let X_t be an order M Markov chain, with alphabet $A = \{0,1\}^k$ and state space $\mathcal{S} = A^M, M < \infty$.

i. For each $s, r \in S$, $s \sim r$ if $P(a|s) = P(a|r)$ $\forall a \in A$.
ii. For each $i \in \{1, ..., k\}$ and $s, r \in \mathcal{S}$, $s \sim_i r$ if $P_i(b|s) = P_i(b|r)$ $\forall b \in \{0,1\}$.

Remark 1 *For each* $s, r \in \mathcal{S}$, $s \sim r \ \Rightarrow \ s \sim_i r \ \forall i \in \{1, 2, ..., k\}$.

The next proposition shows that if all the sources are independent $\forall t$ then, definition 1 (i.) is true if, and only if definition 1 (ii.) is true for all the coordinates $i = 1, ..., k$.

Proposition 1. *Let* $X_t = (X(1)_t, ..., X(k)_t)$ *be an order* M *Markov chain, with alphabet* $A = \{0,1\}^k$ *and state space* $\mathcal{S} = A^M, M < \infty$. *If* $\forall t \ \ \{X(i)_t\}_{i=1}^k$ *are independent, for each* $s, r \in \mathcal{S}$,

$$s \sim r \iff s \sim_i r \ \ \forall i \in \{1, ..., k\}.$$

In the next section we will introduce the notion of partition $\mathcal{L}$ corresponding to $\sim$. Also we will give a summarized introduction to the Partition Markov Models (PMM), for more technical details about those models see [3] and [4].

2.1 Partition Markov Models

We note that the model introduced in this section, can be formulated not only for $A = \{0,1\}^k$, that is the case treated in this paper.

Definition 2. Let X_t be an order M Markov chain, with alphabet A and state space $\mathcal{S} = A^M, M < \infty$. We will say that X_t has partition $\mathcal{L}$ if this partition is the one defined by the equivalence relationship $\sim$ introduced by definition 1 (i.).

We observe that the set of parameters for a Markov chain over the alphabet A with partition $\mathcal{L}$ is given by the set of conditional probabilities

$$\{P(a|L) : a \in A, L \in \mathcal{L}\}.$$

where $P(a|L) = P(a|s)$, for any $s \in L$. If we know the equivalence relationship for a given Markov chain, then we need $(|A| - 1)$ transition probabilities for each part L to specify the model. And the total number of parameters will be $|\mathcal{L}|(|A| - 1)$.
To choose a model (in this case, a partition $\mathcal{L}$) in a consistent way (see [3] for a more comprehensive explanation) we can use a distance in the state space $\mathcal{S}$ formulated from a sample x_1^n of the process X_t.

2.2 Interaction structure on a multivariate PMM

[6] defines the dependence structure between coordinates for VLMC models and shows that this dependence structure can be estimated using BIC criterion in the following way. First is fitted a VLMC and then, for each context, the BIC criterion is used on the transition probabilities corresponding to that context to find a partition of the coordinates on dependent sets. The results in [6] are valid for any family of Markovian models as they only depend on the individual transition probabilities and not on the model structure.

For simplicity in order to introduce the notion of interaction (or not) we will assume that a PMM has been already obtained and the partition $\mathcal{L}$ is the partition corresponding to $\sim$. Our goal is to obtain for each part of the partition of the state space, a partition of the set of coordinates of the multivariate process. This last partition will discriminate the set of coordinates in independent sets. After that, we will put together all the parts of $\mathcal{L}$ with the same partition in the space of coordinates.
Let (X_t) be a Markov chain on $A = \{0, 1\}^k$, with partition $\mathcal{L}$, of the state space $\mathcal{S}$. For a collection of coordinates $u = \{u_1, ...u_l\} \subset \{1, 2, ..., k\}$ and $a = (a_1, ..., a_k) \in A$, define, $a^u = (a_{u_1}, ..., a_{u_l})$ that is a vector composed only by the u coordinates of a. For each part $L \in \mathcal{L}$ define the transition probability from that part to a vector a^u, $P(a^u|L) = \text{Prob}(X_t^u = a^u | X_{t-M}^{t-1} = s)\ \forall s \in L$. The previous definition is allowed because L is a part of the partition $\mathcal{L}$ following definition 1 (i.).
In general, for $A = \{0, 1\}^k$, given $L \in \mathcal{L}$ and a partition of $\{1, 2, ..., k\}$, $\mathcal{I}_L$ of independent coordinates, we have that

$$P(a|L) = \prod_{C \in \mathcal{I}_L} P(a^C|L)\ \forall a \in A,$$

while, the number of parameters needed for the part L will be $\sum_{C \in \mathcal{I}_L} (2^{|C|} - 1)$.

Definition 3. Let X_t be a discrete time, order M Markov chain on a finite alphabet A with $M < \infty$ and partition of the state space $\mathcal{L}$. For each $L \in \mathcal{L}$ define $\mathcal{D}_L$ as the largest partition of $\{1, 2, ..., k\}$ such that $P(a|L) = \prod_{C \in \mathcal{D}_L} P(a^C|L)\ \forall a \in A$. We will say that $\mathcal{D}_\mathcal{L} = \{\mathcal{D}_L\}_{L \in \mathcal{L}}$ is the structure of interaction for the process.

Consider now $I = \cup_{\{L \in \mathcal{L}\}} \mathcal{D}_L$, that will contain each kind of partition of the coordinates appearing in $\mathcal{D}_\mathcal{L}$. Each element of I corresponds to a particular type of dependence identified in the model structure. For each $P \in I$, let be

$$M_P = \cup_{\{L \in \mathcal{L}: \mathcal{D}_L = P\}} L \quad \text{and} \quad \mathcal{M} = \{M_P\}_{P \in I}.$$

$\mathcal{M}$ is a partition of $\mathcal{S}$ such that two sequences $s, r \in \mathcal{S}$ are in the same part of $\mathcal{M}$ if and only if, given each of this two strings, the set of sources interacting is the same. The partition $\mathcal{M}$ tell us for each possible past, which of the different sources interact.

Example 1. Set $A = \{0,1\}^2$ and define the state space as $\mathcal{S} = A^2$. Consider also the following set of conditional probabilities,

s	$P_1(0\|s)$	$P_2(0\|s)$	$P((0,0)\|s)$	$\sim$	$\mathcal{I}$	$\mathcal{F}$
$(0,0),(0,0)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,0),(0,1)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,1),(0,0)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,1),(0,1)$	0.1	0.1	0.01	L_1	M_1	F_1
$(0,0),(1,0)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,0),(1,1)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,1),(1,0)$	0.1	0.1	0.02	L_2	M_2	F_2
$(0,1),(1,1)$	0.1	0.1	0.02	L_2	M_2	F_1
$(1,0),(0,0)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,0),(0,1)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,1),(0,0)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,1),(0,1)$	0.2	0.2	0.04	L_3	M_1	F_1
$(1,0),(1,0)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,0),(1,1)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,1),(1,0)$	0.2	0.2	0.02	L_4	M_2	F_3
$(1,1),(1,1)$	0.2	0.2	0.02	L_4	M_2	F_3

We observe, that for any $s, r \in \mathcal{S}$, $s \sim r$ if, and only if $P_i(0|s) = P_i(0|r)$, $i = 1,2$ and $P((0,0)|s) = P((0,0)|r)$. We will remark two situations, (i) when the sources 1 and 2 interact and (ii) when the sources 1 and 2 are independent.

(i) iff $P_1(0|s) \neq \frac{P((0,0)|s)}{P_2(0|s)}$ iff $1 \neq \frac{P((0,0)|s)}{P_1(0|s)P_2(0|s)}$.
(ii) iff $P((0,0)|s) = P_1(0|s)P_2(0|s)$.

In this example, each marginal equivalence $\sim_i$ has two parts, this means that each marginal state space will be composed by two parts. In addition, the partition $\mathcal{L}$ corresponding to $\sim$ has four parts, the fifth column of the table indicates the part to which each $s \in \mathcal{S}$ (listed in the first column) belongs.
In the parts L_2 and L_4 the two sources are interacting while in the parts L_1 and L_3 the sources are independent. The partition (of coordinates) $\mathcal{M} = \{L_1 \cup L_3, L_2 \cup L_4\}$ indicates when the two sources interact. Note that the partition $\mathcal{M}$ indicates if the sources interact but not how. By complement, the partition $\mathcal{F} = \{L_1 \cup L_3, L_2, L_4\}$ indicates exactly if the two sources interact and it indicates also which kind of interaction occurs.

3 Estimation

In this section we will introduce the methodology of estimation, which consists of distances and algorithms that take advantage of the BIC criterion to produce consistent estimates of the structure of interaction as well as of the parameters of a PMM model.

3.1 Partition Markov Model Estimation

In this section it is assigned a distance to the state space $\mathcal{S}$. It will allow the estimation of the true partition of a PMM model. Based on this distance, it is also proposed the construction of an algorithm that allows to obtain an estimator of the partition which converges almost surely eventually, to the true partition of the state space $\mathcal{S}$.
Consider a sample x_1^n of the process X_t, $a \in A$ and $s \in \mathcal{S}$. We will denote by $N_n(s)$ the number of occurrences of s in the sample and by $N_n(s,a)$ the number of occurrences of s followed by a in the sample,

$$N_n(s) = \sum_{m=M+1}^{n+1} 1_{\{x_{m-M}^{m-1}=s\}}, \quad N_n(s,a) = \sum_{m=M+1}^{n} 1_{\{x_{m-M}^{m-1}=s, x_m=a\}}.$$

The equivalence between two strings coming from the state space, s and r will depend on the sample and denoted by $s \sim_n r$

$$s \sim_n r \iff \frac{N_n(s,a)}{N_n(s)} = \frac{N_n(r,a)}{N_n(r)} \quad \forall a \in A.$$

Definition 4. Let x_1^n be a sample of X_t, for any $s, r \in \mathcal{S}$,

$$\begin{aligned} d_n(s,r) = \frac{2}{(|A|-1)\ln(n)} \sum_{a \in A} \Bigg\{ & N_n(s,a) \ln\left(\frac{N_n(s,a)}{N_n(s)}\right) \\ & N_n(r,a) \ln\left(\frac{N_n(r,a)}{N_n(r)}\right) \\ & - \left(N_n(\{s,r\},a) \ln\left(\frac{N_n(\{s,r\},a)}{N_n(s)+N_n(r)}\right)\right) \Bigg\}, \end{aligned}$$

with $N_n(\{s,r\},a) = N_n(s,a) + N_n(r,a)$.

We note that d_n can be generalized to subsets of the state space $\mathcal{S}$ and it has the property of being equivalent to the BIC Criterion to decide if $s \sim r$ for any $s, r \in \mathcal{S}$ (for details, see [3]).
From the next result will be possible to use the distance d_n as a consistent criterion to allocate the strings of the state space in parts that compound the partition.

Theorem 1 *Let X_t be a discrete time, order M Markov chain on a finite alphabet A with $M < \infty$, with partition $\mathcal{L}$. Let x_1^n be a sample of the process, then for n large enough, for each $s, r \in \mathcal{S}$, $d_n(r,s) < 1$ iff s and r belong to the same part of the partition $\mathcal{L}$.*

Algorithm 1 *(Partition selection algorithm)*
Input*:* $d_n(s,r)\ \forall s,r \in \mathcal{S}$; **Output***:* $\hat{\mathcal{L}}_n$.
$B = \mathcal{S}$
$\hat{\mathcal{L}}_n = \emptyset$
while $B \neq \emptyset$

select $s \in B$
define $L_s = \{s\}$
$B = B \setminus \{s\}$
for each $r \in B, r \neq s$
if $d_n(s,r) < 1$
$L_s = L_s \cup \{r\}$
$B = B \setminus \{r\}$
$\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n \cup \{L_s\}$

Return*:* $\hat{\mathcal{L}}_n = \{\hat{L}_i\}_i$

That means that if the source is Markovian, for n large enough, the algorithm returns the true partition for the source.

Corollary 1. *Under the assumptions of theorem 1,* $\hat{\mathcal{L}}_n$*, given by the algorithm 1 converges almost surely eventually to* $\mathcal{L}$*, where* $\mathcal{L}$ *is the partition of* $\mathcal{S}$ *defined by the equivalence given by definition 1 (i.).*

3.2 Dependence Structure Estimation

In this section we present the maximum likelihood expression that allows the estimation of the underlying dependence structure $D_{\mathcal{L}}$, introduced by definition 3. So, based on an estimate of the partition of the state space $\mathcal{S}$, the BIC criterion enables to obtain an estimated structure that converges eventually almost surely to the true dependence structure.
To estimate the probabilities, we introduce, for $s \in \mathcal{S}$ the number of occurrences of the string s followed by a vector that has the coordinates listed by u equal to a^u $N_n(s,a^u) = \left|\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t^u = a^u\}\right|$. In addition for each part $L \in \mathcal{L}$ $N_n^{\mathcal{L}}(L,a^u) = \sum_{s\in L} N_n(s,a^u)$ and $N_n^{\mathcal{L}}(L) = \sum_{s\in L} N_n(s)$.
In the next paragraph we emphasize the count of two quantities, the number of occurrences of s followed by any vector with i-th coordinate equal to b, and the number of occurrences of s followed by any vector with coordinates listed in u equal to c.

$$N_n(s, a^{\{i\}} = b) = \sum_{t=M}^{n} 1_{\{x_{t-M}^{t-1} = s, x_t^i = b\}},$$

with $i \in \{1,\ldots,k\},\ b \in \{0,1\}$.

$$N_n(s, a^u = c) = \sum_{t=M}^{n} 1_{\{x_{t-M}^{t-1} = s, x_t^{u_j} = c_j,\ 1\leq j\leq l\}},$$

with $c \in \{0,1\}^l$.
Given a sample of the process x_1^n, if we write $P(x_1^n) = \text{Prob}(X_1^n = x_1^n)$, we obtain under the assumption of a hypothetical partition $\mathcal{L}$ of $\mathcal{S}$,

$$P(x_1^n) = P(x_1^M) \prod_{L \in \mathcal{L}, a \in A} \prod_{C \in \mathcal{D}_L} P(a^C|L)^{N_n^{\mathcal{L}}(L,a)}.$$

The maxima for $\prod_{L \in \mathcal{L}, a \in A} \prod_{C \in \mathcal{D}_L} P(a^C|L)^{N_n^{\mathcal{L}}(L,a)}$ is

$$\text{ML}(\mathcal{L}, \mathcal{D}_{\mathcal{L}}, x_1^n) = \prod_{L \in \mathcal{L}, a \in A} \prod_{C \in \mathcal{D}_L} \left(\frac{N_n^{\mathcal{L}}(L, a^C)}{N_n^{\mathcal{L}}(L)} \right)^{N_n^{\mathcal{L}}(L,a)},$$

and the BIC expression under this formulation will be

$$BIC(\mathcal{L}, \mathcal{D}_{\mathcal{L}}, x_1^n) = \ln\left(\text{ML}(\mathcal{L}, \mathcal{D}_{\mathcal{L}}, x_1^n)\right) - \sum_{L \in \mathcal{L}} \sum_{C \in \mathcal{D}_L} (2^{|C|} - 1) \frac{\ln(n)}{2}.$$

For a Markovian source the BIC model selection methodology is consistent as we show in the next result.

Theorem 2 *Let X_t be under the assumptions of theorem 1, with partition of the state space $\mathcal{L}$ and structure of conditional dependence $\mathcal{D}_{\mathcal{L}}$. Define,*

$$\mathcal{D}_{\hat{\mathcal{L}}_n} = \arg\max_{\mathcal{D} \in \mathbf{D}} \{BIC(\hat{\mathcal{L}}_n, \mathcal{D}, x_1^n)\},$$

Where $\mathbf{D}$ is the set of all possible structures of dependences for A and $\hat{\mathcal{L}}_n$, $\hat{\mathcal{L}}_n$ obtained using algorithm 1, then, eventually almost surely as $n \to \infty$,

$$\mathcal{D}_{\mathcal{L}} = \mathcal{D}_{\hat{\mathcal{L}}_n}.$$

3.3 Simultaneous estimation of the partition and the interaction structure

In this section, we will simultaneously estimate the partition of PMM models and the interaction structure using the BIC criterion. A consistent strategy of estimation will be introduced.

We will introduce the following measure of dependence between pairs of coordinates conditioned to a past $s \in \mathcal{S}$,

Definition 5. Let x_1^n be a sample of X_t, for any $s \in \mathcal{S}$, and $i, j \in \{1, 2, ..., k\}$

$$\begin{aligned} d_s^n(i,j) = \frac{2}{\ln(n)} \sum_{b \in \{0,1\}} \Bigg\{ & N_n(s, a^{\{i\}} = b) \ln\left(\frac{N_n(s, a^{\{i\}} = b)}{N_n(s)} \right) \\ & + N_n(s, a^{\{j\}} = b) \ln\left(\frac{N_n(s, a^{\{j\}} = b)}{N_n(s)} \right) \Bigg\} \\ & - \sum_{c \in \{0,1\}^2} \left\{ N_n(s, a^{\{i,j\}} = c) \ln\left(\frac{N_n(s, a^{\{i,j\}} = c)}{N_n(s)} \right) \right\}. \end{aligned}$$

The next theorem shows that this distance between coordinates can be used to find the structure of interactions for a given past $s \in \mathcal{S}$ in a consistent way.

Theorem 3 *Let X_t be under the assumptions of theorem 1, with state space $\mathcal{S}$. For n large enough, for $s \in \mathcal{S}$ and $i, j \in \{1, ...k\}$, $d_s^n(i,j) < 1$ iff i and j are dependent.*

Remark 2 *The concept d_s^n can be extended to parts of the partition defining a PMM, replacing s by L.*

Using the distances in definition 4 and definition 5, we can define the following algorithm to estimate $\mathcal{D}_{\mathcal{L}}$.

Algorithm 2 *(Coordinate partition selection algorithm)*
Input*: for a fixed $s \in \mathcal{S}$, $d_s^n(i,j) \forall 1 \leq i, j \leq k$.*
Output*: $\hat{D}_s^n$;*
$B = \{1, 2, ..., k\}$
$\hat{D}_s^n = \emptyset$
while $B \neq \emptyset$

select $i \in B$
define $D_i = \{i\}$
$B = B \setminus \{i\}$
for each $j \in B, j \neq i$
if $d_s^n(i,j) < 1$
$D_i = D_i \cup \{j\}$
$B = B \setminus \{j\}$
$\hat{D}_s^n = \hat{D}_s^n \cup \{D_i\}$

Return*: $\hat{D}_s^n$*

We will show later, that the distance introduced in the next paragraph can be used to find the PMM and also the dependence structure.

Definition 6. Let x_1^n be a sample of X_t, for any $s, r \in \mathcal{S}$,

$$
\begin{aligned}
& d'_n(s,r) = \\
&= \frac{2}{M(s,r)\ln(n)} \sum_{a \in A} \left\{ \sum_{C \in \hat{D}_s^n} N_n(s,a) \ln\left(\frac{N_n(s,a^C)}{N_n(s)}\right) \right. \\
&+ \sum_{C \in \hat{D}_r^n} N_n(r,a) \ln\left(\frac{N_n(r,a^C)}{N_n(r)}\right) \\
&\left. - \sum_{C \in \hat{D}_{\{s,r\}}^n} (N_n(\{s,r\},a)) \ln\left(\frac{N_n(s,a^C) + N_n(r,a^C)}{N_n(s) + N_n(r)}\right) \right\},
\end{aligned}
$$

with
$M(s,r) = \sum_{C \in \hat{D}_s^n} (2^{|C|} - 1) + \sum_{C \in \hat{D}_r^n} (2^{|C|} - 1) - \sum_{C \in \hat{D}_{\{r,s\}}^n} (2^{|C|} - 1)$.

Following the steps pointed here, can be estimated the dependence structure for each element of the estimated partition.

Step 1. Given the state space $\mathcal{S}$, apply the algorithm 2, obtaining $\hat{D}_s^n,\ \forall s \in \mathcal{S}$;
Step 2. Apply algorithm 1, replacing d_n (def. 4) by d'_n (def. 6), obtaining $\hat{\mathcal{L}}_n = \{\hat{L}_i\}_{i\geq 1}$;
Step 3. Apply algorithm 2, replacing d_s^n (def. 5) by $d_{\hat{L}}^n$ (Remark 2), obtaining $\hat{D}_{\hat{L}}^n,\ \forall \hat{L} \in \hat{\mathcal{L}}_n$.

Once $\mathcal{D}_{\mathcal{L}}$ is estimated, we can identify the specific kind of interaction using standard statistical methods on each interacting set. For example, fixed a part $L \in \mathcal{L}$ if $\mathcal{D}_L = \{C_1, ..., C_{m_L}\}$ (composed by independent sets of coordinates) we only need to work with the marginals $X^{C_i},\ i \in \{1, ..., m_L\}$ to determine the kind of dependence between the coordinates of elements into A.
In addition, recalling $\mathcal{M} = \{M_P\}_{P\in I}$ with $I = \cup_{\{L\in\mathcal{L}\}}\mathcal{D}_L$ and $M_P = \cup_{\{L\in\mathcal{L}:\mathcal{D}_L=P\}}L$, for each part M_P, each marginal $X^C,\ C \in P$, being independent from the others coordinates can be analyzed by itself, which in general requires less data than the simultaneous analysis of all k coordinates.

4 Conclusion

The PMM models are flexible structures that permit the incorporation of concepts of dependence or interaction, as illustrated in this paper. The operation of dividing the state space in parts of a partition $\mathcal{L}$, governed by the equivalence $\sim$ allows that the characterization of the interaction can be made in each part of the partition $\mathcal{L}$, this is, locally on the state space $\mathcal{S}$. Theorem 2 shows how to estimate the structure of interaction, using a consistent estimate of the partition of the state space. Also, in this paper we simultaneously estimate the partition of the space state $\mathcal{L}$ of a PMM model and also the interaction structure $\mathcal{D}_{\mathcal{L}}$, through the Bayesian information criterion. In this way, a consistent strategy of estimation was introduced.

5 ACKNOWLEDGMENTS

The authors gratefully acknowledge the support for this research provided by USP project "Mathematics, computation, language and the brain" and FAPESP's projects (a)"Portuguese in time and space: linguistic contact, grammars in competition and parametric change" 2012/06078-9 and (b)"Research, Innovation and Dissemination Center for Neuromathematics" 2013/ 07699-0, (S. Paulo Research Foundation).

References

1. Csiszár, I. and Talata, Z., "Context tree estimation for not necessarily finite memory processes, via BIC and MDL", *IEEE Trans. Inform. Theory* **52**, 1007-1016 (2006)

2. Galves, A., Galves, C., Garcia, J. E., Garcia, N. L. and Leonardi, F., "Context tree selection and linguistic rhythm retrieval from written texts", *Annals of Applied Statistics*, **6** 1, 186-209 (2012)
3. Garcia, J. and Gonzalez-Lopez, V. A., "Minimal Markov Models". arXiv preprint *arXiv:1002.0729* (2010)
4. Garcia, J. E., Gonzalez-Lopez, V. A. and Viola, M. L. L., "Robust model selection and the statistical classification of languages", *AIP Conference Proceedings*, vol. 1490, p.160 (2012)
5. Rissanen J., "A universal data compression system", *IEEE Trans. Inform. Theory* **29**(5), 656-664 (1983)
6. Viola, M., L. Tópicos em seleção de modelos markovianos. PhD Tesis. `http://www.bibliotecadigital.unicamp.br/document/?code=000844289` (2011)
7. Weinberger, M., Rissanen, J. and Feder, M. "A universal finite memory source", IEEE Trans. Inform. Theory **41**(3) 643-652 (1995)

Detecting regime changes in Markov models

Jesús E. García[1] and V. A. González-López[2]

[1] University of Campinas, Brazil
(e-mail: jg@ime.unicamp.br)
[2] University of Campinas, Brazil
(e-mail: veronica@ime.unicamp.br)

Abstract. Let C be a data collection, indexed by time. $C = \{D(t_1), ..., D(t_n)\}$, where $D(t_i)$ was collected at time $t_i, t_i \leq t_j$ if $i \leq j$. Also, each $D(t_i)$ follows a Markovian model with finite alphabet A, denoted by $M(t_i)$. We device a consistent procedure to detect changes in the model at time t_{i_0} that allows to decide if $D(t_{i_0})$ and $D(t_{i_0-1})$ are coming from the same Markovian source. The procedure is based on the equivalence relationship introduced by the Minimal Markov Models (see [7]), that allows to associate to each Markovian model a minimal number of parameters enough to describe a Markovian source. The model's estimation can be consistently performed through the Bayesian Information Criterion, see for details [1] and [7], also for related topics about those models, see [8], [10] and [5]. Under the possibility of regime change, we can have situations in which $D(t_1), ..., D(t_{i_0-1})$ are coming from a Markovian model, $M(t_{i_0-1})$ different to the Markovian model $M(t_{i_0})$ appropriated for $D(t_{i_0})$. We apply the procedure to detect prosodic changes from classical to modern European Portuguese (see [2], [3], [4]). Taking in consideration that rhythm is a consequence of several characteristics, like number of syllables in the words, position in the word of the stressed syllable, simple and complex syllabic structure, etc., it is possible to look for temporal changes in the rhythm, using written texts. In this context, each $D(t_i)$ is a written text in European Portuguese and t_i is the author's date of birth from 16th century to the 19th century. In this analysis we detect two main change points, the first one at the turn of the 16th century to the 17th century. The second one, in the second half of the 18th century that spreads to the end of the century. Our findings complement the results attained in [3], which study the changes of the European Portuguese in the same period of time, through the analysis of clitic placement. The processing and types of statistical models are completely defined by the differentiated nature of the data. For example, for acoustic signal processing see [6] and for recent research about the statistical modeling see [8], [9] and [10].
Keywords: Minimal Markov models, Model selection, Bayesian information criterion, Historical linguistics.

1 Introduction

Our goal in practical terms is to explore whether using Markov structures can be extracted rhythmic properties of texts written in European Portuguese. We also want to demonstrate that such models may be useful for the study of the

3^{rd} *SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal*
C. H. Skiadas (Ed)

prosodic changes in texts sorted by chronological time.
In [2] it is proved that in fact the European Portuguese has undergone a significant alteration, perceived from the 16th century to the 17th. This finding is consistent with the conjecture that the Portuguese is losing some of its features of "romance language" with the passing of the centuries. In [2] significant changes are verified on two phonological features, the size of the words and the position of the stress. Thus, from the 16th century to the centuries 17th, 18th and 19th, [2] shows: (a) a pattern of increase in the proportion of monosyllables (in the universe defined by words of at most two syllables) and (b) a pattern of increase in the proportion of words with stress on the last syllable (in the universe defined by words with atress positioned in the penultimate or in the last syllable).
This paper aims to investigate the problem by incorporating the Markovian structure inherent to written texts. Supported by a model of this nature we can leverage all the available information about the data. Given that the model used in [2], based on the beta-binomial distribution, requires preprocessing of the data to obtain the independence between the realizations (or words), interfering with the maintenance of the rhythm and at the same time produces a reduction of the sample size of the written texts. So, (i) the number of syllables in each word and (ii) the placement of stress, will be investigated under a richest model that allows to incorporate a dependence structure between words through each text and enables consider jointly, the features (i) and (ii). To achieve a more comprehensive view of linguistic phenomena studied at present, it should be noted that linguistic structures can be studied in their formats "spoken language" and "written language". The processing and types of statistical models are completely defined by the differentiated nature of the data. For example, for acoustic signal processing see [6] and for recent research about the statistical modeling see [8], [9] and [10].
Using the linguistic problem as a motivating basis, we introduce in this paper a consistent method for to find changes of regime in Markov processes. The method takes advantage of the conception of minimal Markov models [7] and was formulated using the Bayesian Information Criterion (BIC). The latter allows to define a rule for deciding whether or not there is a change of regime in Markov process. Finally, we will apply this procedure in the linguistic problem.

2 Historical data

Tycho Brahe corpus is an annotated historical corpus, freely accessible at [4] (`http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html`).
This corpus uses the chronological criterion of the author's birthdate to assign a time for written text. The subset of historical written texts included in this study, listed in Table 1 is composed by 17 texts from 15 authors, coming from four genres. In Table 1 we report also the number of orthographic words (ow) by text. The data collection $C = \{D_{t_1}, ..., D_{t_n}\}$ is now given by the written texts listed in Table 1.

D_t	Gândavo	Pinto	Sousa	Brandão	Vieira	Vieira
t	1502	1510	1556	1584	1608	1608
Type	N	N	N	N	L	S
ow	22850	39941	50218	43192	47888	49275
D_t	Chagas	Bernardes	Oliveira	Aires	Costa	Alorna
t	1631	1644	1702	1705	1714	1750
Type	P	P	L	P	L	L
ow	48670	49479	16629	56055	24538	43318
D_t	Garrett	Garrett	Fronteira	Camilo	Ortigão	
t	1799	1799	1802	1826	1836	
Type	L	N	N	N	L	
ow	30070	45800	54826	20142	27420	

Table 1. Subset of Tycho Brahe corpus used in this study, coming from four genres: narrative (N), letters (L), philosophical (P) and sermons (S).

2.1 Encoding texts

Each written text was processed with a slightly modified version of the perl-code "silaba" that can be freely downloaded for academic purposes at `www.ime.usp.br/ tycho/prosody/vlmc/tools/sil4.pl` . The software was used to extract two components of each orthographic word, denoted by (i, j), where i is the total number of syllables that make up the word, $i = 1, 2, ..., 8$ and j indicates the syllable in which is registered the stress in the word, $j = 0, 1, 2, .., 8$. Where, $j = 0$ means no stress in the word and this just happens in orthographic words with one syllable. The period (final of sentence) was codified as $(0, 0)$.
The alphabet was defined as exposed in Table 2. Note that the set of words represented by $(i, 0), i \geq 2$ corresponds to the empty set.

orthographic word	element alphabet
$(0, 0)$	a
$(1, 0)$	b
$(1, 1)$	c
$(2, 1)$	d
$(2, 2)$	e
$(i, 1)$, $i \geq 3$	f
$(i, 2)$, $i \geq 3$	g
(i, j), $i, j \geq 3$	h

Table 2. Definition of the alphabet A.

3 The Markovian model

The minimal Markov models applied in this paper, were introduced in [7]. Those models are generalizations of Variable Length Markov Chains models, used to discover the differences between branches of the Portuguese in [5].
Let (X_t) be a discrete time (order $M < \infty$) Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$.
For each $a \in A$ and $s \in \mathcal{S}$, $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$.
Let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$, for $a \in A$, $L \in \mathcal{L}$, $P(L, a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a)$, $P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s)$ and $P(a|L) = \frac{P(L,a)}{P(L)}$ with $P(L) > 0$.

Definition 1. Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. We will say that $s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)$ $\forall a \in A$. For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

The previous definition allows to define a Markov chain with a "minimal partition", that is the one which respects the equivalence relationship.

Definition 2. let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$, if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by definition 1.

In a given sample x_1^n, coming from the stochastic process, we denote the number of occurrences of elements into L followed by a for,

$$N_n^{\mathcal{L}}(L, a) = \sum_{s \in L} N_n(s, a), \ L \in \mathcal{L},$$

where the number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $N_n(s, a)$. The accumulated number of $N_n(s)$ for s in L is denoted by,

$$N_n^{\mathcal{L}}(L) = \sum_{s \in L} N_n(s), \ L \in \mathcal{L}.$$

The model, in this context given by the "minimal partition $\mathcal{L}$", can be selected consistently, using the Bayesian Information Criterion. This means, the best partition is the one that maximizes

$$\text{BIC}(x_1^n, \mathcal{L}) = \sum_{a \in A, L \in \mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln \Big(\frac{N_n^{\mathcal{L}}(L, a)}{N_n^{\mathcal{L}}(L)} \Big) - \frac{(|A| - 1)|\mathcal{L}|}{2} \ln(n),$$

over the space of partitions.

3.1 Criterion of remoteness between processes

The BIC allows to compare datasets as we will show in the next result. If two variables X and Y have the same distribution we will assume the next notation $X =^d Y$. Also, if $(x)_{i=1}^n$ and $(y)_{i=1}^m$ are samples of X and Y respectively, we will denote by $(x)_{i=1}^n \perp (y)_{i=1}^m$ the independence between the samples.

Theorem 1. *Given the stochastic process X_{t_i} of order $M < \infty$ with sample $(x_{t_i})_1^{n_i}$ of size $n_i, i = 1, 2$, such that $(x_{t_1})_1^{n_1} \perp (x_{t_2})_1^{n_2}$. $X_{t_1} \neq^d X_{t_2}$ if, and only if*

$$BIC\Big((x_{t_1})_1^{n_1}, (x_{t_2})_1^{n_2}, \mathcal{L}\Big) < \sum_{k=1,2} BIC((x_{t_k})_1^{n_k}, \mathcal{L}^k)$$

with

$$BIC\Big((x_{t_1})_1^{n_1}, (x_{t_2})_1^{n_2}, \mathcal{L}\Big) = \sum_{a\in A, L\in\mathcal{L}} N^{\mathcal{L}}_{n_1+n_2}(L,a) \ln\Big(\frac{N^{\mathcal{L}}_{n_1+n_2}(L,a)}{N^{\mathcal{L}}_{n_1+n_2}(L)}\Big) - \frac{(|A|-1)}{2}|\mathcal{L}| \ln(n_1+n_2).$$

Corollary 1. *Under the assumptions of Theorem 1, $d_{1,2} > 1$, with*

$$d_{1,2} = \frac{2\sum_{a\in A} B(\mathcal{L}^1, n_1, a) + B(\mathcal{L}^2, n_2, a) - B(\mathcal{L}, n_1+n_2, a)}{(|A|-1)\left\{|\mathcal{L}^1| \ln(n_1) + |\mathcal{L}^2| \ln(n_2) - |\mathcal{L}| \ln(n_1+n_2)\right\}}$$

and $B(\mathcal{L}, n, a) = \sum_{L\in\mathcal{L}} N_n^{\mathcal{L}}(L,a) \ln\Big(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\Big)$.

Given the dataset D_{t_i} consider the stochastic process X_{t_i} of order $M < \infty$ generator of D_{t_i}, with sample $(x_{t_i})_1^{n_i}$ of size n_i.
Following the codification given by Table 2, each sample will be composed by the concatenation of symbols from $A = \{a, b, c, d, e, f, g, h\}$. Based on previous works, that investigate similar data (see, for example [5]) the value of M considered here was 4.
Assuming that the data collection is made up of independent texts (which is the case treated here, as each text is a complete work in itself), under the assumption:

$$X_{t_i} =^d X_{t_j} \text{ and } (x_{t_i})_1^{n_i} \perp (x_{t_j})_1^{n_j}, i \neq j$$

$$\text{BIC}\Big((x_{t_i})_1^{n_i}, (x_{t_j})_1^{n_j}, \mathcal{L}\Big) > \sum_{k=i,j} \text{BIC}((x_{t_k})_1^{n_k}, \mathcal{L}^k)$$

and

$$d_{i,j} < 1.$$

That means that both: D_{t_i} and D_{t_j} come from the same model, given by the minimal partition $\mathcal{L}$. In another case D_{t_i} and D_{t_j} come from different models, $\mathcal{L}^i$ and $\mathcal{L}^j$ respectively.
In the next section we use the values of $d_{i,j}$ to measure the distance between the models associated with written texts. Thus, texts that are identified with the same model show no change points in the timeline. When $d_{i,j}$ exceeds the value 1, a change point is identified.

4 Results and Conclusions

In the Figure 1, each horizontal line represents the text written by a particular author. On the line of each text is shown the value of d computed for two consecutive texts in time. Thus, for example, the text titled as "Gândavo (1502)" was compared with the author's text immediately following, that is the text titled as "Pinto (1510)" and the value of d displayed in Gândavo (1502)'s line. Now, when the line shows two points (two values of d), such as the case of the Brandão (1584)'s line, is because there are two texts in the sample of the same year. For instance, those texts are from Vieira (1608):(a) letters and (b) sermons.

In this analysis we detect two main change points, the first one at the turn of

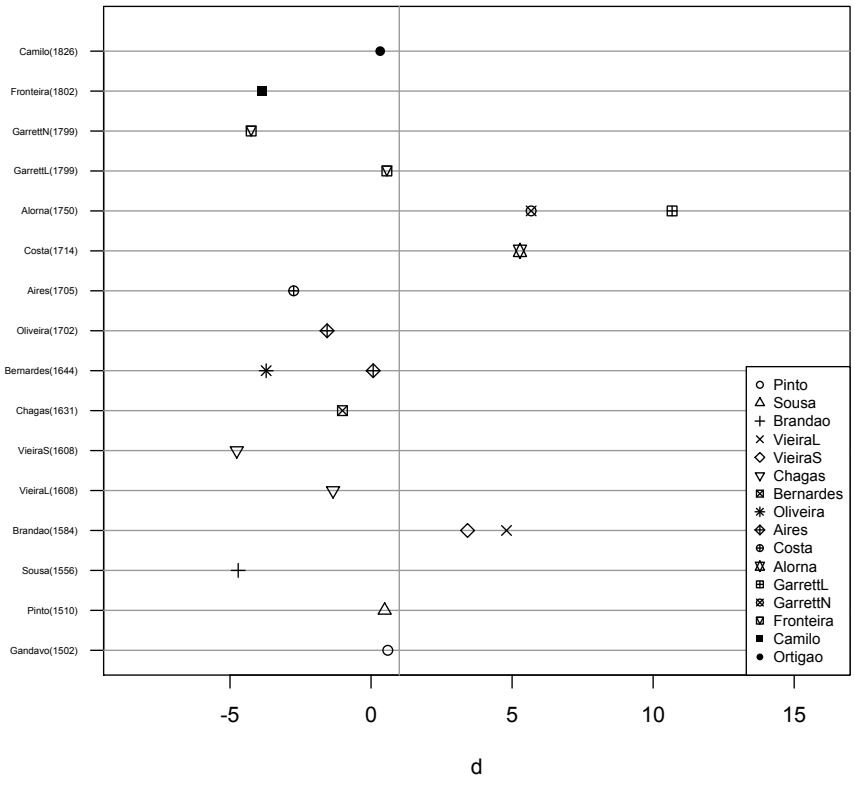


Fig. 1. Each horizontal line represents a written text from European Portuguese, those were ordered by time from down to top. The vertical line represents $d = 1$, greater values of d indicates the presence of a change point.

the 16th century to the 17th century. The second one, in the second half of the 18th century that spreads to the end of the century. Our findings complement the results attained in [3], which study the changes of the European Portuguese in the same period of time, through the analysis of clitic placement.

5 ACKNOWLEDGMENTS

The authors gratefully acknowledge the support for this research provided by (a) FAEPEX-Unicamp, (b) USP's project "Mathematics, computation, language and the brain" and (c) FAPESP's project:"Portuguese in time and space: linguistic contact, grammars in competition and parametric change," grant 2012/06078-9.

References

1. Csiszár, I. and Talata, Z., "Context tree estimation for not necessarily finite memory processes, via BIC and MDL", *IEEE Trans. Inform. Theory*, 52, 1007-1016 (2006).
2. Frota, S., Galves, C., Vigário, M., Gonzalez-Lopez, V. and Abaurre, B., "The phonology of rhythm from Classical to Modern Portuguese", *Journal of Historical Linguistics*, 2(2), 173-207 (2012).
3. Galves, C., Britto, H. and de Sousa, M. C. P., "The Change in Clitic Placement from Classical to Modern European Portuguese", *Journal of Portuguese Linguistics*, 4(1), 39-67 (2005).
4. Galves, C. and Faria, P., *Tycho Brahe Parsed Corpus of Historical Portuguese.* `http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html` (2010)
5. Galves, Antonio, Charlotte Galves, Jesús E. García, Nancy L. Garcia, and Florencia Leonardi,"Context tree selection and linguistic rhythm retrieval from written texts", *The Annals of Applied Statistics*, 6(1), 186-209 (2012).
6. Garcia, Jesus, Ulrike Gut, and Antonio Galves, "Vocale-a semi-automatic annotation tool for prosodic research". In *Speech Prosody 2002, International Conference.* (2002)
7. Garcia, J. and Veronica A. Gonzalez-Lopez, "Minimal markov models", arXiv preprint arXiv:1002.0729 (2010)
8. García, Jesús E., González-López, V. A., Viola, M. L. L., "Robust model selection and the statistical classification of languages". In *AIP Conference Proceedings* vol. 1490, p.160 (2012).
9. García, Jesús E., V. A. González-López, and R. B. Nelsen, "A new index to measure positive dependence in trivariate distributions", *Journal of Multivariate Analysis*, 115, 481-495 (2013) (doi 10.1016/j.jmva.2012.11.007).
10. García, Jesús E., González-López, V. A., Viola, M. L. L., "Robust Model Selection for Stochastic Processes", *Communications in Statistics - Theory and Methods*, 43, 2516-2526 (2014) (doi 10.1080/03610926.2013.851220).

Graphical Probability Modelling of Dynamic Processes

Ali S. Gargoum

Department of Statistics, UAE University
Al Ain,United Arab Emirates, P.O.BOX: 15551
(e-mail: alig@uaeu.ac.ae)

Abstract. Graphical modeling (GM) plays an important role in providing efficient probability calculations in high dimensional problems (computational efficiency). In this paper, we address one of such problems where we discuss fragmenting puff models and some distributional assumptions concerning models for the instantaneous, emission readings and for the fragmenting process. A graphical representation in terms of a junction tree of the conditional probability breakdown of puffs and puff fragments is proposed.
Keywords: Environmental statistics, Graphical modeling, Junction trees.

1 Introduction

Graphical models, as statistical models, embodying a collection of marginal and conditional independencies which may be summarized by means of a graph, are quickly becoming an integral part of modern statistics. The graphical representation of a statistical model can help in many ways: the graph provides an effective means for elicitation and simplification of a problem, it depicts the dependency structure posited in the model and it may be transformed into a structure that can be used for efficient calculations of various quantities of interest. Graphical methods have been used in the early 1980's for the analysis of statistical problems where no decision variables or utilities are explicitly represented. In a series of papers by(Darroch *et al.*[1]; Lauritzen *et al.*[6]; Kliveri *et al.*[5]; Lauritzen *et al.*[6]; Lauritzen and Wermuth[7]) the authors addressed the problem of how graphs such as influence diagrams can help in understanding the conditional independence properties that a given factorization of a probability density implies. Another issue of importance is how graphs can be used to perform efficient probability calculations in high dimensional problems (computational efficiency). This issue is discussed in a number of papers by (Kim and Pearl[4]; Pearl[11]; Lauritzen and Spiegelhalter[8]; Spiegelhalter *et al.*[18]; Smith and Anderson[17]). In Section 2 we give some graph-theoretic results and a background material on graphs, which are necessary for the development of the paper. In Section 3 we show how to propagate information on junction trees. Section 4 describes an environmental application of a high dimensional process, namely, the atmospheric fragmenting puff models. In this section we propose a graphical representation of the conditional probability breakdown of puffs and puff

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

fragments as a junction tree representation of a high dimensional problem. In Section 5 we give an example of clique representation for puff distributions. Section 6 concludes the paper.

2 Background Material

This section introduces some graph-theoretical terms, which will be used in the paper. A network or *graph* is a pair $G = (V, E)$ that consists of a finite set of vertices $V = 1, 2, \ldots, v$ and a set of edges (arcs) $E \subseteq V \times V$ of ordered pairs of distinct vertices. An edge from vertex i (parent) to vertex j (child) is a *directed* edge (arrow) denoted by $i \rightarrow j$ if $(i, j) \in E$ and $(j, i) \notin E$. If both (i, j) and (j, i) are $\in E$, then the edge between i and j is *undirected* (line). If the graph has only undirected edges, it is *undirected* graph and if all edges are directed, the graph is said to be a *directed* graph. A *path* of length $m \geq 0$ from i to j is an ordered sequence $(i = i_1, i_2, \ldots, i_m = j)$ of distinct vertices $i_1, i_2, \ldots, i_m$ such that (i_l, i_{l+1}) is in E for each $l = 1, 2, \ldots, m$. If there is a path from i to j we say that i leads to j. A subset $C \subseteq V$ is said to be a (i, j) *separator* if all paths from i to j intersect C. The subset C is said to separate A from B if it is an (i, j) separator for every $i \in A, j \in B$. For $A \subseteq V$, the set of parents of A denoted by $P_a(A)$ is the set of all these vertices in V, but not in A that have a child in A. An *m-cycle* is a path of length m with the exception that the end points are equal; that is $i = j$. A graph is *acyclic* if it has no cycles.

2.1 Influence diagrams

An influence diagram (ID) is a schematic representation of conditional independence relationships. It is used for deducing new independencies from those used in the construction of the diagram. Influence diagrams were first developed in the mid 1970's by Miller *et al.*[10], Howard and Matheson[3] extended the theory to decision analysis. Shachter[13] gave a procedure for evaluating a decision problem using an influence diagram. In this section we present a brief introduction on how to use influence diagrams, as a modeling framework, that underpins a probability distribution in order to learn about and calculate various quantities of interest efficiently. We begin by defining a chance influence diagram.

In graph-theoretic terms a chance influence diagram or influence diagram (ID) is a directed graph $G = (V, E)$, where V is a set of nodes represented by circles and called chance nodes and E is the set of directed edges or arrows joining these nodes. Chance nodes label random variables (uncertain) quantities relevant to the problem being modeled and directed edges represent probabilistic dependencies.

A chance node labels a random variable X_1 must be a *parent* of a chance node labels a random variable X_2 if and only if the distribution of the random variable X_2 is calculated conditional on the value of the random variable

X_1 and X_1, X_2 are not independent. The generalization to higher dimensions is given below.
Let $\boldsymbol{X} = (X_1, \ldots, X_m)$ be an ordered set of m random variables with a joint probability function

$$p(\boldsymbol{x}) = p(x_1) \prod_{r=2}^{m} p(x_r | x_1, \ldots, x_{r-1}) \tag{1}$$

Suppose $p(x_r|x_1, \ldots, x_{r-1})$ is a function of x_r and the parent set $P(r) \subseteq \{x_1, \ldots, x_{r-1}\}$ only. This will imply that given $P(r)$, X_r is independent of $R(r)$, where

$$R(r) = \{X_1, \ldots, X_{r-1}\} \setminus P(r)$$

is the set of random variables listed before X_r, which do not appear explicitly in the conditional probability function $p(x_r|x_1, \ldots x_{r-1})$. This can be expressed, as in Dawid[2]'s notation

$$X_r \perp\!\!\!\perp R(r) | P(r) \qquad r = 2, \ldots, m \tag{2}$$

Then the graph of an influence diagram over $X_1, \ldots, X_m$ is any directed graph with nodes representing random variables $X_1, \ldots, X_m$ satisfying property (2). Influence diagrams are clearly acyclic, because only nodes of lower index can be connected to nodes of higher index. As a simple illustration, suppose $\boldsymbol{X} = \{X_1, \ldots, X_8\}$. Then from (1)

$$p(\boldsymbol{x}) = p(x_1) \prod_{r=2}^{8} p(x_r | x_1, \ldots, x_{r-1}).$$

Suppose the parents are: $P(2) = \{X_1\}, P(3) = \{X_1, X_2\}, P(4) = \{X_3\}, P(5) = \{X_3, X_4\}, P(6) = \{\phi\}$ (the empty set), $P(7) = \{X_5, X_6\}, P(8) = \{X_7\}$.
The influence diagram (G) of this example is given in Figure 1

2.2 Clique marginal representation

The clique marginal representation is one of many ways of specifying a joint probability distribution (see, for example, Lauritzen and Spiegelhalter[8]; Smith[14]). We start by identifying the cliques of an influence diagram G and $p(\boldsymbol{x})$ by looking at the small sets of variables called *precliques*, see Smith[15] of the form

$$\tilde{C}(r) = \{X_r, P(r)\} \quad (P(1) = \phi), \quad 1 \leq r \leq m.$$

Then we delete from this collection any preclique $\tilde{C}(r)$ for which there exists a $\tilde{C}(k) \quad (k > r)$ such that

$$\tilde{C}(r) \subseteq \tilde{C}(k).$$

The remaining sets of variables after such deletions are called the *cliques* of $p(\boldsymbol{x})$ and G. This set of cliques will be denoted by $\mathcal{C} = \{C(1), \ldots, C(n)\}, \quad 1 \leq$

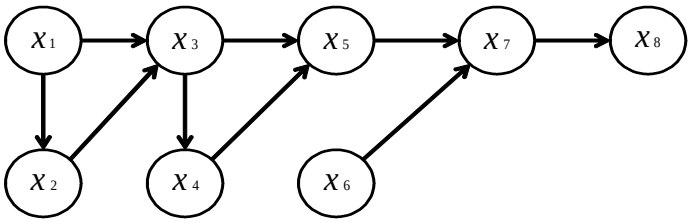


Fig. 1. An Influence Diagram ID I

$n \leq m-1$.
After identifying the cliques, we can determine $p(\boldsymbol{x})$ in terms of the joint probability functions $p_1(\boldsymbol{x}), \ldots, p_n(\boldsymbol{x})$ over the cliques $\{C(1), \ldots, C(n)\}$. A sufficient condition for this is that $p(\boldsymbol{x} \in P(r)) > 0$ for each $\boldsymbol{x} \in P(r), \quad 2 \leq r \leq m$ whenever $P(r) \neq \phi$. Then (1) can be expressed as:

$$p(\boldsymbol{x}) = \frac{\prod_{r=1}^{m} p(\boldsymbol{x} : \boldsymbol{x} \in \tilde{C}(r))}{\prod_{r=2}^{m} p(\boldsymbol{x} : \boldsymbol{x} \in P(r))} \tag{3}$$

where $p(\boldsymbol{x} \in P(r)) = 1$ if $P(r) = \phi$, the empty set.
Since by definition $p(\boldsymbol{x} : \boldsymbol{x} \in \tilde{C}(r))$ (and hence also $p(\boldsymbol{x} : \boldsymbol{x} \in P(r))$ can be obtained from $p(\boldsymbol{x} : \boldsymbol{x} \in C(k))$ where $C(k)$ is a clique of $p(\boldsymbol{x})$ such that $\tilde{C}(r) \subseteq C(k), \quad 2 \leq r \leq m$. Then (3) can be simplified to

$$p(\boldsymbol{x}) = \frac{\prod_{k=1}^{n} p_k(\boldsymbol{x})}{\prod_{k=2}^{n} q_k(\boldsymbol{x})} \tag{4}$$

where $p_k(\boldsymbol{x})$ as defined above and $q_k(\boldsymbol{x}) = p(\boldsymbol{x} : \boldsymbol{x} \in P(r))$ for a $\tilde{C}(r)$ remaining in the clique set, such that $\tilde{C}(r) = C(k), 1 \leq k \leq n$. A set of parents $P(r)$ associated with a clique $C(k)$ is called a *preseparator* and denoted by $\tilde{S}(k), 2 \leq k \leq n$. The clique representation (4) of $p(\boldsymbol{x})$ has many computational advantages as we shall see later on.

2.3 Decomposable influence diagrams

An ID G is called *decomposable* if the set $P(X)$ of direct predecessors of X is completely connected (i.e. each node in $P(X)$ is connected by an edge to another node), this being true for all X in G. Figure 2 illustrates two graphs, one is decomposable and the other is not, since the parent nodes a and b are not joined.

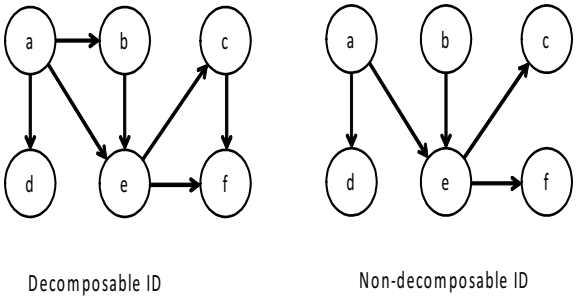


Fig. 2. Graphs of decomposable and non-decomposable ID's

Decomposable influence diagrams have several properties, which make them useful to study. One property is that their structure helps in propagating probabilities as the joint distribution of the system can be stored as margins of cliques. The cliques of a decomposable influence diagram can be ordered. Tarjan and Yannakaskis[19] gave a simple technique for ordering nodes called the maximum cardinality search (MCS), so that in each of its disconnected subgraphs they satisfy the so called *running intersection property* (RIP) which states that: *there exists an ordering* $C[1], \ldots, C[n]$ *of the cliques* $C(1), \ldots, C(n)$ *such that for all* $2 \leq i \leq n$

$$C[i] \cap [\cup_{j=1}^{i-1} C[j]] = S(i) \subseteq C(p_i),$$

for some p_i, $p_i, 1 \leq p_i \leq i-1$.
This means that the intersection of the i^{th} clique with all the preceding ones is a subset of one of the preceding cliques.

3 Junction trees and Probability propagation

The clique representation (4) of $p(\boldsymbol{x})$ can be used efficiently to propagate information through the system, working indirectly with the margins $p_k(\boldsymbol{x})$ and $q_k(\boldsymbol{x})$ successively, updating them rather than updating the whole joint probability function $p(\boldsymbol{x})$ directly. This can be done by passing "simple messages" along the edges of a new graph called a *junction tree*, constructed from the influence diagram of $p(\boldsymbol{x})$. However, in the application cited below, distributions will not always remain decomposable. Because of this we need to define a new graph called *junction graph*, which is an influence diagram on vectors of variables in the original influence diagram of the process. We then show that the definition of a junction tree is just a special case of the undirected version of a junction graph. The use of junction graphs will become apparent later in the paper. A formal definition of a junction graph follows.
A junction graph $\mathcal{G}$ of any density satisfying (4) is a directed graph with n nodes labeling the n cliques $C(1), \ldots, C(n)$. There is an edge to node $C(i)$ from node $C(j), i > j$ if and only if

i) $S(i) \cap C(j) \neq \phi$
ii) there exists no $j' < j$ such that

$$S(i) \cap C(j') \supseteq S(i) \cap C(j).$$

A minimal junction graph $\mathcal{G}$ is a junction graph which has no other junction graph $\mathcal{G}'$ as a proper subgraph.
In general a joint probability function will have several junction graphs and minimal junction graphs over a chosen ordering of its cliques. An influence diagram and its junction graph are shown in Figure 3. The undirected versions of junction graphs are called *junction trees* when the separator of any clique is contained in exactly one previously listed clique or separator. Note that all junction graphs with no unmarried parents and the same undirected version (junction tree) embody an equivalent set of conditional independence statements.
In the case when $p(\boldsymbol{x})$ is decomposable, a collection of disconnected junction trees will be called *a junction forest*.

3.1 Propagation of information on junction trees

Let $\mathcal{C} = \{C(1), \ldots, C(n)\}$ denote the set of cliques of the joint probability function $p(\boldsymbol{x})$. Suppose we learn the values of some or all of the variables lying in some arbitrary clique $C(1) \in \mathcal{C}$ and we want to compute the conditional distribution of all variables in the system given a subset of variables in $C(1)$. To described a propagation algorithm paralleling that given in Lauritzen and Spiegelhalter[8]. It is clear that we can obtain a new probability function

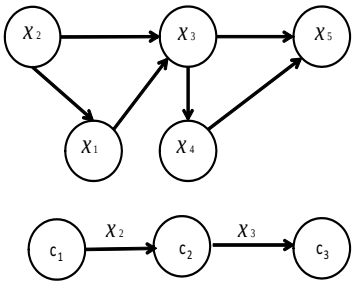


Fig. 3. An influence diagram (ID) and its Junction graph

$p^*(\boldsymbol{x})$ of the variables $\boldsymbol{x}(1)$ in $C(1)$ from $p(\boldsymbol{x}(1))$ its original probability function using Bayes rule. Smith[15] shows how to update probabilities over the variables in the other cliques given the values of some of the variables in $C(1)$. The updating is possible using the junction tree of the system. For detailed discussions, see the above references.

4 Graphical representation of Puff models

4.1 Puff Models

Most interest in the study of probabilistic networks has centered around problems where the junction tree (or variables in that tree) are fixed. However, there is a whole class of spatial temporal processes on which the efficient probability propagation algorithms developed for static networks can be used. For example, one of the methods of modeling atmospheric dispersion after an accidental release of radioactive pollutants is called the puff model Mikkelsen*et al.*[9] According to this model, instead of assuming a continuous release from a source, it is assumed that the mass is released in a series of discrete puffs. These puffs can then be transported and dispersed around the local terrain based on the current wind field and local terrain. This method has been incorporated into the RisØ-Meso-scale Puff model, RIMPUFF Thykier-Nielsen and Mikkelsen[20]. To add to the accuracy of the RIMPUFF model, its designers added a further level of detail, puff splitting or *pentafurcation*. As puffs are released and transported over the local terrain, they grow in size. When their diameter reaches a chosen threshold,

they can split into five smaller puffs. The mass associated with the parent puff is distributed amongst the children which are also smaller in size. In such examples, new variables (puffs) are being continually added so that, at any time in the process, the joint density of all variables up to that time satisfy equation (1).

4.2 Dynamic fragmenting of Puff models

The fragmenting Puff model described above can be reconstructed as a dynamic junction tree Smith *et al.*[16]. In this section we describe briefly the reconstruction procedure starting with notation and distributional assumptions.
Let $m(t, \mathbf{l}) = m(t, l_1, \ldots, l_k)$ be the puff fragment which is the l_kth child of the l_{k-1}th child, ..., of the l_1th child of the puff released at time t. In RIMPUFF $1 \leq l_i \leq 5, 1 \leq i \leq k$. The index k relates to the number of fragmentations that have taken place before fragment $m(t, \mathbf{l})$ appears. Let:
I_T denote the set of all puffs(puff fragments) appearing on or before time T.
$Q(\mathbf{l})$ denote the true mass under $m(t, \mathbf{l})$.
$\bar{\boldsymbol{Q}}(\mathbf{l})$ denote the vector of true masses under the set of the children of $m(t, \mathbf{l})$.
$\boldsymbol{Q}(\mathbf{l}) = (Q(\mathbf{l}), \bar{\boldsymbol{Q}}(t, \mathbf{l}))^T$. Here we consider the following process.

The observation process: Let $\boldsymbol{Q}_T$ be the vector of masses of all puffs and puff fragments emitted on or before time T. Let $\boldsymbol{Y}(t, \boldsymbol{s})$ denote a vector of observations taken at time t at a selection of site(s) $\boldsymbol{s}$. Assume that $\boldsymbol{Y}(t, \boldsymbol{s})|\boldsymbol{\theta}(t, \boldsymbol{s})$ is independent of all other variables in the system. Here $\boldsymbol{\theta}(t, \boldsymbol{s})$ can be interpreted as a random vector relating to the actual mass at time t on site $\boldsymbol{s}$. As a simple process, $\boldsymbol{Y}(t, \boldsymbol{s})|\boldsymbol{\theta}(t, \boldsymbol{s})$ is defined to have a Gaussian distribution with mean $\boldsymbol{\theta}(t, \boldsymbol{s})$ and a fixed covariance matrix V. An important feature of puff models is that at all points $(t, \boldsymbol{s})$ of the observation grid, $\boldsymbol{\theta}(t, \boldsymbol{s})$ can be written as

$$\boldsymbol{\theta}(t, \boldsymbol{s}) = F(t, \boldsymbol{s})\boldsymbol{Q}_t + \boldsymbol{\epsilon}(t, \boldsymbol{s})$$

The matrix $F(t, \boldsymbol{s})$ is a very complicated but known function of $(t, \boldsymbol{s})$, which defines the density of contamination contributed at sites $\boldsymbol{s}$ by each puff or puff fragment at time t. Each row of this matrix corresponds to the weighings used in a dispersal model at a site which is a component of the vector of sites. Notice that $F(t, \boldsymbol{s})$ has non-zero components only on fragments that still exist and have not fragmented further. In practice it is found that only a few puff fragments will be observed at a site at a given time, which implies that for most $(t, \boldsymbol{s})$ many components of each row of $F(t, \boldsymbol{s})$ will be zeros. The error process $\boldsymbol{\epsilon}(t, \boldsymbol{s})$ will be Gaussian with zero mean and fixed covariance matrix U. In the particular case of observations at source $\boldsymbol{s} = \mathbf{0}$, where $\boldsymbol{\theta}(t, \boldsymbol{s})$ is a scalar we set $\boldsymbol{\theta}(t, \boldsymbol{s}) = Q(t)$ and hence $\boldsymbol{\epsilon}(t, \boldsymbol{s}) = 0$. To specify the joint distribution of $\mathbf{Q}_t$ at any time T we need to specify the following processes.

The fragmentation process: This process assumes that a vector of mass fragments (children) $\bar{\boldsymbol{Q}}(\mathbf{l})$ of a parent $m(t, \mathbf{l})$ is independent of all masses

$\boldsymbol{Q}_t$ given the mass $Q(\mathbf{l})$. This can be written as

$$\bar{\boldsymbol{Q}}(\mathbf{l}) \perp\!\!\!\perp \{\boldsymbol{Q}_t \setminus Q(\mathbf{l})\}|Q(\mathbf{l}).$$

Thus, the masses inherited by fragments depend only on the mass of the parent unfragmented puff and no other puff. Thus to specify the joint distribution of puff fragments, it is only necessary to specify the conditional distribution of $\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})$ for each puff/puff fragment $m(t,\mathbf{l})$. To model the dispersal of gas, these conditional distributions are usually chosen to conserve mass. For example, in RIMPUFF model we set

$$E[\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})] = \boldsymbol{\alpha}Q(\mathbf{l}),$$
$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_5)^T,$$
$$\sum_{i=1}^{5} \alpha_i = 1, \alpha_i > 0,$$

and

$$Var[\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})] = B^*,$$

where $\mathbf{1}^T B^* \mathbf{1} = 0$ and $\mathbf{1}$ denotes a vector of ones.
Obviously, if $\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})$ is chosen to be conditionally Gaussian, then this uniquely defines the joint distribution of $\boldsymbol{Q}_t$.

The Emission process: The emission process is modeled as a Dynamic Linear Model (DLM) West and Harrison [21] with state space $(Q(t), \boldsymbol{\psi}_t)^T$ where $\boldsymbol{\psi}_t$ is a vector of dummy variables. Special cases of these models set $\boldsymbol{\psi}(t)$ as null when the process becomes 1-dimensional; $Q(t)|Q(t-1) \sim N[Q(t-1)+\mu(t)-\mu(t-1), W]$ where W is a fixed variance and $\mu(t)$ is a trend term which is a function of time t. This is just a standard state space model on the univariate process $\{Q(t), t = 1, 2, \ldots, \}$. Here, setting the conditional variance $V(t,0)$ of $Y(t,0)$, the source readings, given $Q(t)$ large relative to W gives a process, which after source readings are taken, still preserves strong relationship between masses $Q(t)$ and $Q(t-1)$. On the other hand, if $V(t,0)$ is set to be negligible relative to W, this assumes source readings $Y(t,0), 1 \leq t \leq T$, are very accurate. As a consequence it is not hard to prove that after observing $Y(1,0), \ldots, Y(T,0), \{Q(1), \ldots, Q(T)\}$ are independent and future source emissions $Q(T+k), k = 1, 2, ...$ have expectation $\mu(T+k) - G^k[E[Q(T)] - \mu(T)]$ (say). When the shape of the emission profile is very vague, this can be modeled by setting $\mu(t) = 0, t = 1, 2, ...$ (a steady model). Here the forecast future emission $E[Q(T+k)] = E[Q(T)], k = 1, 2, ...$ i.e. constant. If $Y(T,0)$ is very accurate i.e. $V(T,0)$ is very small relative to W then $E[Q(T+k)] \simeq Y(T,0)$, the last observed emission.

4.3 Clique representation of Puff distributions

Let $\boldsymbol{X}_T$ denote a vector of state random variables of interest (vector of mass emissions and their fragments in our context) existing on or before time T. It

is easy to check that because of the conditional independencies in the system, the joint density $p_T(\boldsymbol{x})$ of $\boldsymbol{X}_T$ can be written as

$$\begin{aligned} p_T(\boldsymbol{x}) = & \, p(Q(1), \boldsymbol{\psi}(1)) \\ & \cdot \prod_{t=2}^{T} p(Q(t), \boldsymbol{\psi}(t) | Q(t-1), \boldsymbol{\psi}(t-1)) \\ & \cdot \prod_{I_T} p(\bar{\boldsymbol{Q}}(\mathbf{l}) | Q(\mathbf{l})) \end{aligned} \tag{5}$$

where $Q(t), \boldsymbol{\psi}(t), \bar{\boldsymbol{Q}}(\mathbf{l}), Q(\mathbf{l})$ and I_T are as defined above. The density can be expressed in a suitable form, namely, the clique marginal representation form of equation (4). For an efficient propagation of probabilities.
Let

$$\begin{aligned} C^*(t) &= \{Q(t), \boldsymbol{\psi}(t), Q(t+1), \boldsymbol{\psi}(t+1)\}, \\ C(\mathbf{l}) &= \{\boldsymbol{Q}(\mathbf{l}), Q(\mathbf{l}, l_1), \ldots, Q(\mathbf{l}, l_5)\}, \mathbf{l} \in I_T \end{aligned}$$

where $C^*(t), C(\mathbf{l})$ are cliques, $1 \leq t \leq T-1$.
Applying equation (4), $p_T(\boldsymbol{x})$ can be written as

$$p_T(\boldsymbol{x}) = \frac{\prod_{1 \leq t \leq T-1} p(C^*(t)) \prod_{\mathbf{l} \in I_T} p(C(\mathbf{l}))}{\prod_{2 \leq t \leq T-1} p(S(t)) \prod_{\mathbf{l} \in I_T} [p(Q(\mathbf{l}))]^{r_T(\mathbf{l})}} \tag{6}$$

where $p(C^*(t))$ and $p(C(\mathbf{l}))$ denote respectively the joint densities of the variables in the cliques $C^*(t)$ and $C(\mathbf{l}), S(t) = \{Q(t), \boldsymbol{\psi}(t)\}$ and $r_T(\mathbf{l})$ is the number of offsprings of $Q(\mathbf{l})$ produced before or at time T. Using this simplified representation, the joint density $p_T(\boldsymbol{x})$ can be stored as a moderate number of joint densities of low dimension instead of a single density of a high dimension.

5 An Illustrative Example

The structure of the joint density $p_T(\boldsymbol{x})$ can be represented by a dynamic influence diagram see, for example, Queen [12] and Smith *et al.*[16]. The nodes of the ID are the random variables (or vectors) defined on the cliques. For example the ID given in Figure 4 represents the conditional probability breakdown of puff and puff fragments in the early stages of an accidental release. As an example, let us assume that a source has emitted 4 puffs at time T, the first puff has pentificated, the 2nd and 5th fragments have then pentificated and further fragmentation has occurred on the 2nd offspring of the 2nd fragment. The second puff has also pentificated and its 2nd puff also split into 5. The 3rd and 4th puffs have yet to fragment.
Here we note that it is easy to check that the ID of Figure 4 is *decomposable* (all parents of a given child are connected) with its cliques having the running

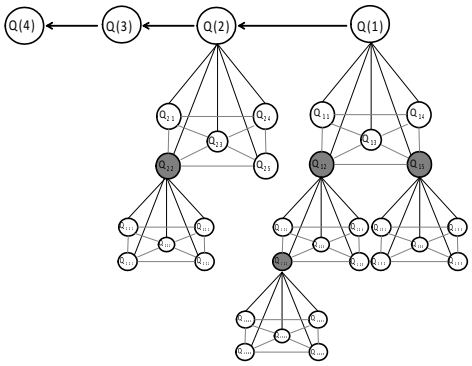


Fig. 4. An ID of early emissions

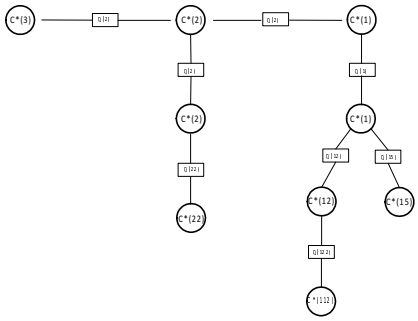


Fig. 5. A junction tree of an ID of early emissions

intersection property (RIP), that is at any time T, the cliques can be ordered as $C[1], \ldots, C[9]$ such that

$$C[i] \cap [\cup_{j=1}^{i-1} C[j]] = S[i] \subseteq C[p_i] \quad 2 \leq i \leq 9.$$

for some $p_i, 1 \leq p_i \leq i-1$. Also we note the following:

(i) If $C[i] = C^*(t)$ then $C[p_i] = C^*(t-1)$ and $S[i] = Q(t)$.
(ii) If $C[i] = C(\mathbf{l})$, if $\mathbf{l} = t$ then $C[p_i] = C^*(t)$ and $S[i] = \{Q(t)\}$, if $\mathbf{l} = (t, l_1, l_k)$ then $C[p_i] = C(t, l_1, l_{k-1})$ and $S[i] = \{Q(\mathbf{l})\}$.

Since the ID is decomposable we can form a junction tree whose nodes are the cliques of $p_T(\boldsymbol{x})$ and whose node $C[i]$ is attached to node $C[p_i]$ by an edge represents a separator $S[i]$. The junction tree which corresponds to the ID of

Figure 4 is shown in Figure 5. A typical clique $\bar{C}[i]$ of this junction tree will have a probability defined conditionally in terms of a particular separator $\bar{S}[i]$ of the junction tree. That separator will take one of the forms:

(a) When $\bar{C}[i] = C^*(t)$ it will take the form $S(t)$ of equation (4).
(b) when $\bar{C}[i] = C(\mathbf{l})$ it will take the form $Q(\mathbf{l})$.

Now an exact algorithm for quick absorption of information on such junction trees which evolve dynamically can easily be adopted.

6 Conclusion

In this work we showed how the continuous release of gas or radioactive material can be described as a series of puffs of contaminated mass emitted sequentially at discrete times and then dispersed and diffused (puff models). We also described a stochastic version of these dispersal models. This version made it possible to incorporate and adjust to uncertain information about contamination readings at different sites. We then proceeded to show that all relevant uncertainties could be modeled by describing the evolution of puffs and puffs fragments within the system by a high dimensional Gaussian process exhibiting many conditional independencies. Finally, a graphical representation (a clique representation) of these fragmentation processes was described. This representation is suitable for an efficient propagation of evidence as it arrives.

References

1. Darroch, J. N., Lauritzen, S. L., and Speed, T. P., *Markov fields and loglinear interaction models for contingency tables*, Annals of Statistics, **8**, 522 - 539, 1980.
2. Dawid, A. P., *Conditional independence in statistical theory*, J. Roy. Statist. Soc. (Ser B), **41** 1 - 31, 1979.
3. Howard, R. A. and Matheson, J. E., *Influence diagrams*, In R. A. Howard and J. E. Matheson Eds. Reading on the Principles and Applications of Decision Analysis, Vol II. Strategic Decisions Group, Menlo Park, Calif. pp 719 - 762, 1981.
4. Kim, J. and Pearl, J., *A computational model for combined causal and diagnostic reasoning in inference systems*, Proc. 8th International Conference on Artificial Intellegence, pp 190 - 193, 1983.
5. Kliveri, H., Speed, T. P., and Carlin, J. B., *Recursive causal models*, J. Austral Math. Soc. (Ser A). **36**, 30 - 51, 1984.
6. Lauritzen, S. L., Speed, T. P., and Vijayan, K., *Decomposable graphs and hypergraphs*, J. Austral. Math. Soc. (Ser. A), **36**, 12 - 29, 1984.
7. Lauritzen, S. L. and Wermuth, N., *Mixed interaction models*, Research Report R 84 - 8, Institute for Elektroniske Systems, Aalborg Universitetscenter, Denmark, 1987.

8. Lauritzen, S. L. and Spiegelhalter, D. J., *Local computations with probabilities on graphical structures and their application to expert systems*, (with discussion), J. R. Statist. Soc. (Ser.B) **50**, 157 - 224, 1988.
9. Mikkelsen, T. Larsen, S. E. and Thykier-Nielsen, S. *Description of RISO Puff Diffusion Model*, Nucl. Safety, **67**, 56 - 65, 1984.
10. Miller, A. C., Merkhofer, M. W., Howard, R. A., Matheson, J. E., and Rice, T. R. *Development of automated aids for decision analysis*, Stanford Research Institute, Menlo Park, Calif, 1976.
11. Pearl, J., *Fusion, Propagation and Structuring in Belief Networks*, AI Journal, **29** (3), 241 - 288, 1986.
12. Queen, C. M., *Bayesian Graphical Forecasting Models for Business Time Series*, Ph.D thesis, Department of Statistics, University of Warwick, 1991.
13. Shachter, R. D., *Evaluating influence diagrams*, Oper. Res., **34** (6), 871 - 882, 1986.
14. Smith, J. Q., *Decision Analysis: A Bayesian Approach*, Chapman and Hall, London, 1988.
15. Smith, J. Q., *Handling Multiple sources of variation using influence diagrams*, European Journal of Operational Research. **86** 189 - 200, 1995.
16. Smith, J. Q., French, S., and Ranyard, D., *An efficient graphical algorithm for updating the estimates of the dispersal of gaseous waste after an accidental release*, Proceedings of Adaptive Computing and Information Processing. Unicom Seminar Ltd., 583 - 610, 1995.
17. Smith, J. Q., and Anderson, P. E., *Conditional independence and Chain Event Graphs*, Artificial Intelligence, 172, 1, 42 - 68, 2008.
18. Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G., *Bayesian Analysis in Expert Systems*, Statistical Science, Vol. **8**, No. 3, 219 - 283, 1993.
19. Tarjan, R. E., and Yannakakis, M. *Simple linear-time algorithms to test chordality of graphs, text acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs*, SIAM J. Comput., **13**, 566 - 579, 1984.
20. Thykier-Neilsen, S., and Mikkelsen, T *RIMPUFF User Guide: Version 30*, National Laboratory, Roskilde, Germany, 1991.
21. West, M., and Harrison, P. J., *Bayesian Forecasting and Dynamic Linear Models*, Spring-Verlag, 1997.

Time Operator and Innovation. Applications to Financial Data

Ilias Gialampoukidis[1] and Ioannis Antoniou[2]

[1] Department of Mathematics, Aristotle University of Thessaloniki
Thessaloniki, 54124, Greece
(E-mail: iliasfg@math.auth.gr)
[2] Department of Mathematics, Aristotle University of Thessaloniki
Thessaloniki, 54124, Greece
(E-mail: iantonio@math.auth.gr)

Abstract. The theory of Time Operators has recently been applied into real life problems with the estimation of innovation probabilities. Based on the assumption that the asset values follow Geometric Brownian Motion with constant variance within each trading day, the internal Age of an asset turns out to be a new statistical index, assessing the average innovations. Moreover, the unpredictability of the t-th observation X_t is estimated by the distribution of innovations of X_t. The innovation probabilities and internal Age are estimated using nonlinear stochastic variance models.
Keywords: Time Operator, Innovation, Financial Data, Stochastic Variance Models.

1 Introduction

The Time Operator of Dynamical Systems [1–3] has been extended to stochastic processes and has been related to the complexity of the stock price dynamics [4]. The application presented in [4] refers to a specific stock from the Athens stock market during the important Greek elections of June 2012, where the distribution of innovations within the eigenspaces of the Time Operator has been computed. In this work, we propose specific models from the literature for the prediction of the distribution of innovations of an asset, within a predefined trading period.

In section 2, we present the Time Operator associated with a stochastic process $X_t, t = 1, 2, \ldots$, through the construction of its eigenprojections, with eigenvalues the times $t = 1, 2, \ldots$. The average value of the Time Operator (Rayleigh quotient) defines the internal Age of the process (section 3). When the process is the evolution of an asset's price, it is shown [4] that the internal Age is a function of the variances of each trading day. The values Open O_τ, High H_τ, Low L_τ and Close C_τ are known for $\tau = 1, 2, \ldots, T$ so using the

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

more efficient daily variance estimator we find, in section 4, the distribution of innovations and their mean, i.e. the internal Age.

The novelty of this work is the estimation of the innovation probabilities p_τ from future variances σ_τ^2 where Open, High, Low, Close values are unknown. In section 5, we shall employ nonlinear stochastic variance models for the evolution of σ_τ^2 and obtain an explicit innovation probability formula corresponding to each model.

In section 6, we present an application to a stock market index. The estimated variances σ_τ^2 from stock market data are transformed using the Box-Cox transformation [5]. The nonlinear stochastic variance model is selected, among the existing ones, as the one with the best fit to the data.

2 Innovations and Time Operator of Stochastic Processes

Consider a stochastic process $X_1, X_2, \ldots$ with the correlation scalar product $< X, Y >= E[XY]$ and denote by $\mathfrak{S}$ the σ-algebra generated by the random variables $X_1, X_2, \ldots$. Assuming that $X_1, X_2, \ldots$ have finite mean and finite correlations, they live in the Hilbert space $L^2(\Omega, \mathfrak{S}, \mu)$, where Ω is their sample space.

The random variables $X_1, X_2, \ldots, X_t$ generate the σ-algebras $\mathfrak{S}_t, t = 1, 2, \ldots$ which define the natural filtration $\{\Omega, \emptyset\} = \mathfrak{S}_0 \subseteq \mathfrak{S}_1 \subseteq \mathfrak{S}_2 \subseteq \cdots \subseteq \mathfrak{S}$ of the stochastic process $X_1, X_2, \ldots$. From the natural filtration $\mathfrak{S}_t, t = 1, 2, \ldots$ of the stochastic process we construct the corresponding sequence of subspaces of $L^2(\Omega, \mathfrak{S}, \mu)$:

$$\mathcal{H}_0 = span\{1_\Omega\}, \quad \mathcal{H}_t = L^2(\Omega, \mathfrak{S}_t, \mu), \quad t = 1, 2, \ldots \tag{1}$$

where $\mathcal{H}_0$ is the Hilbert space of constant random variables. The orthocomplement of $\mathcal{H}_0$ is the Hilbert space $\mathcal{H}$ of fluctuations $\mathcal{H} = L^2(\Omega, \mathfrak{S}, \mu) \ominus \mathcal{H}_0$. Every random variable in $\mathcal{H}$ has the form:$X - E[X]1_\Omega, X \in L^2(\Omega, \mathfrak{S}, \mu)$. The family $\mathcal{H}_t, t = 1, 2, \ldots$ is a resolution of the identity of the Hilbert space $\mathcal{H}$, i.e. $\wedge_{t\in\mathbb{N}}\mathcal{H}_t = \emptyset, \vee_{t\in\mathbb{N}}\mathcal{H}_t = \mathcal{H}$ and $\mathcal{H}_t \subseteq \mathcal{H}_{t+1}, t \in \mathbb{N}$.

The projections: $E_t : \mathcal{H} \to \mathcal{H}_t, t = 0, 1, 2, \ldots$ onto the spaces $\mathcal{H}_t$ are the conditional expectations:

$$\mathbb{E}_t := E[.|\mathfrak{S}_t], t = 0, 1, 2, \ldots \tag{2}$$

and define the resolution of identity operator in $\mathcal{H} : \mathbb{E}_t, t = 0, 1, 2, \ldots$.

Definition 1. The self-adjoint operator with spectral projections the conditional expectations $\mathbb{E}_t$ (2) on the space of fluctuations $\mathcal{H}$ is called *the Time Operator of the stochastic process* $X_t, t = 1, 2, \ldots$:

$$\mathbb{T} = \sum_{t=1}^{\infty} t(\mathbb{E}_t \ominus \mathbb{E}_{t-1}) = \sum_{t=1}^{\infty} t\mathbb{P}_t \tag{3}$$

The eigenspaces of the Time Operator are the Hilbert spaces $\mathcal{N}_t := \mathcal{H}_t \ominus \mathcal{H}_{t-1}, t = 1, 2, \ldots$ and they are called *Innovation Spaces*. The projections $\mathbb{P}_t = \mathbb{E}_t \ominus \mathbb{E}_{t-1}$ onto the innovation spaces $\mathcal{N}_t, t = 1, 2, \ldots$ quantify the innovative part $\mathbb{P}_t Z$ of a random variable Z at time step t.

For example, The Time Operator of Bernoulli Processes [3,4] is applied to the one-dimensional non-stationary random walk

$$X_0 = 0, \quad X_t = Z_1 + Z_2 + \cdots + Z_t, \quad t = 1, 2, \ldots \tag{4}$$

as follows [4]:

$$\mathbb{T} X_t = \sum_{\tau=1}^{\infty} \tau \mathbb{P}_\tau X_t = \sum_{\tau=1}^{\infty} \tau (Z_\tau - E[Z_\tau]) \tag{5}$$

where

$$Z_t = \begin{cases} 1 & \text{with probability } \pi_t \\ -1 & \text{with probability } 1 - \pi_t \end{cases}, \quad t = 1, 2, \ldots \tag{6}$$

3 The Internal Age of an Asset

The innovation probability of a random variable A at time t, is defined as the probability to observe the random variable A in the innovation space $\mathcal{N}_t$:

$$p_t(A) = prob\{A \in \mathcal{N}_t\} = \frac{\|\mathbb{P}_t A\|^2}{\|A - E[A]\|^2} = \frac{Var[\mathbb{P}_t A]}{Var[A]} \tag{7}$$

The Rayleigh quotient (expectation) of the Time Operator $\mathbb{T}$ for the random variable A is called the internal Age of A and is given by the formulas [4]:

$$Age(A) = \frac{< A - E[A], \mathbb{T}(A - E[A]) >}{\|A - E[A]\|^2} = \sum_{t=1}^{\infty} t p_t(A) \tag{8}$$

The internal Age is the average innovation time of the random variable A and $p_t(A), t = 1, 2, \ldots$ is the distribution of innovations within the eigenspaces of $\mathbb{T}$.

When the Time Operator of Bernoulli Processes is applied to one-dimensional non-stationary random walk X_t (4), the internal Age of X_t is a function of the variance of the increments (6) [4, Theorem 5.2]:

$$Age(X_t) = \sum_{\tau=1}^{t} \tau \frac{\sigma_\tau^2}{\sum_{\nu=1}^{t} \sigma_\nu^2}, \quad t = 1, 2, \ldots . \tag{9}$$

where $\sigma_\tau^2 = Var[Z_\tau]$, and the innovation probabilities of the random walk X_t are:

$$p_\tau = \frac{\sigma_\tau^2}{\sum_{\nu=1}^{t} \sigma_\nu^2}, \quad \tau = 1, 2, \ldots, t \tag{10}$$

Formulas (9) and (10) allow estimation of the innovation probabilities and internal Age through the estimation of the variances σ_τ^2.

In order to estimate the internal Age and the innovation probabilities of an asset (stock, currency, etc.), we assume that the asset's prices $X_t, t = 1, 2, \ldots$ are a non-stationary random walk and the index set $X_t, t = 1, 2, \ldots$ of the random walk observations X_t refers to trading days. Some hours of the day the market is open (trading period) and the rest of the day the market is closed. The values Open O_τ, Close C_τ, High H_τ and Low L_τ of the τ-trading day are the available price information of each trading day. Moreover, we assume that the prices of an asset follow Geometric Brownian Motion within each trading day [6]. The variance σ_t^2 is assumed constant with each trading day, but variable from one trading day to another.

At time $t = T$ the observation X_T corresponds to the present asset's price, at the end of today's trading period, so that the values Open O_τ, Close C_τ, High H_τ and Low L_τ are available. Hence, $t = T+1$ will stand for "tomorrow", i.e. the following trading day.

4 Innovation Probability Estimators from High, Low, Open and Closing Prices

In this section we discuss the estimation of the variance $\hat{\sigma}_\tau^2 = \hat{Var}[Z_\tau]$ of the increment Z_τ for each trading day τ for times $\tau \leq T$. In previous work [4], we have presented five popular unbiased estimators, namely the close-to-close estimator $\hat{\sigma}_{CC}^2$ [7], the high-low Parkinson estimator $\hat{\sigma}_P^2$ [7], the Garman-Klass estimator $\hat{\sigma}_{GK}^2$ [8], the Rogers-Satchell estimator $\hat{\sigma}_{RS}^2$ [9] and the Yang-Zhang estimator $\hat{\sigma}_{YZ}^2$ [10]. Among these known variance estimators:

- Parkinson's estimator $\hat{\sigma}_P^2$ [7] and the classic close-to-close estimator $\hat{\sigma}_{CC}^2$ [7] are less efficient than the Rogers-Satchell estimator $\hat{\sigma}_{RS}^2$ [9]
- The Yang-Zhang estimator $\hat{\sigma}_{YZ}^2$ [10] is not able to estimate the variance using data from only one trading day.
- The Garman-Klass estimator $\hat{\sigma}_{GK}^2$ [8] assumes that there is no upward or downward trend, while the Rogers-Satchell estimator $\hat{\sigma}_{RS}^2$ [9] does not.

Due to the above three reasons, the Rogers-Satchell estimator $\hat{\sigma}_{RS}^2$ [9] is the most efficient drift-independent variance estimator, allowing intraday estimations. Hence, the variance of the τ-trading day is given as follows:

$$\hat{\sigma}_{RS(\tau)}^2 = u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau) \tag{11}$$

where $u_\tau = \ln H_\tau - \ln O_\tau, \quad d_\tau = \ln L_\tau - \ln O_\tau, \quad c_\tau = \ln C_\tau - \ln O_\tau$.

Corollary 1. *The innovation probabilities of the asset's price at day t, $t = 1, 2, \ldots, T$ are estimated as follows:*

$$\hat{p}_\tau = \frac{u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)}{\sum_{\tau=1}^{t} u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)} \tag{12}$$

The internal Age of the asset's price at day t is given by:

$$\hat{Age}(X_t) = \sum_{\tau=1}^{t} \tau \frac{u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)}{\sum_{\tau=1}^{t} u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)} \tag{13}$$

Corollary 1, may be used for the estimation of the internal Age and distribution of innovations of an asset's price at day t, when $t \leq T$. In the following section, we modify Eq. (12) and Eq. (13), using nonlinear stochastic variance models for estimations at day $t = T + 1$, i.e. the following trading day.

5 Innovation Probability Estimators from Nonlinear Stochastic Variance models

We shall estimate the distribution of innovations and the internal Age of the random variable X_{T+1}. The formula which gives the innovation probability of X_{T+1} for the following trading day $(T + 1)$ is:

$$\hat{p}_{T+1} = \frac{\hat{\sigma}^2_{T+1}}{\hat{\sigma}^2_{T+1} + \sum_{\tau=1}^{T} \hat{\sigma}^2_{\tau}} \tag{14}$$

where $\hat{\sigma}^2_{T+1}$ is estimated from a stochastic variance evolution model and $\hat{\sigma}^2_{\tau}$ are estimated from Open, High, Low and Closing values $\tau = 1, 2, \ldots, T$. The values $O_{T+1}, C_{T+1}, H_{T+1}$ and L_{T+1} have not been observed yet.

In [11] we find a classification of stochastic variance (volatility squared) models. They all assume that an asset's price, or stock market index, follows a Geometric Brownian Motion with variance σ^2_{τ} evolving according to its own stochastic process. We list the models and the corresponding innovation probability estimators $\hat{p}_{T+1}$, in Table 1.

Reference	**Stochastic Variance Model**	**Innovation Probability** $\hat{p}_{T+1}$
[12]	$\ln \sigma^2_{T+1} = \alpha + \beta\epsilon_T$	$\frac{\exp(\alpha+\beta\epsilon_T)}{\exp(\alpha+\beta\epsilon_T)+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
[13]	$\ln \sigma^2_{T+1} = \gamma(\ln \sigma^2_T - \alpha) + \alpha + \beta\epsilon_T$	$\frac{\exp(\gamma(\ln \sigma^2_T-\alpha)+\alpha+\beta\epsilon_T)}{\exp(\gamma(\ln \sigma^2_T-\alpha)+\alpha+\beta\epsilon_T)+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
[14,15]	$\ln \sigma^2_{T+1} = \ln \sigma^2_T + \alpha + \beta\epsilon_T$	$\frac{\sigma^2_T \exp(\alpha+\beta\epsilon_T)}{\sigma^2_T \exp(\alpha+\beta\epsilon_T)+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
[16–18]	$\sigma_{T+1} = \gamma(\sigma_T - \alpha) + \alpha + \beta\epsilon_T$	$\frac{(\gamma(\sigma_T-\alpha)+\alpha+\beta\epsilon_T)^2}{(\gamma(\sigma_T-\alpha)+\alpha+\beta\epsilon_T)^2+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
[19]	$\sigma^2_{T+1} = \gamma(\sigma^2_T - \alpha) + \alpha + \beta\epsilon_T$	$\frac{\gamma(\sigma^2_T-\alpha)+\alpha+\beta\epsilon_T}{\gamma(\sigma^2_T-\alpha)+\alpha+\beta\epsilon_T+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
[11]	$h(\sigma^2_{T+1}, \delta) = \alpha + \gamma(h(\sigma^2_T, \delta) - \alpha) + \beta\epsilon_T$	$\frac{g(\alpha+\gamma(h(\sigma^2_T,\delta)-\alpha)+\beta\epsilon_T)}{g(\alpha+\gamma(h(\sigma^2_T,\delta)-\alpha)+\beta\epsilon_T)+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$
	$h(\sigma^2_{T+1}, \delta) = h(\sigma^2_T, \delta) + \alpha + \beta\epsilon_T$	$\frac{g(h(\sigma^2_T,\delta)+\alpha+\beta\epsilon_T)}{g(h(\sigma^2_T,\delta)+\alpha+\beta\epsilon_T)+\sum_{\tau=1}^{T}\sigma^2_{\tau}}$

Table 1. Forecasting the innovation probability of the following trading day using stochastic variance models

In Yu *et al.* [11] the model involves the Box-Cox transformation of the variance σ^2_{τ}:

$$h(\sigma^2_{\tau}, \delta) = \begin{cases} \frac{(\sigma^2_{\tau})^{\delta}-1}{\delta} & \text{if } \delta \neq 0 \\ \ln \sigma^2_{\tau} & \text{if } \delta = 0 \end{cases} \tag{15}$$

and the innovation probability p_{T+1} the inverse Box-Cox transformation of the variance σ_τ^2:

$$g(h(\sigma_\tau^2,\delta)) = \begin{cases} (1+\delta\cdot h(\sigma_\tau^2,\delta))^{\frac{1}{\delta}} & \text{if } \delta \neq 0 \\ \exp(h(\sigma_\tau^2,\delta)) & \text{if } \delta = 0 \end{cases} \tag{16}$$

The latest model of Yu *et al.* [11] is a generalization of the previous, most popular models:

- For $\delta = 0$, the model of Yu *et al.* [11] becomes the Wiggins model [13].
- For $\delta = 0.5$, the model of Yu *et al.* [11] becomes the Stein model [17].
- For $\delta = 1$, the model of Yu *et al.* [11] becomes the Andersen model [19].

Combining the most appropriate model (the one with the best fit to our data) with the Rogers-Satchell estimator:

$$\hat{Age}(X_{T+1}) = (T+1)\hat{p}_{T+1} + \sum_{\tau=1}^{T} \tau \frac{u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)}{\hat{\sigma}_{T+1}^2 + \sum_{\tau=1}^{T} u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)} \tag{17}$$

where $\hat{p}_{T+1}$ are given from Table 1 and u_τ, c_τ, d_τ are the quantities of Eq. (12).

In the case of the general model of Yu *et al.* [11] the transformed variances are assumed to follow an Ornstein-Uhlenbeck process [11], which is a mean-reverting process with parameters: α is the long-run variance, γ is the rate at which the transformed variance $h(\sigma_\tau^2,\delta)$ reverts to α, β is the constant variance of the Gaussian increment:

$$h(\sigma_\tau^2,\delta) - \gamma(h(\sigma_{\tau-1}^2,\delta) - \alpha) - \alpha \tag{18}$$

In Zhang and King [20] we find Monte Carlo simulation techniques for the estimation of the parameters $\alpha, \beta, \gamma, \delta$. The process with the independent increments (18) assumes that the variances of an asset, or an index, revert to a constant value α in the long-run. This is an empirical assumption [6] and for statistically significant values of γ estimated to be close to 1 (as in [20]) the model (19) can be simplified to assume that the variance increments

$$h(\sigma_\tau^2,\delta) - h(\sigma_{\tau-1}^2,\delta) \tag{19}$$

are normally distributed, with constant mean and variance. This is the case we examine in the following section. The most related work to this model in the literature is the NARCH model of Higgins and Bera [21,22] where the transformed errors of a time series follow an AR(p) process, generalizing the ARCH and GARCH models of Engle [23] and Bollersev [24] respectively.

6 Application to Athens Stock Market General Index

We assume that the variance σ_τ^2 within each trading day is constant and we use the Rogers-Satchell estimator (11) [9] which is more efficient than the classic close-to-close estimator.

δ	**p-value**	**Decision**	δ	**p-value**	**Decision**
0,20	0,019	Reject	0,33	0,106	Accept
0,21	0,029	Reject	0,34	0,112	Accept
0,23	0,06	Accept	0,36	0,092	Accept
0,24	0,069	Accept	0,37	0,086	Accept
0,25	0,092	Accept	0,38	0,076	Accept
0,26	0,112	Accept	0,39	0,061	Accept
0,27	0,122	Accept	0,40	0,061	Accept
0,28	0,148	Accept	0,41	0,051	Accept
0,29	0,143	Accept	0,42	0,043	Reject
0,30	0,164	Accept	0,43	0,036	Reject
0,31	0,14	Accept	0,44	0,025	Reject
0,32	0,106	Accept	0,45	0,02	Reject

Table 2. The normality of the increments is tested for several values of δ using the Kolmogorov-Smirnov test

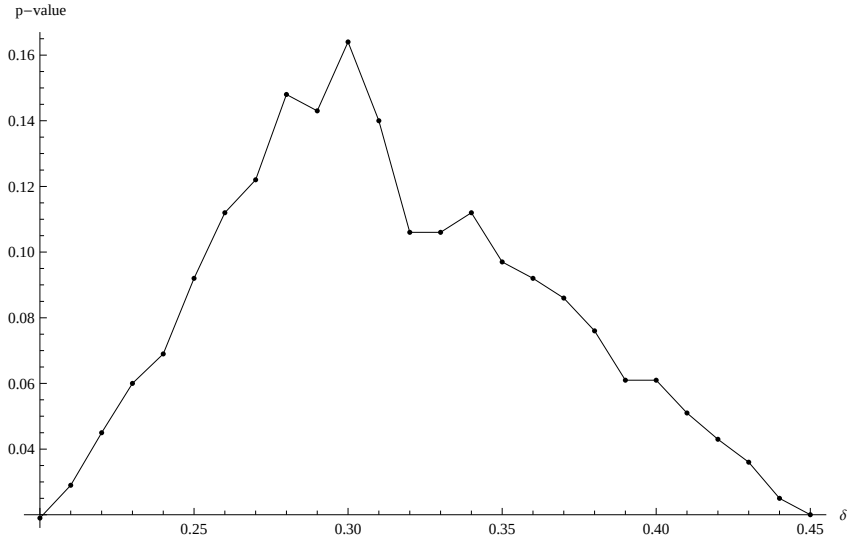


Fig. 1. The region $\delta \in [0.23, 0.41]$ involves transformed variances which are normally distributed (with significance level 0.05). The highest p-value (0.164) is attained at $\delta = 0.30$

We found evidence that the variances σ_τ^2 the Athens General Stock Market Index from 18$^{\text{th}}$ February 2009 to 17$^{\text{th}}$ February 2014 are determined by the model:

$$h(\sigma_\tau^2, \delta) - h(\sigma_{\tau-1}^2, \delta) = \alpha + \beta \cdot \epsilon_\tau \tag{20}$$

where $\hat{a} = 0.0001, \hat{\beta} = 0.1016, \hat{\delta} = 0.30$ and ϵ_τ are Gaussian zero centered, unit variance random variables.

We could not reject the hypothesis that the transformed variances $h(\sigma_\tau^2, \delta) - h(\sigma_{\tau-1}^2, \delta)$ of the Athens General Stock Market Index from 18$^{\text{th}}$ February 2009 to 17$^{\text{th}}$ February 2014 are normally distributed for $\delta \in [0.23, 0.41]$. The

Kolmogorov-Smirnov tests and their corresponding decision are shown in Table 2. Figure 1 shows the p-value of each Kolmogorov-Smirnov test corresponding to each selection of the power δ. Data have been obtained from [25] and the computations have been done in SPSS Statistics 20.

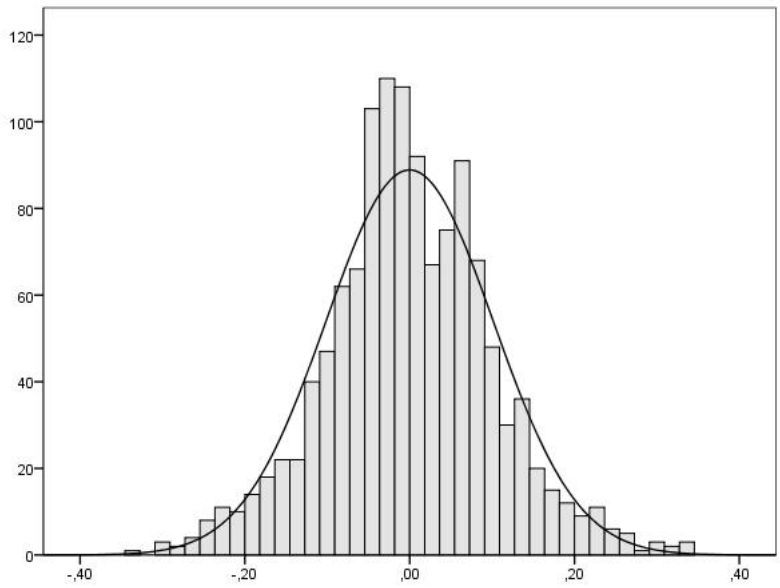


Fig. 2. Histogram of the transformed variances with the estimated normal curve, for $\delta = 0.30$

We report some of the statistics of the variable $h(\sigma_\tau^2, 0.30) - h(\sigma_{\tau-1}^2, 0.30)$. The sample has 1245 values having mean -0.0001, standard deviation 0.1016, skewness 0.077 and kurtosis 0.556. The spectrum of the sample is in the interval $[-0.3348, 0.3339]$.

Corollary 2. *The innovation probability p_{T+1} for the Athens stock market general index is:*

$$p_{T+1} = \frac{\sigma_{T+1}^2}{\sigma_{T+1}^2 + \sum_{\tau=1}^{T} u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)} \tag{21}$$

where $\sigma_{T+1}^2 = \sqrt[0.30]{(u_T(u_T - c_T) + d_T(d_T - c_T))^{0.30} + 0.03\epsilon_T}$.

Proof. The highest p-value (0.164) is attained at $\delta = 0.30$, so the variance evolution model is:

$$\frac{(\sigma_\tau^2)^\delta - 1}{\delta} - \frac{(\sigma_{\tau-1}^2)^\delta - 1}{\delta} = \alpha + \beta\epsilon_\tau \tag{22}$$

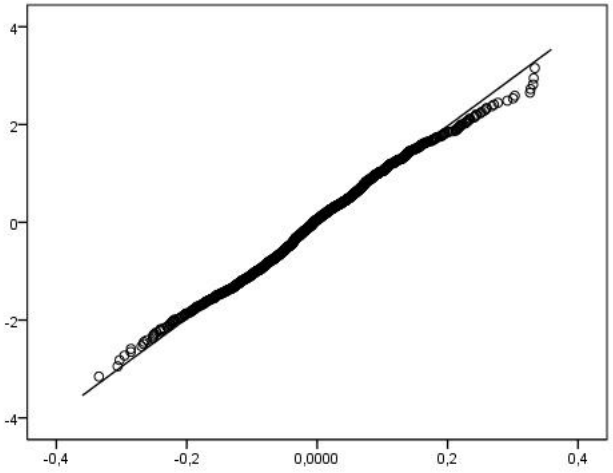


Fig. 3. Q-Q plot of the transformed variances for $\delta = 0.30$

where $\alpha = -0.0001, \beta = 0.1016, \epsilon_\tau$: zero mean, unit variance normally distributed random variable. Equivalently, $(\sigma_\tau^2)^\delta - (\sigma_{\tau-1}^2)^\delta = \delta\alpha + \delta\beta\epsilon_\tau$. Therefore, the variance is $\sigma_\tau^2 = \sqrt[\delta]{(\sigma_{\tau-1}^2)^\delta + \delta\alpha + \delta\beta\epsilon_\tau}$ and the innovation probability p_{T+1} (21) is proved after substitution of the Rogers-Satchell estimator (11) to the variances $\sigma_\tau^2, \tau = 1, 2, \ldots, T$ and taking into account that $\delta\alpha = -0.00003 \cong 0$.

From the definition of the internal Age (8), formula (9) and Corollary 2, the internal Age computation is straightforward:

Corollary 3. *The internal Age of the following trading day is:*

$$\hat{Age}(X_{T+1}) = (T+1)\hat{p}_{T+1} + \sum_{\tau=1}^{T} \tau \frac{u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)}{\hat{\sigma}_{T+1}^2 + \sum_{\tau=1}^{T} u_\tau(u_\tau - c_\tau) + d_\tau(d_\tau - c_\tau)} \tag{23}$$

where $\sigma_{T+1}^2 = \sqrt[0.30]{(u_T(u_T - c_T) + d_T(d_T - c_T))^{0.30} + 0.03\epsilon_T}$ *and* $\hat{p}_{T+1}$ *is given by Corollary 2.*

The transformed variance increments $(\sigma_\tau^2)^\delta - (\sigma_{\tau-1}^2)^\delta$ are expected to be practically zero:

$$E[(\sigma_\tau^2)^\delta - (\sigma_{\tau-1}^2)^\delta] = E[\delta\alpha + \delta\beta\epsilon_\tau] = \delta\alpha + \delta\beta E[\epsilon_\tau] = \delta\alpha = -0.00003 \cong 0$$

Therefore, we may apply Corollaries 2 and 3 in any period of T trading days for the estimation of the innovation probability based on Corollary 2 (expected innovation probability) and compare to the innovation probability estimations based on Open, High, Low and Close prices. Moreover, we may also estimate the internal Age, based on Corollary 3 (expected internal Age) and compare to the internal Age estimations based on Open, High, Low and Close prices.

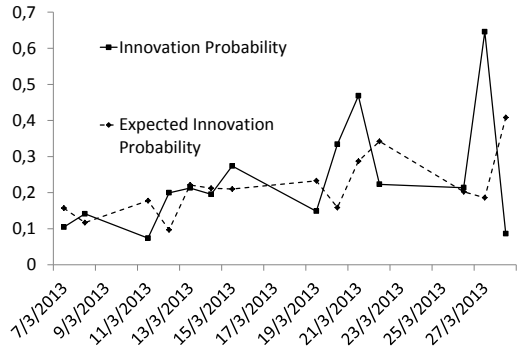


Fig. 4. The innovation probability of the following trading day based on Open, High, Low and Close prices (solid line). The expected innovation probability (dashed line) is based on Corollary 2.

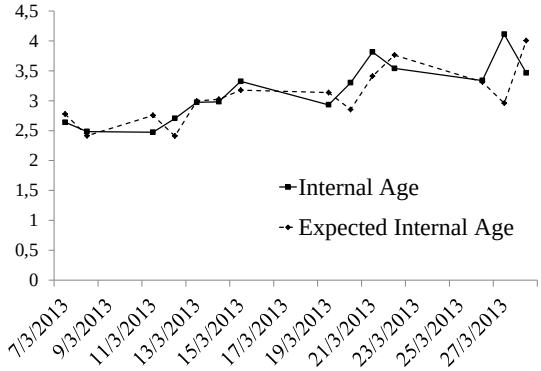


Fig. 5. The internal Age of the following trading day based on Open, High, Low and Close prices (solid line). The expected internal Age (dashed line) is based on Corollary 3.

We apply previous analysis on March 2013. This month has an extreme value related to the Cyprus bailout deal where large deposits were seized and the second-largest bank closed [26].

Based on the previous 3 trading days and the available information at the end of the present trading day ($T = 4$), we find the innovation probability and the internal Age of the following trading day. The values Open, Close, High, Low of the 5^{th} trading day are not known yet and the expected innovation probability and the expected internal Age are estimated from Corollaries 2 and 3 respectively. At the end of the 5^{th} trading day, we estimate the innovation probability and the internal Age using the Rogers-Satchell estimator (Corollary 1). The expected innovation probabilities (Corollary 2) are compared to the innovation probabilities (Corollary 1) in Figure 4. The expected internal Age (Corollary 3) is compared to the internal Age (Corollary 1) in Figure 5.

The internal Age of 5 successive trading days in March 2013 attains its maximum value at 27$^{\text{th}}$ March, i.e. one day after the announcement trading day date (26$^{\text{th}}$ March 2013). Greek firms with large deposits in Cypriot banks and projects running in Cyprus caused a strong impact on the Athens stock market index. This impact is quantified through the innovation probabilities and the internal Age of the corresponding dates demonstrating the high complexity of this trading period.

The estimations close to the 27$^{\text{th}}$ of March 2013 are not satisfactory, due to the external force that changed the dynamics of the stock value process. As long as the prices fluctuate according to its own innovations, the prediction of the innovation probabilities is satisfactory. The change in the dynamics was caused by a political decision, unpredictable so far for many scientists.

7 Concluding Remarks

We used the Rogers-Satchell estimator of the daily variance, which is more efficient than the classic close-to-close estimator and drift-independent. The most recent variance estimator of Yang and Zhang [10] uses data from more than one trading days.

From Corollary 3 we see that internal Age estimations are computed from past variances, using the Rogers-Satchell estimator, and future variances, using a model from Table 1. The most appropriate model is selected as the one with best fit to the corresponding data.

In case the independent increments Z_τ of Eq. (4) of the asset price dynamics are not Gaussian, Mandelbrot [27] proposed the Pareto distribution to model the changes in the logarithm of cotton prices. These distributions have infinite variance:

$$Var[Z_\tau] = \sigma_\tau^2 = \infty \tag{24}$$

In case Eq. (24) is true, i.e. the increments Z_τ of Eq. (4) have infinite variance, the innovation probability is always 100%:

$$\lim_{\sigma_{T+1}^2 \to \infty} p_{T+1} = \frac{\sigma_{T+1}^2}{\sigma_{T+1}^2 + \sum_{\tau=1}^{T} \sigma_\tau^2} = \frac{1}{1 + \frac{\sum_{\tau=1}^{T} \sigma_\tau^2}{\sigma_{T+1}^2}} \to 1 \tag{25}$$

The prediction in such a complex environment is not expected to be successful, no matter how many previous observations are used.

As shown in Figure 6, the survival function in logarithmic scale does not fit to a straight line. Therefore, the Pareto distribution does not fit to the differences of the logarithms of the closing prices.

In case the predicted variance $\hat{\sigma}_{T+1}$ of the following trading day is estimated from a constant known value c (for example the 100-year average variance), Eq. (14) is of the form:

$$\hat{p}_{T+1} = \frac{\hat{\sigma}_{T+1}^2}{\hat{\sigma}_{T+1}^2 + \sum_{\tau=1}^{T} \hat{\sigma}_\tau^2} = \frac{c}{c + x} \tag{26}$$

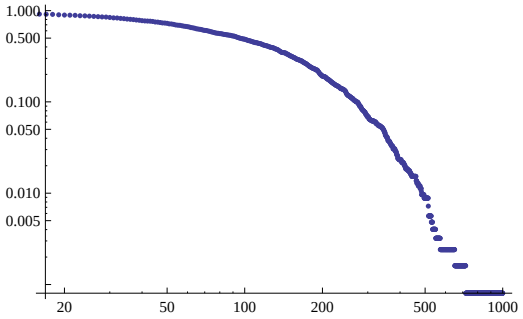


Fig. 6. LogLogPlot of the survival function $Prob[\ln C_t - \ln C_{t-1} > k]$. Not power-law scaling.

Eq. (24) shows that periods of low uncertainty and variance ($x \to 0$) imply high innovation probability $\hat{p}_{T+1} \to 1$ the following trading day. Moreover, periods of high uncertainty and variance ($x \to \infty$) imply low innovation probability $\hat{p}_{T+1} \to 0$.

We estimated the innovation probabilities and the internal Age of the Athens general stock market index observations during March 2013. The significant event associated with this month is the recent bailout program of Cyprus, resulting to a severe local downward trend at the Athens General stock market index. The high complexity (measured here from the innovation probabilities and the internal Age) of specific trading dates is not affected by the sign of the local drift, i.e. it does not matter whether the local trend is upward or downward. We have recently illustrated the increasing distribution of innovations as we approach the important Greek elections of June 2012 [4], leading to an upward trend. The average innovation time is important for the risk assessment of specific trading days.

Acknowledgements

We thank the Aristotle University of Thessaloniki and especially the Research Committee for supporting one of us (Ilias Gialampoukidis) by awarding him the Excellence Scholarship 2013. We also thank Karl Gustafson for useful discussions and comments.

References

1. Misra B., Prigogine I. and M. Courbage. From deterministic dynamics to probabilistic descriptions. *Physica A*, 98, 1, 1–26, 1979
2. I. Prigogine. *From Being to Becoming*. Freeman, New York, 1980
3. M. Courbage and B. Misra. On the equivalence between Bernoulli dynamical systems and stochastic Markov processes. *Physica A*, 104, 3, 359–377, 1980
4. Gialampoukidis I., Gustafson K. and I. Antoniou. Financial Time Operator for Random Walk Markets. *Chaos, Solitons & Fractals* 57, 62–72, 2013

5. Box G.E.P, and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society* B, 211–252, 1964
6. J. Hull. *Options, Futures, and other Derivatives*, seventh ed., Pearson, New Jersey, 2009
7. M. Parkinson. The extreme value method for estimating the variance of the rate of return. *Journal of Business* 53, 61–65, 1980
8. M. Garman and M.J. Klass. On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business* 53, 67–78, 1980
9. L. C. G. Rogers and S.E. Satchell. Estimating Variance from High, Low and Closing Prices. *Annals of Applied Probability* 1, 504–12, 1991
10. D. Yang and Q. Zhang. Drift Independent Volatility Estimation Based on High, Low, Open and Close Prices. *Journal of Business* 73, 477–491, 2000
11. Yu J., Yang Z. and X. Zhang. A class of nonlinear stochastic volatility models and its implications for pricing currency options. *Computational Statistics & Data Analysis* 51, 2218–2231, 2006
12. P. K. Clark. A subordinated stochastic process model with finite variance for speculative price. *Econometrica* 68, 135–155, 1973
13. J. B. Wiggins. Option values under stochastic volatility: Theory and empirical estimates. *Journal of financial economics* 19, 2, 351–372, 1987
14. Hull J. and A. White. The pricing of options on assets with stochastic volatilities. *The journal of finance* 42, 281–300, 1987
15. Johnson H. and D. Shanno. Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis* 22 ,143–152, 1987
16. L. O. Scott. Option pricing when the variance changes randomly: theory, estimation and an application. *Journal of Financial and Quantitative Analysis* 22, 419–439, 1987
17. Stein E. M. and J. C. Stein. Stock price distributions with stochastic volatility: an analytical approach. *Review of financial studies* 4, 727–752, 1991
18. S. L. Heston. A closed-form solution for options with stochastic volatility, with application to bond and currency options. *Review of financial studies* 6, 327–343, 1993
19. T. Andersen. Stochastic autoregressive volatility: a framework for volatility modeling. *Mathematical finance* 4, 75–102, 1994
20. Zhang X. and M. L. King. Box-Cox stochastic volatility models with heavy-tails and correlated errors. *Journal of Empirical Finance* 15, 3, 549–566, 2008
21. Higgins M. L. and A. K. Bera. A class of nonlinear ARCH models. Unpublished manuscript (Department of Economics, University of Illinois, Champaign, IL), 1989
22. Higgins M. L. and A. K. Bera. A class of nonlinear ARCH models. *International Economic Review* 137–158, 1992
23. R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*. 987–1007, 1982
24. T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31.3, 307–327, 1986
25. http://www.naftemporiki.gr/finance/quote?id=GD.ATH
26. http://www.independent.co.uk/news/world/europe/cyprus-bailout-large-deposits-will-be-seized-and-secondlargest-bank-closed-in-10bn-eu-bailout-deal-8547104.html
27. B. Mandelbrot. The Variation of Certain Speculative Prices. *The Journal of Business* 36, 394–419, 1963

On the probabilistic structure of power TGARCH models and applications to real data

E. Gonçalves[1], J. Leite[2], and N. Mendes-Lopes[1]

[1] CMUC, Department of Mathematics, University of Coimbra, Portugal
(e-mail: esmerald@mat.uc.pt, nazare@mat.uc.pt)
[2] CMUC, Instituto Politécnico de Coimbra, ISCAC, Portugal
(e-mail: jleite@iscac.pt)

Abstract. Power TGARCH models are a natural extension of threshold GARCH processes that allows taking into account both long memory and asymmetry in the stochastic volatility. In Gonçalves, Leite, Mendes-Lopes[5] such models, with real power δ and general error process, have been developed by establishing their main probabilistic properties. The aim of this paper is to enhance the practical interest of real power models by showing their adequacy to describe a physical time series, namely the areas of the plage regions of the Sun, that is, the bright regions in the chromosphere of the Sun, measured by the percentage area of the regions of solar activity in one of the hemispheres relatively to its visible area. With this goal, after recalling the main probabilistic properties of the real power models, we describe the dynamical behavior of daily solar activity modeling the evolution of the plage regions observed in the South solar hemisphere and measured in the Ca II K3 Coimbra's spectroheliograms between 1976 and 2006.

Keywords: Stochastic modeling, Time series, Power TGARCH models..

1 Introduction

Time series modeling has undergone an important development in last years. The linear formulation present in the classical autoregressive moving average (ARMA) models have been found to be insufficient to describe adequately some data, like financial, monetary and physical one. In fact, this kind of time series presents features of non-linearity behavior, particularly the fact that its conditional volatility depends strongly on the past. In order to best describe this fact, several conditional heteroscedastic models appeared in the literature following the seminal paper of Engle[4].

Another fact often found in those time series is the asymmetrical reaction of the volatility according to the sign of past observations, namely its different behavior during a rising or falling period. This feature is taken into account in the threshold ARCH models in which the conditional standard deviation of the process at time t is a piecewise linear function of negative and positive values of past observations. Similarly, the presence of long memory in the shocks of the conditional variance contributed to the proposal of power conditional

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

heteroscedastic models, among others, by Ding, Granger and Engle[3]. Following this original idea, we developed in Gonçalves, Leite and Mendes-Lopes[5] a natural extension of TGARCH processes that allows to take into account both long memory property and the asymmetry in the stochastic volatility namely, the TGARCH model with real power δ (δ-TGARCH).

In this paper we present this model referring some of their principal probabilistic properties, namely stationarity, ergodicity and δ-order moments existence; moreover a representation of the conditional volatility as function of the present and past observations is considered.

We apply these processes to the study of the dynamical behavior of daily solar activity, based on the plage region areas observed in the South solar hemisphere and measured in the Ca II K3 Coimbra's spectroheliograms between 1976 and 2006 analogously to what is done in Gonçalves et al.[6]. Our study indicates that the temporal evolution of this series is well described by ARMA processes with δ-TGARCH errors and that the conditional volatility, strongly present in solar activity, is not well reproduced by these models with integer power.

2 $\delta-$ TGARCH processes

2.1 Definition

Let $X = (X_t, t \in \mathbb{Z})$ be a real stochastic process and, for any $t \in \mathbb{Z}$, let us consider $X_t^+ = X_t \mathbb{I}_{\{X_t \geq 0\}}$, $X_t^- = -X_t \mathbb{I}_{\{X_t < 0\}}$ and $\underline{X}_t$ the $\sigma-$ field generated by $(X_{t-i}, i \geq 0)$.

The stochastic process $X = (X_t, t \in \mathbb{Z})$ is said to follow a δ-power threshold generalized autoregressive conditional heteroscedastic (δ-TGARCH) mo- del with orders p and q ($p, q \in \mathbb{N}$) if, for every $t \in \mathbb{Z}$, we have

$$\begin{cases} X_t = \sigma_t \varepsilon_t \\ \sigma_t^\delta = \omega + \sum_{i=1}^{p} \left[\alpha_i \left(X_{t-i}^+ \right)^\delta + \beta_i \left(X_{t-i}^- \right)^\delta \right] + \sum_{j=1}^{q} \gamma_j \sigma_{t-j}^\delta \end{cases}$$

for some real constants $\delta \neq 0$, $\omega > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$, $i = 1, ..., p$, $\gamma_j \geq 0$, $j = 1, ..., q$, and where $\varepsilon = (\varepsilon_t,\ t \in \mathbb{Z})$ is a sequence of independent and identically distributed real random variables such that ε_t is independent of $\underline{X}_{t-1}$, for every $t \in \mathbb{Z}$. If $\delta < 0$ we consider the following convention: $\left(X_t^+\right)^\delta = 0$ if $X_t \leq 0$ and $\left(X_t^-\right)^\delta = 0$ if $X_t \geq 0$, for every $t \in \mathbb{Z}$. ε is called the generator process of X.

If $\gamma_j = 0$, $j = 1, ..., q$, the $\delta-$TGARCH(p, q) model is simply denoted $\delta-$TARCH(p).

We observe that for these processes the TGARCH equation propagates not just the conditional standard deviation but, more generally, the absolute moments of order δ.

With this general formulation we include the principal conditional heteroscedastic models, namely:

1) GARCH (Engle[4], Bollerslev[2]): considering $\delta = 2$ and $\alpha_i = \beta_i$, $i = 1, ..., p$.

2) TGARCH (Zakoian[10]): considering $\delta = 1$.

3) $\delta-$GARCH, $\delta > 0$ (Mittnik, Paolella and Rachev[7]): considering $\alpha_i = \beta_i, i = 1, ..., p$. In fact, $\alpha_i \left[\left(X_{t-i}^{+} \right)^{\delta} + \left(X_{t-i}^{-} \right)^{\delta} \right] = \alpha_i \left[X_{t-i}^{+} + X_{t-i}^{-} \right]^{\delta} = \alpha_i \left| X_{t-i} \right|^{\delta}$.

4) APARCH (Ding, Granger and Engle[3]), considering $\alpha_i = a_i \left(1 - \tau_i\right)^{\delta}$ and $\beta_i = a_i \left(1 + \tau_i\right)^{\delta}$, where $a_i \geq 0$, $|\tau_i| \leq 1$, $i = 1, ..., p$, and $\delta > 0$.

Some essential probabilistic properties of these processes are referred in the following section.

2.2 Probabilistic structure

This class of power-transformed and threshold GARCH models was considered by Pan, Wang and Tong[8] for $\delta > 0$, for which they established conditions of the strict stationarity and the existence of moments. In a more general framework, in what concerns the power δ, the coefficients and generator process distributions (no moments assumptions and not necessarily symmetric generator process) we establish (Gonçalves, Leite, and Mendes-Lopes[5]):

i) the existence of an unique strict stationary and ergodic solution,

ii) a necessary and sufficient condition of existence of the order δ moment under which the strict stationarity is satisfied,

iii) the weak stationarity up to the δ-order.

To establish the existence and unicity of a strict stationary and ergodic solution it is crucial to find a Markovian representation of the model, involving a strictly stationary and ergodic process from which that solution is deduced. Following the idea present in Mittnik, Paolella and Rachev[7], the following vectorial representation

$$Y_{t+1} = A_t Y_t + B$$

is obtained considering the $\mathbb{R}^m$-vectorial stochastic process $Y = (Y_t, t \in \mathbb{Z})$, $m = \max(p, q)$, where the k-component, $Y_t^{(k)}$ is

$$\begin{cases} Y_t^{(1)} = \sigma_t^{\delta} \\ Y_t^{(k)} = \sum\limits_{i=k}^{m} \left[\alpha_i \left(X_{t-i+k-1}^{+} \right)^{\delta} + \beta_i \left(X_{t-i+k-1}^{-} \right)^{\delta} + \gamma_i \sigma_{t-i+k-1}^{\delta} \right], \; k = 2, ..., m, \end{cases}$$

with $(A_t, \; t \in \mathbb{Z})$ a sequence of independent and identically distributed random square matrices of order m and where B is a determinist vector of $\mathbb{R}^m$ given by

$$A_t = \begin{bmatrix} \sum\limits_{i=1}^{m-1} \left[\alpha_i \left(\varepsilon_t^{+} \right)^{\delta} + \beta_i \left(\varepsilon_t^{-} \right)^{\delta} + \gamma_i \right] e_i & I_{m-1} \\ \alpha_m \left(\varepsilon_t^{+} \right)^{\delta} + \beta_m \left(\varepsilon_t^{-} \right)^{\delta} + \gamma_m & 0_{m-1}^{T} \end{bmatrix}, \qquad B = \begin{bmatrix} \omega e_1 \\ 0 \end{bmatrix};$$

($e_1, ..., e_{m-1}$ is the canonical base of $\mathbb{R}^{m-1}$, I_{m-1} the identity matrix of $m-1$ order and 0_{m-1} the null vector of $\mathbb{R}^{m-1}$).

We remark that the probabilistic analysis developed has enormous impact on statistical applications of such models, in particular, since we ensure the existence of stationary and ergodic solutions under conditions of great simplicity expressed in terms of the model coefficients. In the following theorem we summarize the particular results obtained.

Supposing that

(**H1**): $E\left(|\varepsilon_t|^{\delta}\right) < +\infty$ and $P(\varepsilon_t = 0) \neq 1$,

and denoting $E\left(|\varepsilon_t|^{\delta}\right) = \phi_{\delta}$, $E\left[\left(\varepsilon_t^{+}\right)^{\delta}\right] = \phi_{1,\delta}$ and $E\left[\left(\varepsilon_t^{-}\right)^{\delta}\right] = \phi_{2,\delta}$, we have the following result.

Theorem 1:

1. Under (**H1**), $E\left(|X_t|^{\delta}\right)$ exists and is independent of t if and only if $\sum_{i=1}^{m}\left(\alpha_i\phi_{1,\delta} + \beta_i\phi_{2,\delta} + \gamma_i\right) < 1$. Moreover X is weak stationary up to the δ order.

2. Under (**H1**) and if $\sum_{i=1}^{m}\left(\alpha_i\phi_{1,\delta} + \beta_i\phi_{2,\delta} + \gamma_i\right) < 1$, then X is strictly stationary and ergodic.

We finalize this brief review on the $\delta-$TGARCH processes probabilistic structure, presenting the representation of the conditional volatility as function of the present and past observations of the process.

Let us consider $G(x) = 1 - \gamma_1 x - ... - \gamma_q x^q$. If $\gamma_1 + ... + \gamma_q < 1$ we may introduce the coefficients d_i such that $\frac{1}{G(x)} = \sum_{i=0}^{+\infty} d_i x^i$, $|x| \leq 1$ and define for $j \in \mathbb{N}$, $c_j = \alpha_1 d_{j-1} + ... + \alpha_p d_{j-p}$, $\widetilde{c}_j = \beta_1 d_{j-1} + ... + \beta_p d_{j-p}$, where $d_{k-j} = 0$ if $j > k$.

Considering a $\delta-$TGARCH process X with identically distributed components such that

(**H2**): $E\left(\log^{+}|\varepsilon_0|\right) < +\infty$ and $E\left(\log^{+}\sigma_0\right) < +\infty$,

the following result is established.

Theorem 2: If $\gamma_1 + ... + \gamma_q < 1$ then

$\sigma_t^{\delta} = c_0 + \sum_{i=1}^{+\infty} c_i\left(X_{t-i}^{+}\right)^{\delta} + \sum_{i=1}^{+\infty}\widetilde{c}_i\left(X_{t-i}^{-}\right)^{\delta}$, almost surely,

with coefficients c_i and $\widetilde{c}_i$ that decrease exponentially. If, in addition, ε_0^{+} and ε_0^{-} are non-degenerated random variables, the given representation is unique.

We note that the condition $\gamma_1 + ... + \gamma_q < 1$ is a necessary condition of stationarity; so, for a strictly stationarity $\delta-$TGARCH process X satisfying (**H2**), a δ-TARCH(∞) representation exists and as the coefficients, c_j and $\widetilde{c}_j$, of this representation decrease exponentially to zero, as $j \longrightarrow +\infty$, σ_t is approximated, in a convergent way, using a finite sample of X.

2.3 Application to real data

Solar activity features has been extensively studied using the classical ARMA models as for the study of the temporal evolution of sunspots numbers. Considering another solar feature, the plage regions areas, we develop a temporal analysis on the dynamical behavior of daily solar activity (Gonçalves et al.[6]). This study is based on the areas of the plage regions observed in each one of the solar hemispheres and measured in the Ca II K3 Coimbra's spectroheliograms between 1976 and 2006.

To illustrate the importance of considering $\delta-$TGARCH processes to describe the dynamical evolution of this kind of data, we consider here the series of plage region areas daily observed in South solar hemisphere. This series is analyzed using a generalization of the classical Box-Jenkins methodology in order to take into account the features of conditional volatility that we have detected in the residual series of the model firstly deduced by the classical procedure.

All statistical analysis were performed using the statistical software Eviews.

2.3.1. The sample: south plage region areas

Let us consider the series of the plage region areas daily observed in South solar hemisphere between 1976 and 2006.

As the temporal evolution analysis requires observations equally spaced in time, a random average methodology is implemented to estimate some missing observations due to weather conditions. In the Figure 1 we present the trajectories of the observed plage regions areas series (in blue) and of the corresponding completed series (in red), in the South hemisphere. These series are denoted respectively by SOUTHOBS and SOUTHCOMP.

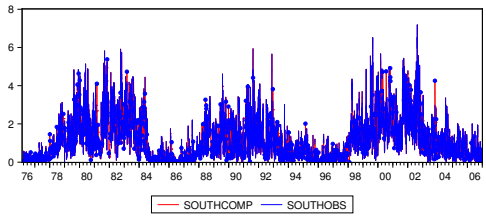


Fig. 1. Plage regions areas on the South hemisphere (in percent of the total area of the hemisphere): observed series (blue) and completed series (red).

The probabilistic equivalence between the observed and the completed series was confirmed considering descriptive summaries (histogram and numerical parameters), comparison tests of means, medians and variances, kernel density estimation, quantile-quantile chart and Chi-squared test. To illustrate these

studies, we present in Figures 2 and 3, respectively, the histograms and the estimated densities graphic representations obtained with the two sets of data. This density estimation is performed by the nonparametric kernel method (Silverman[9]) using the Epanechnikov kernel with bandwidth $h = 0.3401$.

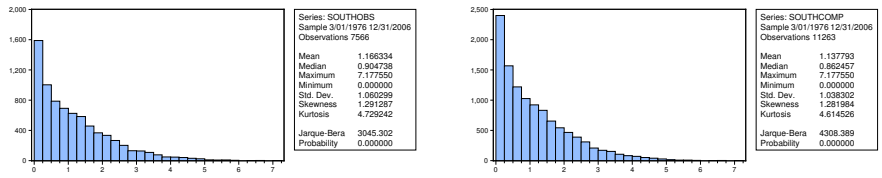


Fig. 2. Descriptive summaries of the observed and completed series of plage regions areas.

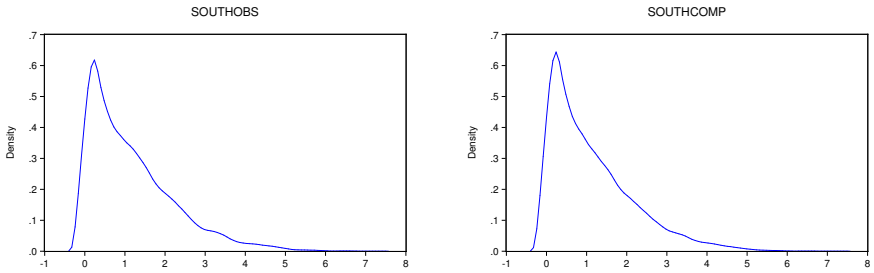


Fig. 3. Estimated densities for the observed and completed series of faculae regions areas.

Moreover, using the χ^2-test, the equality of the empirical distributions of the two series is accepted with p-value 0.32 as referred in Table 1.

To discard a possible loss of information by aggregation, we have repeated this analysis for the recorded and estimated observations, in each year, and similar conclusions were obtained. So, the random average methodology used to complete the series respects the original probabilistic structure.

2.3.2. Temporal evolution: south plage region areas

To characterize the temporal evolution of the series in study and according to the Box-Jenkins methodology, we begin by analyzing the autocorrelation and

	Observed N	Expected N	Residual
]0.0,0.5]	2591	2663.2	-72.2
]0.5,1.0]	1478	1505.6	-27.6
]1.0,1.5]	1208	1180.3	27.7
]1.5,2.0]	824	802.0	22.0
]2.0,3.0]	937	923.1	13.9
]3.0,4.0]	359	340.5	18.5
]4.0,6.0]	169	151.3	17.7
Total	7566		
Test statistics			
	Chi-squared		7.005
	Degrees of freedom		6
	p-value		0.320

Table 1. Chi-squared test of equality of the empirical distribution of the observed and completed series of plage regions areas.

partial autocorrelation functions of the series. From these functions, presented in Table 2, it is easy to conclude that this data is well fitted by an AR(2) model.

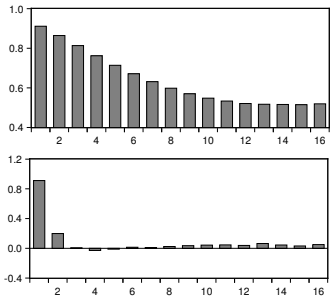
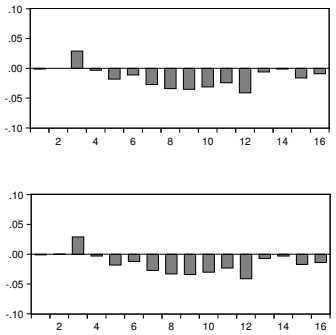


Table 2. Empirical autocorrelation and partial autocorrelation coefficients for the areas of plage regions series.

The estimation of this AR model leads to an heteroscedastic residual series presented in Table 3. In fact, applying the ARCH-LM test to this residual, the null hypothesis of homoscedasticity is rejected with p-value 0.0000. A generalization of the Box-Jenkins methodology must be used.

The SOUTHCOMP series is reanalyzed considering the class of AR(2) models with general δ-TGARCH(1) error processes. In fact, the correlogram and partial correlogram of the residuals squared exhibit a step one dependence (Table 4). The parameters of the model and the corresponding standard error produced by Eviews are presented in Table 5.



Heteroskedasticity Test ARCH			
F-statistic	996.9546	Prob. F(1,11258)	0.0000
Obs*R-squared	916.0139	Prob. Chi-Square(1)	0.0000
Test Equation			
Dependent Variable:RESID^2			
Method:Least Squares			
Sample(adjusted) 3/04/1976 12/31/2006			
Included observations: 11260 after adjustments			

Table 3. Empirical autocorrelations and partial autocorrelations of the residual series of AR(2) model estimation and Output of ARCH-LM test.

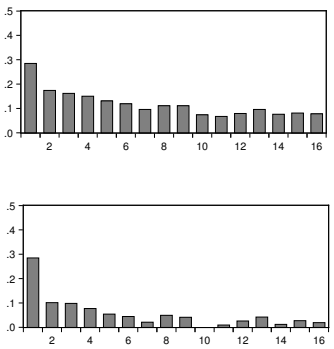


Table 4. Empirical autocorrelations and partial autocorrelations of the residual square after AR(2) estimation.

Denoting by $S = (S_t, t \in \mathbb{Z})$ the AR(2) process and $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ the corresponding error process, the evolution of South series is well fitted by the following AR(2)-δ-TARCH(1) model:

$$S_t = 0.211 + 0.721 S_{t-1} + 0.262 S_{t-2} + \varepsilon_t$$

where

$$\begin{cases} \varepsilon_t = \sigma_t Z_t \\ \sigma_t^{0.081} = 0.68 + 0.2932 (\varepsilon_{t-1}^{+})^{0.081} + 0.2811 (-\varepsilon_{t-1}^{-})^{0.081} \end{cases},$$

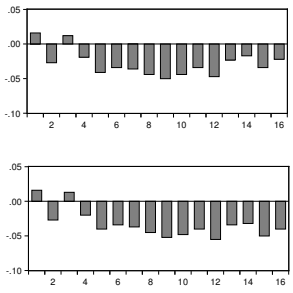
and $(Z_t, t \in \mathbb{Z})$ are independent real random variables with a centered and reduced Gaussian distribution.

According to Theorem 1, it is easy to establish the strictly stationarity and ergodicity of this process. The residual series associated has properties of homoscedastic error process. In fact, the residual correlogram is compatible

Dependent Variable: SOUTHCOMP				
Method: ML-ARCH (Marquardt)-Normal distribution				
Included observations: 11261 after adjustments				
Convergence achieved after 35 iterations				
Variance backcast: ON				
@SQRT(GARCH)^C(7)=C(4)+C(5)*(ABS(RESID(-1))-C(6)*(RESID(-1))^C(7)				
	Coefficient	Std error	z-statistics	Prob
C	0.211178	3.44E-08	6140565	0.0000
AR(1)	0.721457	5.00E-05	14442.70	0.0000
AR(2)	0.262416	6.83E-05	3540.143	0.0000
Variance Equation				
C(4)	0.680386	0.006421	105.9555	0.0000
C(5)	0.288426	0.004804	61.45183	0.0000
C(6)	-0.249003	0.011459	-21.7298	0.0000
C(7)	0.080675	0.006538	12.33854	0.0000
R-squared	0.834036	Mean dependent var		1.137987
Adjusted R-squared	0.833948	S.D. dependent var		1.038293
S.E. of regression	0.423099	Akaike info criterion		0.775714
Sum squared resid	2014.610	Schwarz criterion		0.783270
Log-likelihood	-4377.552	F-statistic		9426.001
Durbin-Watson stat	1.952839	Prob (F-statistic)		0.00000
Inverted AR Roots	0.99	-0.27		

Table 5. Estimates of model AR(2) with δ-TARCH errors.

with that of a white noise and the ARCH-LM test applied to that series accepts the null hypothesis of homoscedasticity with p-value 0.797887 (Table 6).



ARCH Test:			
F-statistic	0.065581	Prob F(1,11258)	0.797887
Obs*R-squared	0.065593	Prob. Chi-Square(1)	0.797876
Test Equation			
Dependent Variable: WGT_RESID^2			
Method: Least Squares			
Sample (adjusted): 3/04/1976 12/31/2006			
Included observations: 11260 after adjustments			

Table 6. Empirical autocorrelations and partial autocorrelations of the residual series of AR(2)-δ-TARCH(1) model estimation and Output of ARCH-LM test.

This model fitting leads us to the three trajectories present in Figure 4, namely the completed series (in red), the series estimated by the model mentioned above (in green) and the corresponding residues trajectory (in blue). We point out the fitting quality as the temporal model captures very well the evolutionary characteristics of the observed series.

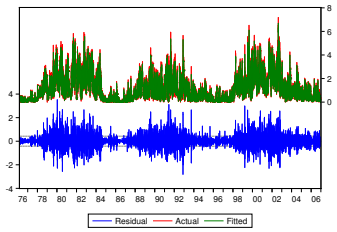


Fig. 4. The series of the plage regions areas, the fitted series and the residual trajectories.

Finally, we should stress that ARMA-TGARCH models with integer power do not capture well the heteroscedasticity of the residual series that must correspond to a long memory property in the shocks of the plage regions areas volatility (Ding, Granger and Engle[3]).

References

1. Berkes, I., L. Horváth, P. Kokoszka, "GARCH processes: structure and estimation", *Bernoulli* 9, 2, 201-227, 2003.
2. Bollerslev, T., "Generalized autoregressive conditional heteroskedasticity", *J. Econometrics* 31, 307-327, 1986.
3. Ding, Z., C.W. Granger, R.F. Engle, "A long memory property of stock market returns and a new model", *J. Empirical Finance* 1: 83-106, 1993.
4. Engle, R.F., " Autoregressive conditional heteroskedasticity with estimates of the variance of the UK inflation", *Econometrica* 50: 987-1008, 1982.
5. Gonçalves, E., J. Leite, N. Mendes-Lopes, "On the probabilistic structure of power threshold generalized ARCH stochastic processes", *Stat. Prob. Lett.* 82: 1597-1609, 2012.
6. Gonçalves, E., N. Mendes-Lopes, I. Dorotovic, J.M. Fernandes, A. Garcia, "North and South Hemispheric Solar Activity for Cycles 21-23: Asymmetry and Conditional Volatility of Plage Region Areas", *Solar Physics* 289: 6, 2283-2296, 2014.
7. Mittnik, S., M.S. Paolella, S.T. Rachev, "Stationarity of stable power-GARCH processes", *Journal of Econometrics* 106, 97-107, 2002.
8. Pan, J., H. Wang, H. Tong, "Estimation and tests for power-transformed and threshold GARCH models". *Journal of Econometrics* 142, 352-378, 2008.
9. Silverman, Density estimation for Statistics and Data Analysis, Chapman and Hall, London, 1998.
10. Zakoian, J.M., "Threshold heteroskedasticity models". *Journal of Economic Dynamics and Control* 18, 931-955, 1994.

Modeling errors in temperature forecasts

Rui Gonçalves[1]

LIAAD - INESC TEC
and Faculty of Engineering of the University of Porto, Portugal
(e-mail: rjasg@fe.up.pt)

Abstract. In this work we use data transformations to find a probability density function (pdf) to forecasting errors in daily maximum and minimum temperatures. This kind of data is not Gaussian and has features of the so called nearly Gaussian random variables (NG), see Lefebvre[6]. A NG random variable can be obtained by simply raising a Gaussian random variable to an exponent c, where $c = \{(2k + 1)/(2j + 1)\}$, $k, j = \{0, 1, \ldots\}$. To the pdf's obtained in this way from the Gaussian pdf we call the power normal pdf family. For the NG variables it is possible to relate the value of the kurtosis coefficient to the particular exponent of the transformation, c. We analyze the daily temperature forecast errors in the city of Porto during the year 2011. We compare the fit of the power normal model pdf to the fit of different non Gaussian models such as the Laplace and the Pearson type IV. We conclude that the power normal model gives the best fitting results with exponents close to those obtained by Lefebvre[6]. For the case of errors in minimum temperature forecasts we found that the data is already approximately Gaussian.
Keywords: Nearly Gaussian random variable, kurtosis, power transformation.

1 Introduction

Forecasting in temperature and precipitation is important to agriculture, to estimate the demand of certain goods on over coming days. On daily basis, people use weather forecasts to determine what clothe to wear, to plan outdoors activities and for the protection of life and goods. The problem of modeling forecasts errors of temperatures has been addressed in Lefebvre[6] and Wilks[8] among others. In temperature forecasts, common sense tell us that we should expect a rather symmetrical error distribution with small deviation. For instance, on a forecast of 15 degrees Celsius it is not expected to occur a temperature of 35 degrees Celsius. In this work we follow Lefebvre[6] that considered power transformations as a way to find a pdf for NG variables. The power transform gives rise to the power normal pdf's family that is useful to model NG variables. The data that we use in application of this theory is the daily forecast errors in maximum and minimum temperature in the city of Porto.[1] (one observation and respective forecast per day) and for the year 2011. On section (2) we present the nearly Gaussian random variable. We calculate the ordinary central moment and the kurtosis coefficient of a Gaussian variable X with zero mean raised to a power c. We

[1] The data was collected by Instituto Português do Mar e da Atmosfera (IPMA)

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

also describe a method to find the specific pdf of the power normal family that models data. In section (3) we apply the method described on section (2) to the data. Firstly, we consider the errors in maximum temperatures. Using the Lilliefors and the Shapiro-Wilk test we conclude that normality is rejected. Using the sample kurtosis and a table of the theoretical kurtosis of a power normal random variable we found the exponents 9/11 and 7/9 to be appropriated to transform the original data in a sample with Gaussian distribution. This means that the original data can be well fitted by a Gaussian distribution raised to the power 11/9 or 9/7. In the case of the one day ahead forecast of minimum temperature errors we found that normality is not rejected. In section (4), we compare the fit of the power normal distribution to other distributions such as the Laplace and the Pearson type IV and we compare the results of the Qui-square goodness-of-fit test to conclude that all pvalues observed are greater than the usual significance levels but the pvalue associated with the power normal is significantly greater than the others.

2 Power transformation and the power normal

In this section we will follow Lefebvre[6]. The pdf of a random variable resulting form the power transformation of a normal random variable is given in the following proposition.

Proposition 1. *If $Y = X^c$ is gaussian, $Y \sim N(\mu, \sigma^2)$, then the pdf of the power transformation of a Gaussian variable, $X = Y^{1/c}$ is,*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} c|x^{c-1}| \exp\left[-\frac{1}{2}\left(\frac{x^c - \mu}{\sigma}\right)^2\right].$$

This transformation is related to the Box-Cox transformation. We will call it the power normal pdf. The more interesting case for applications is when μ is zero and $c \in]0, 1[$ that is the range of c values for which Y is a nearly Gaussian random variable. An example pdf is given in figure (2). In order to identify the exponent c from experimental data we must relate c with some statistical measure. If X is a Gaussian random variable with parameters $\mu = 0$ and variance σ^2, then for $c > 0$

$$E(X^c) = \int_{-\infty}^{\infty} \frac{x^c}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{2^{c/2}}{2\sqrt{\pi}} \sigma^c [1 + (-1)^c] \Gamma\left(\frac{c}{2} + \frac{1}{2}\right),$$

where Γ is the gamma function. Hence, the following proposition (see Lefebvre[6]) may be stated

Proposition 2. *The kurtosis coefficient of the random variable $Y = X^c$ when $X \sim N(0, \sigma^2)$ is given by*

$$\beta_2(c) = \frac{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^{4c} e^{-x^2/2} dx}{\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^{2c} e^{-x^2/2} dx\right)^2} = \sqrt{\pi} \frac{\Gamma(2c + \frac{1}{2})}{\Gamma^2(c + \frac{1}{2})}.$$

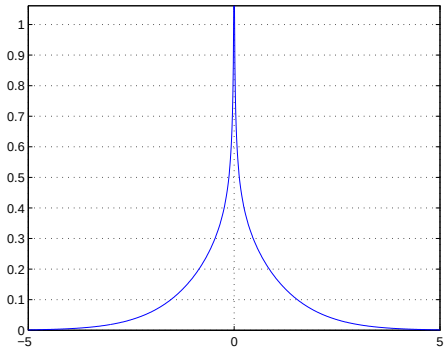


Fig. 1. Pdf of a Gaussian variable raised to the power 0.7.

c	$\beta_2(c)$	c	$\beta_2(c)$
7/9	2.237	17/15	3.58
9/11	2.358	15/13	3.08
11/13	2.447	11/15	2.11
13/15	2.514	19/21	2.643
15/17	2.566	23/25	2.697
17/19	2.6	29/31	2.753
13/11	3.828	31/33	2.767
23/19	3.979	21/23	2.672
11/9	4.042	35/37	2.79
9/7	4.404	37/39	2.8

Table 1. Kurtosis of the random variable $Y = X^c$ where $X \sim N(0,1)$ for a few values of c.

We present the table (1) that shows $\beta_2(c)$ for some values of c. This table will be used to select an exponent c, close to 1, so that, raising the values of the nearly Gaussian sample to $1/c$ one obtains an approximately Gaussian sample. Note that the identification of the exponent c requires that the variable should have mean and skewness close to zero. To estimate the power normal parameter for a given a sample $x_1, \ldots, x_n$ we center the data defining

$$z_i = x_i - \bar{x},$$

so that the mean of the z_i is 0. Assuming that the skewness coefficient is close to zero and the kurtosis is not close to 3 to be Gaussian then, by selecting an appropriate c from table (1), we find the transformation $W = Z^{1/c}$ that is likely to transform the data into a an approximately Gaussian.

3 Application to temperature forecasts

In this section we apply the method described in the last section. Our goal is to find statistical models for the forecasting errors of minimum and maximum temperatures. The size of our data set is 347 (there is a few missing data). We define X as $X = T_F - T_O$ where T_F is the forecast and T_O the observed temperatures. Firstly, we consider the one day ahead maximum temperature forecasts during the year 2011.

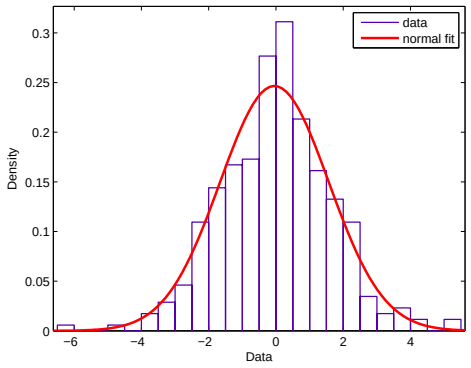


Fig. 2. Histogram and Gaussian fit to the errors X in maximum temperature forecasts.

The figure (2) is an histogram for the maximum temperature forecast errors data with the Gaussian fit on top. We note that there are more observations in the center when comparing to the Gaussian density. This feature is typical of the nearly Gaussian random variables. Hence, we will try to fit an appropriate member of the power normal family of distributions. But, before applying any statistical test we must make sure that there is none or very little temporal correlation among the data. A way of measuring temporal correlation is by computing the sample autocorrelation function (ACF). The data with small temporal correlation has an ACF within the 95% confidence bounds for white noise. Since we've found 3 out of 20 values outside the bounds then we decided to take a subset of the data. We kept only one in each pair of forecasting errors. This procedure reduces the data size to an half (173). Using the software SPSS we performed the Lilliefors and Shapiro-Wilk tests and we obtained pvalues of 0.023 and 0.1 respectively. The pvalue of Lilliefors test is less that the usual significance levels (0.05 to 0.1) used in statistical tests. So, we conclude that the Gaussian distribution is not a good model for X. Next, we tried classic models for the forecast errors such as the Laplacian and the Student's T and in both cases it turned out that the

models were not acceptable. Now, we turn our attention to the power normal family. To find the right transformation we must compute some statistics of the data first. Let us define the mean of X by

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n},$$

and

$$\hat{\mu}_k = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^k}{n-1},$$

for $k = 1, 2, ...,$ is the estimated k-th order central moment. The sample standard deviation $s_x = \sqrt{\hat{\mu}_2}$. The sample skewness coefficient is

$$b_1 = \frac{\hat{\mu}_3}{s_x^3},$$

and the sample kurtosis is

$$b_2 = \frac{\hat{\mu}_4}{s_x^4}.$$

The observed statistics for the error in the maximum temperature are:

$$\bar{x} = -0.0458;\ s_x = 1.616\ b_1 = 0.035;\ b_2 = 3.567.$$

The mean is close to zero so there is no need to center the data. The kurtosis of the reduced data set is 4.0583. Based on the kurtosis in table (1), we tried the transformation

$$w_k = x_k^{9/11}.$$

We found that the gaussian distribution is an acceptable model to the w_k's. Applying the Lilliefors test (with Matlab) the p-value increased from 0.023 before the transformation to at least 0.2 and the p-value of the Shapiro-Wilk p-value test is 0.439. Because the values of the sample kurtosis are not exactly equal to those of the table we also tried another exponent close to 9/11. In fact, trying

$$w_k = x_k^{7/9},$$

and applying the Lilliefors and Shapiro-Wilks tests (with SPSS) we obtain the p-values 0.2 and 0.439 which are equal to those obtained for the former exponent. Hence, we can use a Gaussian distribution raised to the power 11/9 to model the data. Raising the original data to the power 9/11 we obtain an approximately Gaussian distribution with mean -0.027 and standard deviation 1.311

$$X^{(9/11)} \approx N(-0.027, 1.311^2).$$

4 Fitting results

In this section, we will compare the power normal family fitting results to those of the symmetric Laplace and the Pearson IV distributions. Firstly, we consider the Laplace distribution

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right).$$

The estimate for μ is of course zero and the maximum likelihood estimate for $b = \sum_{i=1}^{n} |X_i - \hat{\mu}|$. Secondly, we considered the Pearson type IV distribution, its pdf is

$$f_X(x) = k\left[1 + \left(\frac{x-\lambda}{a}\right)^{-m}\right] \exp\left\{-\nu \arctan\left(\frac{x-\lambda}{a}\right)\right\},$$

for $x \in \mathbb{R}$ and the parameters m, $m > 1/2$, ν, a and λ are real constants. A normalizing constant k required, see Lefebvre([6]). Using the estimators given by the method of moments (see Stuart and Ord[7]) we obtained

$$\hat{m} = 5.5666, \hat{\nu} = -1.6843, \hat{a} = 4.2831, \hat{\lambda} = -0.8211.$$

The results of the chi-square tests are presented in table (2). We see that the

interval	n_j	X	Lap	Pearson IV
$(-\infty$, -2.5)	8	10.6870	9.4782	4.4115
(-2.5, -2)	11	7.01236	5.2715	7.9234
(-2, -1.5)	12	10.4451	8.20341	12.8885
(-1.5, -1,0)	11	14.8661	12.7660	18.4764
(-1, -0.5)	16	20.4931	19.8660	22.9571
(-0.5, 0)	31	32.2387	30.9149	24.566
(0,0.5)	31	25.1018	30.9149	22.7495
(0.5, 1)	21	17.5664	19.8659	18.4937
(1,1.5)	10	12.4670	12.7659	13.4767
(1.5,2)	9	8.50495	8.20342	8.996
(2, 2.5)	7	5.54012	5.27154	5.6398
(2.5, ∞)	6	7.92306	9.47825	3.3908
d^2		8.63832	14.0892	13.1724
p-value		0.47131	0.11919	0.15496

Table 2. Chi-Square goodness of fit test to determine whether a Gaussian $N(-0.027, 1.311^2)$ distribution raised to the power 11/9 is a good model for the raw data. The five columns give the chosen subintervals, the number n_j of observations in each subinterval and the expected e_j number of observations for each subinterval and for the Gauss, Laplace and Pearson IV distributions in this order.

p-value of the observed statistic for the power transformation of the normal

is considerably better than the others. Next, we turn our attention to the forecasts errors of minimum temperatures. Like in the case of the maximum temperature when computing the ACF there are 2 values outside the 95% confidence bounds therefore we decided to eliminate one value in each two. The observed statistics for the error in the minimum temperature for the reduced data set are:

$$\bar{x} = 0.099;\ s_x = 1.52\ b_1 = -0.04;\ b_2 = 2.67.$$

Applying the Lilliefors and the Shapiro-Wilk tests (again with SPSS) we found a pvalue of at least 0.2 for the Lilliefors test and a pvalue of 0.439 for the Shapiro-Wilk test. This means that the data is already approximately Gaussian and there is no need for transformations.

5 Concluding remarks

In this paper, we show that a using an appropriate exponent of the form $(2k+1)/(2j+1)$, $k, j = 0, 1, ...$ the power transformation of a nearly Gaussian random variable can be Gaussian. The transformation is bijective so it may be used in both positive and negative data. We applied the method described in section (2) to data consisting of the one day ahead forecast errors in daily maximum and minimum temperatures. In the case of errors in maximum temperatures we used both Lilliefors and Shapiro-Wilk tests and we concluded that normality was rejected. Afterwards, using the sample kurtosis and the table (1) we found the appropriate exponents 9/11 and 7/9 that transform the original data in a sample with Gaussian distribution. Fitting results of the power normal, Laplace and Pearson IV distributions were compared and the pvalue of the power normal was found to be significantly greater than the other two. Surprisingly, in the case of the one day ahead forecast for daily minimum temperature errors, normality was not rejected.

6 Acknowledgments

This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project "FCOMP-01-0124-FEDER-037281".

References

1. Box, G.E.P and Cox.D. R. "An analysis of transformations", *J. Roy. Statist. Soc. Ser. B*, 26, pp. 211-243, 1964.

2. Gonçalves, R., Pinto, A. and Stollenwerk, N. "Cycles and universality in sunspot numbers fluctuations", *The Astrophysical Journal*, 691, pp.1583-1586, 2009.
3. Gonçalves, R., Ferreira, H.,Pinto, A. and Stollenwerk, N. " Universality in nonlinear prediction of complex systems" *Journal of Difference Equations and Applications*, 15, 11-12, pp.1067-1076, 2009.
4. Gonçalves, R., Ferreira, H. and Pinto, A. " Universality in the stock exchange market", *Journal of Difference Equations and Applications*, 17,7 pp. 1049-1063, 2010.
5. Gonçalves, R., Ferreira, H., Stollenwerk, N. and Pinto, A. " Universal fluctuations of the AEX index" *Physica A: Statistical Mechanics and its Applications*, 389,21, pp. 4776-4784, 2010.
6. Lefebvre, M. " Nearly Gaussian Distributions and Application" *Communications is Statistics - Theory and Methods*, 39, pp 823-836, 2010.
7. Stuart, A. and Ord, K. *Kendall's Advanced Theory of Statistics*, vol 1, 6th, ed. Oxford University Press, 2010.
8. Wilks, D. *Statistical Methods in the atmospheric Sciences* vol. 59, Academic Press, 1995.

Quality control of GNSS-Receivers by accuracy-based analysis

Federico Grasso[1], Michal Hodoň[2], Jana Púchyová[2] and Eckehard Schnieder[1]

[1] Institute for Traffic Safety and Automation Engineering, Technische Universität Braunschweig, Germany.
(Email: grasso@iva.ing.tu-bs.de)

[2] Faculty of Management Science and Informatics. University of Žilina, Slovakia.
(Email: jana.puchyova@fri.uniza.sk)

Abstract. In this paper we present a quality control method for Global Navigation Satellite System (GNSS) receivers. A statistical quality control (SQC) approach for accuracy is proposed, focused on quantitative trueness, precision and location availability analysis of GNSS receivers', based on an independent reference system. The location availability is described as the percentage of the total received data that can be considered precise under n-σ boundaries; being n the level of requested precision. As part of this accuracy-based location availability analysis several filter techniques are tested, in order to select the most reliable for this specific quality control method. A traditional SQC method is compared with Mahalanobis Ellipses Filter (MEF) method, while both are provided by particle filter (PF) position estimation, as the independent reference. The quality control methods are depicted in graphical representation. And the results are analysed from an end-user point of view. Finally a detailed description of the receiver's characteristics and conditions of the measurements are presented as part of a case study. Significant differences between the presented approaches are shown and a quality-oriented assessment is proposed.

Keywords: Quality control, GNSS receivers, Mahalanobis Ellipses Filter, Particle Filter.

1 Introduction

The quality control for GNSS receivers is an important feature to all GNSS-based applications. In this paper we describe the development of a quality control methodology by means of accuracy analysis. Accuracy is described by quantitative values of trueness and precision of the GNSS dataset; while the location availability is described as the percentage of the total received data that can be considered precise under n-σ boundaries; where n is the level of requested precision. Two approaches are presented. First a traditional statistical quality control (SQC) trial, by means of quality control chart (QCC) of the module deviation analysis and easting-northing (E/N) bivariate deviation analysis. Then a new Mahalanobis Ellipses Filter (MEF) approach is tested; by means of Mahalanobis distance evaluation trial of the deviation dataset.

All three trials of both approaches and their correspondent results are used for outlier detection as the proposed quality control methodology. Also a reference based on Particle Filter (PF) is developed and tested for both approaches.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

Fig. 1: GNSS quality description.

A. GNSS Receivers quality control

More and more GNSS-based applications are available for localisation purposes. But no quality control methodology for the GNSS receivers has been developed. One step to the future certification of GNSS receivers is to develop a multiple receivers' quality methodology, focusing on the user side.

In Fig. 1 it can be seen how both software and hardware quality control analysis are possibilities from the system point of view. However from the user point of view, a methodology focused on accuracy and precision will be more representative of the quality of the receiver, as presented in Hodon [1].

Previous studies of quality by means of these mentioned characteristics can be found in Hodon [1] and Grasso et al. [2]. A more detailed method is presented in Grasso et al. [3], providing the theoretical basis for the presented accuracy-based quality control methodology.

GNSS receivers' accuracy, described in Grasso et al. [3], is presented in Fig. 2. Based on the True Score Theory model from Trochim [4], accuracy by means of trueness and precision of the GNSS receiver is based on the deviation analysis:

$$\boldsymbol{Location = Reference + Error}$$

From where deviation is defined as the error term and it can also be divided into two significant components:

$$\boldsymbol{error = deterministic\ error + stochastic\ error}$$

Where deterministic error is the intrinsic error of the receiver's behaviour and stochastic error (or non-deterministic error) is randomly added error to the receiver's behaviour. Location availability is then defined as "the percentage of the GNSS data provided by the system that is considered precise, after filtering with an n-σ from a defined precision threshold":

$$LocAv_{N\sigma} = \frac{N\sigma\ filtered\ number\ of\ samples}{Total\ number\ of\ samples} * 100\%$$

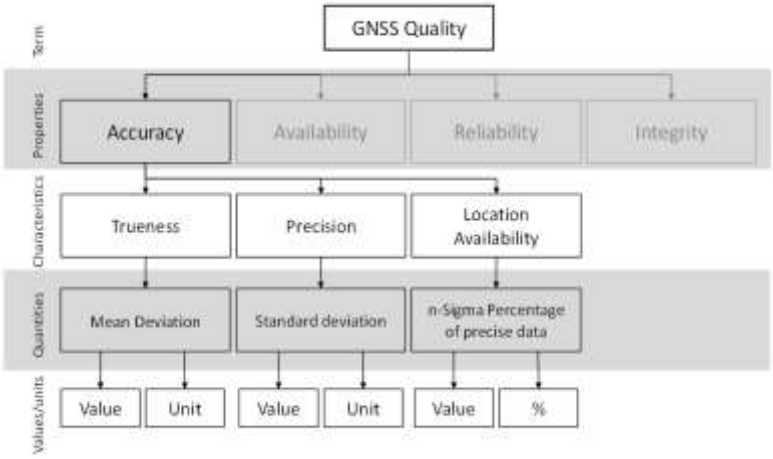


Fig. 2: GNSS Quality attributes hierarchy, focused on accuracy.

B. Particle Filter approach for reference estimation

Particle filter (PF) is a kind of probabilistic suboptimal nonparametric filters, whose main idea is the implementation of sequential Monte Carlo estimation, using particle representation of the probability density function (pdf), as described in Doucet et al. [5] and Arulampalam et al. [6]. Advantages such as the possibility of using PF also for nonlinear systems and the ability of PF to filter any error probability distribution make them not only limited to normal Gaussian probability distribution errors. This is why this filter can be well suited for the problematic of target localization, as seen in Púchyová [7] [8], and GNSS receiver error filtration. Our aim is to estimate the position of the receiver with some error following discrete-time stochastic model:

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}, \boldsymbol{v}_k)$$

where *f* is the known function of the state x_{k-1} and v_k is process noise sequence. The measurements have relationship with the state of the receiver through measure equation:

$$\boldsymbol{z}_k = h(\boldsymbol{x}_k, \boldsymbol{u}_k),$$

where *h* is known function and u_k is the process noise sequence.
Noise sequence v_k and u_k are independent. The filtrated estimation x_k based on the sequence of all the available measurements $\boldsymbol{Z}_k \triangleq \{\boldsymbol{z}_i, i = 1,\ldots,k\}$ up to time k is searched, so it is necessary to construct the posterior pdf p(xk|Zk). Then in principle, pdf p(xk|Zk) can be reached recursively in three steps: prediction, update and resampling, involving update of pdf prediction with Bayesian rule:

$$p(\boldsymbol{x}_k|\boldsymbol{Z}_k) = p(\boldsymbol{x}_k|\boldsymbol{z}_k, \boldsymbol{Z}_{k-1}) = \frac{p(\boldsymbol{z}_k|\boldsymbol{x}_k, \boldsymbol{Z}_{k-1})p(\boldsymbol{x}_k|\boldsymbol{Z}_{k-1})}{p(\boldsymbol{z}_k|\boldsymbol{Z}_{k-1})} = \frac{p(\boldsymbol{z}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{Z}_{k-1})}{p(\boldsymbol{z}_k|\boldsymbol{Z}_{k-1})}$$

For our filtration, a Sampling Importance Resampling (SIR) algorithm from Gordon et al. [9] was used, where the new particles are estimated from the prior $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ so this step is independent from the measurement vector Z_k.
The main advantage of this kind of filter is that the significant weights are easily accessed and therefore the pdf can be easily sampled.
The measurement z_k is used when weights of each particle are set:

$$w_k^i \propto p(\boldsymbol{z}_k|\boldsymbol{x}_k^i)$$

where index *i* expresses the *i*−th particle in the PF. The resampling phase is executed on each time step k.

C. Particle Filter receiver's behaviour estimator

The PF in the present paper aims to the position estimation of the receiver. In order to provide the PF a probability distribution, the measurements from the first day dataset were computed to find the deviation distribution of the receiver.
For both PF-Easting Estimator and PF-Northing Estimator the fitting values were a normal distribution with the mean value and sigma (σ) values describing the easting and northing behaviour of the receiver.
These values resulted from the deviation between the actual reference point from the antenna and the receiver's output from the first day dataset.
Based on Arulampalam et al. [6] a two part PF-based estimator was developed to be used in combination with the two quality control proposed approaches, as a reference frame for the receiver's behaviour. These two filters estimate easting and northing positions of the evaluated receiver.
As seen in Fig. 3 the inputs and outputs are:
Inputs: 1) Deviation distribution of the location of the antenna: this is calculated from the actual location of the antenna and the actual deviation value. Mean value and standard deviation are used for the deviation distribution. 2) Position data: the value for the position provided by the receiver.
Output: 1) Estimated reference: estimation based on provided deviation distribution. The used reference for the further quality control methodologies is the composition of both filters' outputs, and it will be referred as PF-Estimator for the rest of the paper.
Fig. 4 presents a short example of the PF adaptive period, showing the filter evolution of the Gauß-Krüger Northing part of the PF-Estimator.

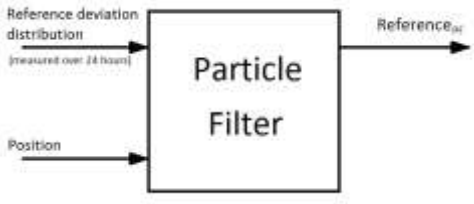


Fig. 3: Input-output diagram of PF.

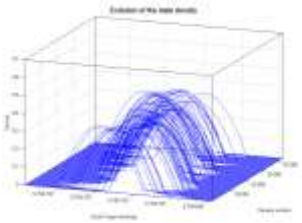


Fig. 4: Evolution.

Fig. 5 displays the improvement (in percentage of $LocAv_{1\sigma}$ added) for the MEF quality control methodology, with the PF-Estimator for different number of particles. Due to computational complexity of PF is affected by rising number of particles; a compromise relation had to be found with the improvement of the data for quality control approaches. For this purpose the selected number of particles for each component of the PF-Estimator was 400.
The MEF quality control methodology results and its improvements related to the usage of the PF-Estimator will be explained in detail in the section three of the present paper. Mathematically the PF-Estimator reference results in:

$$\boldsymbol{Reference_{PF} = Reference_{REAL} + deterministic\ error}$$

In Fig. 6 the resulting 777600 samples (i.e. nine days of estimated reference) used for the quality control methodologies comparison are presented. These are estimated by the PF-Estimator, after the one-day (19.05.2011) adaptation period composed by 86400 samples.
Using the PF-Estimator reference deviation can be studied as the direct function of the stochastic error of the receiver.
This is the selected characteristic for determining the quality of the receiver by means of uncertainty measurement.
The location provided by the receiver can be defined as:

$$\boldsymbol{Location = Reference_{PF} + stochastic\ error}$$

In section three the proposed quality control methodologies, based on SQC and MEF approaches, are compared using deviation calculated with and without the developed PF-Estimator.

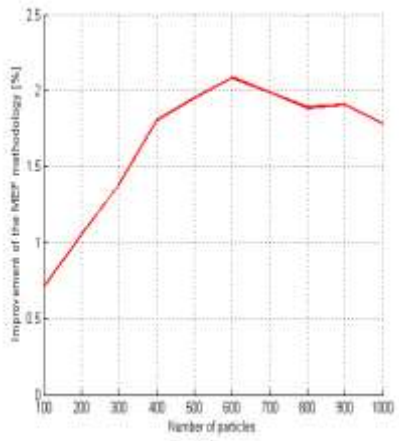


Fig. 5: MEF improvement.

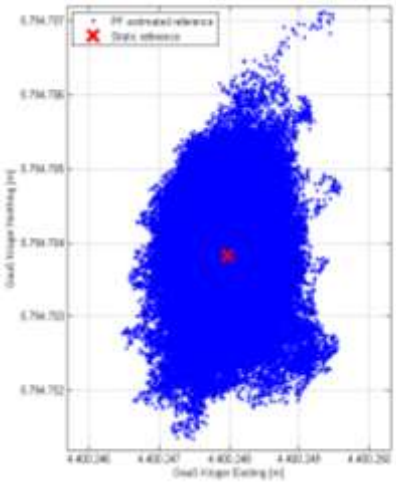


Fig. 6: Reference estimated by PF.

D. Statistical Approach

Statistical quality control (SQC) is a set of statistical tools used by quality professionals that can be divided into three categories according to Reid [10]:
1) Descriptive statistics; 2) Statistical process control (SPC); and 3) Acceptance sampling.
For the GNSS receiver accuracy-based quality control methodology developed in the present paper the proposed SQC approach focuses on descriptive statistics (description of quality characteristics and relationships by statistics measurements such as the mean, standard deviation and distribution of data) of the totality of the receivers' output dataset for further outliers' detection by means of deviation analysis. Two different trials are tested within this approach, focused on two separated calculations of the deviation. The first calculates the deviation by means of the Euclidean distance, and it's called module deviation calculation. And the second calculates the deviation in two separated components, called the easting northing (E/N) deviations calculation.

E. MEF approach

In order to achieve a meaningful description of the quality of a localisation system a new approach has been developed, in the frame of the extended accuracy-based evaluation developed in Grasso et al. [11] Mahalanobis Ellipses Filter (MEF) is a filtering technique that allows a better understanding of the nature of the deviation datasets, and therefore a better ground for GNSS data validation, based on Mahalanobis [12].
MEF methodology focuses on the finding of outliers from the deviation dataset while describing the behaviour of the system (composed by the GNSS-receiver and the reference system) by means of the resulting Mahalanobis ellipses.
MEFs provide not only a description of the bivariate (easting and northing) deviation, but also the resulting rotated ellipses describe the correlated behaviour of the deviation dataset. In Grasso et al. [11] it is proposed that quality control methods as well as validation procedures for certification of GNSS receivers can be performed by MEFs. In this paper the comparison between this new filter approach and the traditional SQC approach is conducted by means of outlier detection in the deviation resulting dataset. Results of this comparison and conclusions are presented in section three of the presented paper.

2 Dataset description

This section is focused on a short description of the used datasets for the quality control methodologies comparison, and their correspondent statistical analysis.

A. Measurement description

Using a receiver u-blox EVU-6H with EGNOS turn on, assembled with the antenna Novatel GPS-702-GG the location was determined geodetically on the roof of the Institut für Verkehrssicherheit und Automatisierungstechnik.

Fig. 7: Installed equipment and Google view of the reference.

Fig. 7 presents the picture of the installed equipment and the Google Maps reference. Measurements of 10 days were collected with a 1 Hz frequency in the period between 19.05.2011 and 01.06.2011, resulting in 864000 positions.

B. Statistical analysis of collected datasets

All 10 datasets present similar characteristics: Each one has 86400 samples over a period of 24 hours, with a number of visible satellites between 7 and 12 (average of 10) and a Horizontal Dilution of Precision (HDOP) value between 0.69 and 1.51 (average of 0.91). According to Ming [13], these characteristics describe the scenario as ideal. Table I shows the statistical analysis for the 10 datasets deviation analysis, referred to the PF-Estimator reference. The tenth dataset is a global dataset composed from the other 9 datasets. The first day (19.05.2011) dataset used for the PF adaptation period is not considered for the rest of the analysis.
In section three these distribution fittings will be used to calculate the limits for the QCC, as part of the SQC approach. Fig. 8 presents the fittings for the deviation between the PF-Estimator reference and the position measurements from the global dataset. Fig. 8A displays a lognormal distribution for the module deviation, while Fig. 8B and 8C display normal distributions for the E/N deviations, independently. This fitting evaluation presented in Fig. 8 has already been used as the basis for reliability margins definition in Grasso et al. [2].

3 Quality control process

This section focuses on the results for both SQC and MEF approaches. Both quality control methodologies are tested by three separated trials, using both the actual location of the receiver and the developed PF-Estimator, in order to conclude on its usefulness.

A. SQC results

The results presented here are provided by both the real reference and the PF-Estimator reference regarding the deviation global dataset.

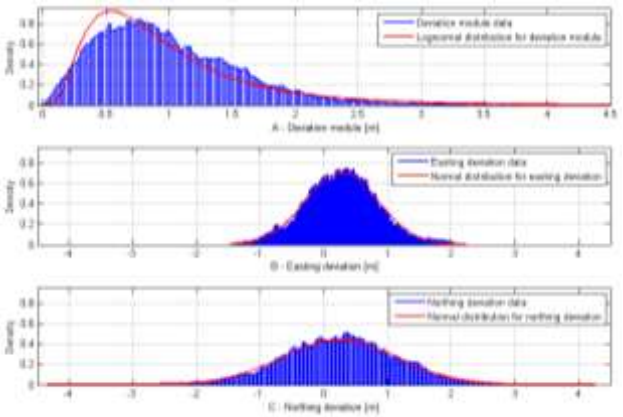


Fig. 8: Fitted distributions for global deviation analysis.

The first trial is focused using the SQC methodology with module deviation, while the second trial uses the E/N deviation.

B. SQC methodology with module deviation analysis

The control limits of the QCC for the module deviation analysis are calculated considering the lognormal distribution, as shown in set of equations (1).

$$\begin{aligned} UCL_M &= e^{\mu_M + nSigmas * \sigma_M} \\ CL_M &= e^{\mu_M} \\ LCL_M &= 0 \end{aligned} \tag{1}$$

The fitting parameters presented in Table I for the global dataset are used and the control limits are calculated for 1σ result in:

$$LCL_M = 0,\ CL_M = 0.8218,\ UCL_M = 1.5782$$

Fig. 9 shows a QCC in the left side for the module deviation analysis, based on the lognormal distribution fitting. UCL_M separates accepted and rejected samples. Also Fig. 9 presents a in the right side a scatter-plot with the detailed marked limits. Using the limits from Table I from the lognormal distribution and the set of equations (1), the resulting location availability for 1-σ for the module deviation analysis is:

$$LocAv_{1\sigma M} = \frac{SQC\ module(1\sigma)\ filtered\ number\ of\ samples}{Total\ number\ of\ samples} * 100\%$$

The result without PF-Estimator is: $LocAv_{1\sigma M} = 86.0739\%$.
The result with PF-Estimator is: $LocAv_{1\sigma M} = 85.7486\%$.
This presents a decrease of the accuracy of the 1-σ filter of 0.3253 %.
The lognormal limits for QCC show that module deviation analysis is too permissive for the analysed dataset. And also the PF-Estimator proves that the estimated reference behaviour in the SQC approach spreads the acceptance threshold even outside the data cloud.

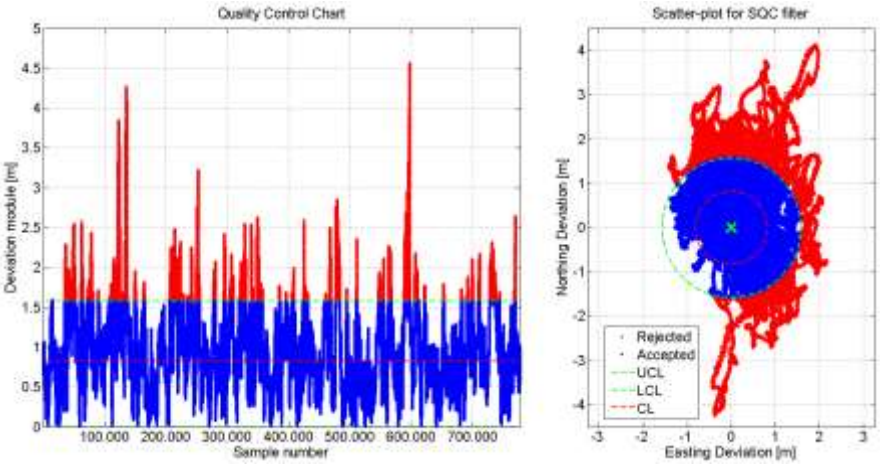


Fig. 9: QCC and Scatter-plot for 1σ SQC filter based deviation module.

C. SQC methodology with easting-northing deviation analysis

The control limits of the QCC for the E/N deviation analysis are calculated considering the normal distribution, as shown in set of equations (2).

$$\begin{aligned} UCL_E &= \mu_E + nSigmas * \sigma_E \\ CL_E &= \mu_E \\ LCL_E &= \mu_E - nSigmas * \sigma_E \\ UCL_N &= \mu_N + nSigmas * \sigma_N \\ CL_N &= \mu_N \\ LCL_N &= \mu_N - nSigmas * \sigma_N \end{aligned} \qquad (2)$$

The fitting parameters presented in Table I for the global dataset are used and the control limits are calculated for 1σ resulting in:

LCL_E = -0.2842, CL_E = 0.2831, UCL_E = 0.8503
LCL_N = -0.6555, CL_N = 0.2533, UCL_N = 1.1620

Fig. 10 shows in the left side two QCCs for the E/N deviations, based on the normal distribution fittings. Also Fig. 10 presents in the right side a scatter-plot with the detailed marked limits.
Using the limits from Table I from the normal distribution and the set of equations (2), the resulting location availability value for 1-σ for the E/N deviation analysis is:

$$LocAv_{1\sigma EN} = \frac{SQC\ E/N(1\sigma)\ filtered\ number\ of\ samples}{Total\ number\ of\ samples} * 100\%$$

The result without PF-Estimator is: $LocAv_{1\sigma EN}$ = 49.4520%.
The result with PF-Estimator is: $LocAv_{1\sigma EN}$ = 50.9816%.
This presents an improvement of the accuracy of the 1-σ filter of 1.5296 %.

TABLE I: STATISTIC ANALYSIS OF THE DATASET

Dataset date	Trueness Module [cm]	Deviation Module Lognormal Fitting		Easting Deviation Normal Fitting		Northing Deviation Normal Fitting	
		μ_M	σ_M	μ_E [m]	σ_E [m]	μ_N [m]	σ_N [m]
23.05	51.1417	-0.1591	0.6491	0.3124	0.5140	0.4049	0.8691
24.05	37.4525	-0.0512	0.6771	0.0763	0.5310	0.3667	1.1880
25.05	44.5379	-0.1277	0.6587	0.1331	0.6841	0.4250	0.8676
26.05	51.8165	-0.1668	0.5992	0.2969	0.5940	0.4247	0.7660
28.05	41.5153	-0.2046	0.6285	0.1621	0.4599	0.3822	0.8786
29.05	30.8001	-0.2979	0.6666	0.3069	0.4852	0.0256	0.9118
30.05	31.3785	-0.2955	0.7471	0.2811	0.5390	0.1394	1.0128
31.05	33.0559	-0.3721	0.5855	0.2459	0.4700	0.2210	0.6875
01.06	74.1256	-0.0919	0.5743	0.7331	0.5342	-0.1099	0.7181
Global	37.9860	-0.1963	0.6526	0.2831	0.5673	0.2533	0.9088

E/N deviation analysis proves to be better than the module deviation analysis, due to the consideration of the deviation on each component of the horizontal position plane independently. Also the increase of the percentage of the $LocAv_{1\sigma EN}$ with PF-Estimator results from the reduction of stochastic error in the deviation calculation, making the quality control analysis based only on the stochastic error of the receiver.

D. MEF results

Since the Mahalanobis distance measures the number of sigmas that separate all samples from the rest of the group, the MEF 1-σ filter provides the number of samples within 1-σ of Mahalanobis distance related to the group.
Therefore, the location availability for 1-σ with the MEF Methodology is:

$$LocAv_{1\sigma MEF} = \frac{MEF(1\sigma)\ filtered\ number\ of\ samples}{Total\ number\ of\ samples} * 100\%$$

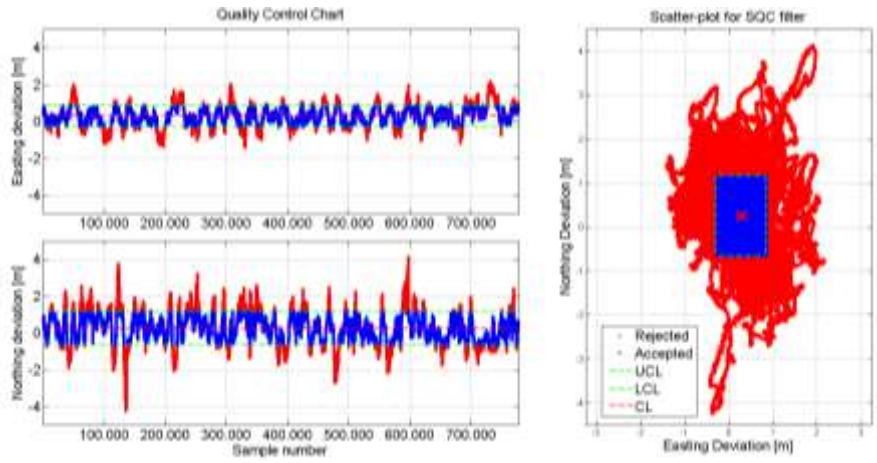


Fig.10: QCC and Scatter-plot for 1σ SQC filter based on easting and northing.

TABLE II: MEF DEVIATION ANALYSIS

Ellipses Center [m]	**Easting**	0.2533
	Northing	0.2831
Semi-radii [m]	**A**	0.529
	b	0.9316
Rotation	**[deg]**	15.497
LocAv 1σ	**with PF**	**44.0213**
LocAv 1σ	**without PF**	**42.2172**

Fig. 11 shows the Mahalanobis distance for each sample in the left side. The 1σ distance is marked by the UCL, separating the rejected samples from the accepted.

Fig. 11 presents as well a scatter-plot in the right side with the detailed marked limits. The deviation mean value is marked in Fig. 11 as the intersection between semi axes. Also the rotation of those axes with respect to the coordinate system describes the degree of correlation between E/N deviations. These characteristics are numerically presented in Table II.

The Mahalanobis distance is a descriptive statistic that provides a scale independent measurement of the distance of each sample with respect to the entire dataset, normalised with respect to σ. Therefore UCL_{MEF} for 1σ corresponds to a unitary Mahalanobis distance.

The Euclidean distance that is used in the SQC module deviation analysis and presented in Fig. 9, is not a descriptive statistic and is not scale-invariant. It measures the distance of independent position samples with respect to the reference.

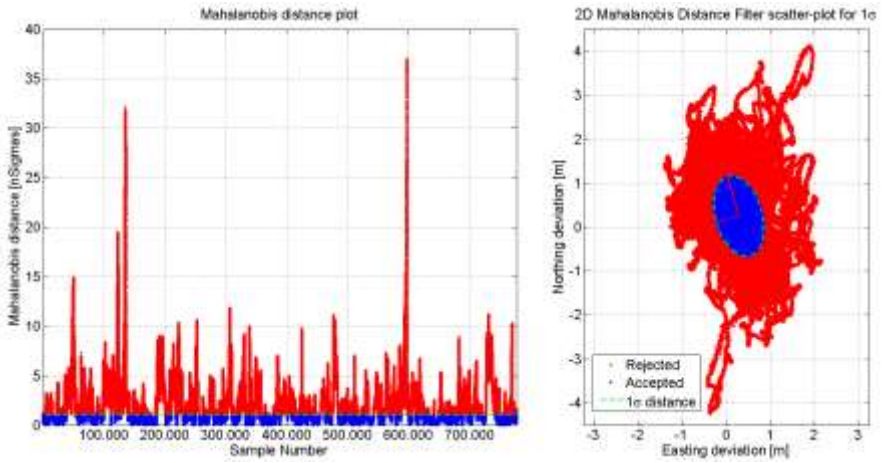


Fig. 11: Mahalanobis distance plot and Scatter-plot for 1σ MEF filter.

TABLE III: SUMMARY OF QUALITY CONTROL METHODOLOGIES

SQC Module	**LocAv 1S (with PF)**	***85.7486***
	LocAv 1S (without PF)	86.0739
SQC E/N	**LocAv 1S (with PF)**	***50.9816***
	LocAv 1S (without PF)	49.452
MEF	**LocAv 1S (with PF)**	***44.0213***
	LocAv 1S (without PF)	42.2172

This distance is fitted to a lognormal distribution in order to perform a statistical analysis of the samples, and the UCL_M for 1σ depends on the parameters of the fitted distribution.
Table II shows the characteristics of the produced Mahalanobis Ellipse of 1-σ and also the location availability for 1-σ with and without PF-Estimator. The increase of the 1.8041% with PF-Estimator results from the reduction of stochastic error in the deviation calculation.

E. Comparison of results

Table III and Fig. 12 present the results of the comparison between the three analysed quality control trials for both SQC and MEF methodologies.
SQC Module, SQC E/N and MEF results are presented by percentages values of the 1-σ outlier detection function with and without PF-Estimator as reference. The SQC Module filter test considers only the module of the deviation. It does not discriminate the deviation direction of the location samples, resulting in a high $LocAv_{1\sigma}$ that does not describe accurately the behaviour of the receiver. The SQC E/N filter test considers an independent evaluation of easting and northing deviations. It discriminates between two main deviation direction components, resulting in a lower $LocAv_{1\sigma}$ that represents better the receiver's behaviour when low correlation between easting and northing deviation is present.
Finally the MEF test performs a simultaneous evaluation of easting and northing deviations, taking into account the deviation correlation. The MEF approach also works considering all possible deviation directions by means of the normalised Mahalanobis distance; resulting in a lower $LocAv_{1\sigma}$ that represents better than the SQC E/N filter the receivers' behaviour.

Further work and conclusions

As seen in Fig. 12, the SQC approach is separated in two trials: SQC module trial, testing the trueness and precision, regarding the receiver's deviation; and SQC E/N trial that solves the trueness problem while improves the precision, although it is still not able to discriminate borderline outliers.
On the other hand the MEF approach proves to be the best filter for the receiver's deviation datasets, according to the receiver's behaviour and coinciding with the theorised ellipse, described in Kaplan and Hegarty [14].

The presented $LocAv_{1\sigma}$ comparison of outliers' detection for GNSS-receiver quality control proves that the MEF methodology is better than SQC methodology for the three following reasons: 1) Even though the SQC E/N test was a better description than SQC module, only MEF approach has a simultaneous evaluation of easting and northing deviations, while considering their correlation. 2) Also the MEF approach is the only one from the tested approaches considering all possible deviation directions by means of the normalised Mahalanobis distance. 3) The MEF methodology follows the elliptic behaviour predicted theoretically for accuracy evaluation. In the accuracy metrics section of Kaplan and Hegarty [14] the probability of a measurement to be in the 1-σ ellipse is defined as 39 %; while the probability of being in the 2-σ ellipse is 86 %. These theoretical values are calculated in contrast to the one-dimensional Gaussian result of the probability of being within ±1-σ of the mean value being 68 %.

Based on these theoretical values proposed by Kaplan and Hegarty [14] the MEF methodology is proven to be a sufficient representation of the receiver's behaviour and enough for validating the quality of the receiver tested in the present paper (both with and without the developed PF-Estimator).

Also it has been proven that the usage of a PF-Estimator as estimated reference is effective for stochastic error reduction of the receiver.

The results with 400 particles for the presented static case, in the 1-σ MEF case proves to improve up to 1.8041 % the inclusiveness of the location.

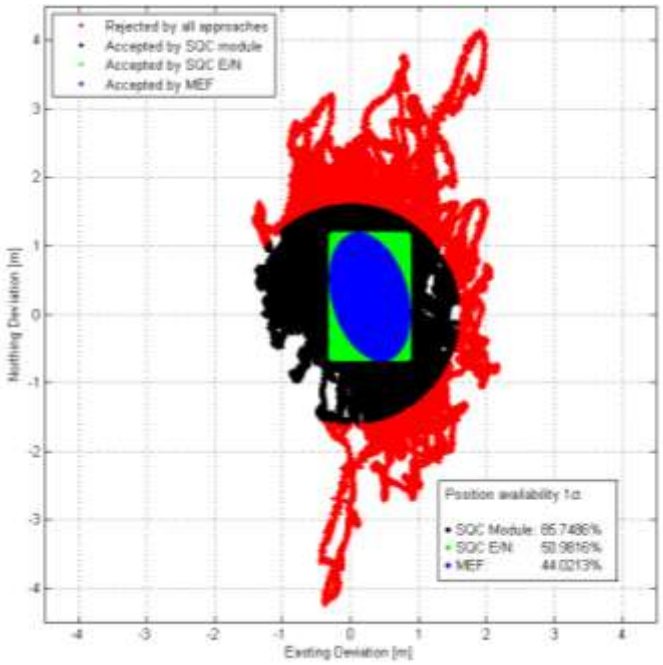


Fig. 12: Comparison between all approaches with 1σ filters.

It is suggested as further work to test the PF-Estimator for dynamic cases. In dynamic scenarios the actual reference value must be calculated from an on-board independent reference.
The developed PF-Estimator approach will be a necessary part of quality control processes, as well as further validation and certification processes for static and dynamic reference system. Finally, as a general conclusion of the present paper, the combination of the MEF methodology and the developed PF-Estimator approach seems to be the best representation of the behaviour of the GNSS-Receiver, and therefore the best base for its quality description.

Acknowledgments

The authors wish to thank QualiSaR EU Project Nr. 287187, StandOrt Project BMWI Nr. 01FS12046 and D.A.A.D. for the provided support.

References

1. M. Hodon: “GNSS receiver quality investigation estimated from the implemented standards analysis.” (2012).
2. F. Grasso Toro et al.: “Accuracy evaluation of GNSS for a precise vehicle control.” IFAC-CTS (2012)
3. F. Grasso Toro et al.: “Extended accuracy evaluation of GNSS for dynamic localisation in railways.”(2013).
4. W. Trochim: “Research Methods Knowledge Base 3e“. Cornell. (2004)
5. A. Doucet, et al.: “Sequential Monte CarloMethods In Practice.” New York: Springer-Verlag, (2001)
6. M.S. Arulampalam et al.: “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174-188. (2002)
7. J. Púchyová: “Minimizing of Target Localization Error using Multi-robot System and Particle Filters.” ICDIPC 2013 International Conference on Digital Information Processing, Turkey. World Academy of Science, Engineering and Technology. ISSN 2010-376X. - Iss. 78, s. 856-860. (2013)
8. J. Púchyová: “Behaviour of multiagent system with defined goal.” Information Sciences and Technologies: bulletin of the ACM Slovakia. ISSN 1338-1237-Vol. 5, no. 4, s. 15-25. (2013)
9. N.J. Gordon, D.J. Salmond and A. Smith: "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." Radar and Signal Processing, IEE Proceedings. Vol.140, no.2, pp.107-113. (1993)
10. R. D. Reid, N. R. Sanders: “Operations Management, 5th Edition” (2012).
11. F. Grasso Toro et al.: “New filter by means of Mahalanobis distance for accuracy evaluation of GNSS.” (2013).
12. P.C. Mahalanobis: On the generalised distance in statistics. (1936)
13. F. Y. Ming: DILUTION OF PRECISION CALCULATION FOR MISSION PLANNING PURPOSES, NAVAL POSTGRADUATE SCHOOL. (2009)
14. D. Kaplan and Christopher J. Hegarty: Understanding GPS. Principles and applications, 2nd Aufl., Artech Verlag, Boston, op. Page 328-332. (2006)

Employment and Fertility – A Comparison of the Family Survey 2000 and the Pairfam Panel

Andreas Groll[1] and Jasmin Abedieh[2]

[1] Ludwig-Maximilians-University
Department of Mathematics, Munich, Germany
(E-mail: groll@math.lmu.de)
[2] Ludwig-Maximilians-University
Institute of Statistics, Munich, Germany
(E-mail: jasmin.abedieh@hotmail.de)

Abstract. The major objective of this work is the analysis of the relationship of employment and fertility in Germany, also regarding causality. Based on Germany's current panel analysis of intimate relationships and family dynamics (pairfam), Cox's proportional hazards model is used to investigate the influence of labor force participation of women on the transition into motherhood. The obtained results serve as validation of an earlier study presented in Schröder and Brüderl [25], where the effect of employment on the fertility is analyzed for women based on the data of the West-German Family Survey 2000, using a proportional hazards model with a piecewise constant baseline hazard. In general, the estimated effects for the Cox model based on the pairfam data are surprisingly consistent with the results from Schröder and Brüderl [25], whereas indirect causality test results disagree.
Keywords: Pairfam, Employment, Fertility, Event data analysis, Cox's proportional hazards model.

1 Introduction

Today, there exist already several empirical studies in the literature, which clearly indicate that there is evidence for an influence of female labor force participation on the fertility. In this context, Schröder and Brüderl [25] mention several works which use event data analysis for different western industrial nations to show that employed women have a lower transition rate for delivering a (further) child than non-working women, see e.g. Felmlee [11] and Budig [6] for the US or Liefbroer and Corijn [23] for Flanders and the Netherlands. Apart from a few studies such as Kohlmann and Kopp [17], Kreyenfeld [18], Dornseiff and Sackmann [10], Lauer and Weber [20] or Kreyenfeld [19], which partly have a different analytical focus or exhibit some methodical problems, the work of Schröder and Brüderl [25] is the first study that explicitly analyzes if and to what extent there is a relationship between the labor force participation of women and their fertility in Germany, based on the West-German Family

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

Survey 2000. This study is replicated and validated here for the territory of the reunified Germany based on Germany's current panel analysis of intimate relationships and family dynamics (pairfam), release 4.0 (Nauck et al. [24]). A detailed description of the study can be found in Huinink et al. [16]. So the main focus in this work is the analysis of the influence of labor force participation of women on the transition into motherhood. Besides, like Schröder and Brüderl [25] we also investigate the causality of a possible (negative) effect of employment on the fertility, using a similar indirect causality test as proposed there. Note that our analyses are also restricted on transitions of childless women into motherhood, i.e. women delivering their first child.

The rest of the article is structured as follows. The most important sociological theories concerning employment and fertility are shortly summarized in Section 2. In Section 3 we discuss some theoretical aspects concerning the causality of a potential negative effect of female labor force participation on the fertility and propose a suitable indirect causality test. The data, the used methods and the results are presented in Section 4, before we finally conclude in Section 5.

2 Sociological theories concerning employment and fertility

Though lively discussed in media and social sciences, according to Schröder and Brüderl [25] only few theoretical approaches concerning the explicit mechanisms of employment and fertility exist. Schröder and Brüderl [25] provide a compact summary of the existing sociological theories and hypothesis in this context. Among the most important and relevant theories are the following two:

The hypothesis of incompatibility of roles: the roles of a woman as mother on the one hand and as employee on the other hand are generally incompatible, as simultaneous childcare and labor force participation would either reduce the productivity of the job performance or the quality of childcare.

The hypothesis of substitution: both of these roles are linked with certain rewards or incentives of e.g. emotional, social or financial kind; furthermore, the rewards that go along with one role can partially be substituted by those of the other role.

However, the gist of both theories does not directly explain why labor force participation thus necessarily has a negative effect on the fertility, because according to Schröder and Brüderl [25] employed women could simply give up their role as employee by the time they want to have children. At this point another theory has to be mentioned, which plays a major role in this context.

The economic theory of fertility: this theory is embedded in the well-known rational choice framework for understanding and modeling social and economic behavior. Here, the main idea is that couples are regarded as consumers, who take their decision with regard to the number of children they want to have after an extensive cost-benefit-assessment. Among the most famous protectionists of the economic theory of fertility are Leibenstein and Becker.

Leibenstein [21] considers children to implicate three different types of benefit: a consume benefit, as children are a general enrichment for parents, bringing affection and personal gratification to them; an income benefit, arising from the productive activities of the children; and finally, an insurance benefit, as children care and assist their parents in their old ages. At the same time, children cause direct (food, clothes, education etc.) and indirect costs (the raising of children goes along with a huge expenditure of time, strongly limiting the engagement of the parents in other activities) for their parents. While nowadays the last two types of benefit became more or less obsolete, at least in western industrial nations, where child labor is illegal since many decades and the requirement of insurance is transferred as far as possible to responsible institutions (compare Huinink and Konietzka [15]), the consume benefit has remained rather consistent and can already be achieved by a small number of children, according to Leibenstein [21]. At the same time, with increasing economic wealth, the costs of children have generally increased. Consequently, by the theory of Leibenstein the number of children is decreasing with increasing economic wealth.

A similar approach is presented in Becker [3], where children are regarded as consumer products, offering psychological benefit to their parents. Both the quantity and quality of children are included, the quality of children covering several characteristics such as education, health or future income. For Becker quantity and quality of children can (at least partly) be substituted, creating an incentive for parents to invest into the quality of their children, i.e. to spend more efforts on care and education, rather than to realize a higher number of children. On the other hand, similar to Leibenstein's indirect costs of children, Becker associates the costs that arise by the time spent for children. The idea is that child education is highly time-consuming and hence competes with other activities, e.g. employment. The time used for education could instead be used for employment, and the corresponding loss of earnings generates the so-called opportunity costs. This aspect is especially relevant for an employed woman. As soon as she stops working, even if only temporarily, her opportunity costs increase. Besides, the higher the wage rate the higher the opportunity costs (see Huinink and Konietzka [15]). Finally, as for Becker quantity and quality of children are more or less exchangeable, an employed women can realize her psychological benefit by investing in the quality of a child instead of deciding to get another child. Accordingly, with more and more women being employed and increasing income levels, also Becker's theory indicates a general decline in the number of children in developed nations. For more information about the economic theory of fertility, see also Hotz et al. [14]. A useful introduction and summary regarding important highlights of the attempts to develop an "economic" theory of human fertility are found in Leibenstein [22].

Several models exist, which consider the connection between the decision of women with respect to labor force participation and a demand for children, see e.g. Willis [28]. In most of these models the decisions relating to fertility and time allocation depend on basic economic variables such as man's income and woman's wage rate. As in these models the decisions relating to the number of

children and to the time that a woman spends for labor force participation are usually ultimately determined at the beginning of the marriage, these models are called *static life time models*.

As pointed out by Schröder and Brüderl [25], so-called *dynamic life cycle models* are more realistic, where the whole life time is divided into periods and then for each period the time is determined that a woman spends for child education and employment (or leisure time, depending on the model) together with the corresponding fertility decision. The major assumption in these models is that the previous employment history and the current work effort have an influence on the income. Consequently, employed women are able to achieve higher wage rates than non-working women and hence, these models expect a causal negative effect of employment on the fertility.

As already stated in the introduction, in fact several studies exist that confirm this hypothesis. In particular, Schröder and Brüderl [25] have found that practically all studies that base on event data analysis and investigate the influence of female labor force participation and fertility in western industrial nations have found such a negative effect, which is independent from the country and from the parity of the child. Apparently, the existing empirical studies confirm the theoretical considerations presented in this section. However, in spite of the results of existing studies, following Schröder and Brüderl [25] one has to be careful when making statements with regard to causality of this negative effect and a more sophisticated analysis seems necessary, see next section.

3 Causality

In this section we discuss some theoretical aspects concerning the causality of the negative effect of female labor force participation on the fertility. According to Schröder and Brüderl [25], one cannot directly conclude from the results of the presented existing studies that the effect is causal, i.e. the reason for the probability of birth being lower for employed women than for non-working women is in fact their labor force participation. If so, reversely, this would require that the decisions related to the labor force participation are made independent from the fertility decisions. But Schröder and Brüderl [25] point out that it is also conceivable that fertility decisions may have an influence on the labor force participation. Some studies have tried to account for this problem by considering suitable control indicators for the fertility and employment intentions, see e.g. Budig [6] or Cramer [9], but unfortunately the operationalization of these variables is quite imprecise. However, in most analyses the fertility intentions are not controlled at all. Otherwise, the results of two studies for Sweden (Hoem and Hoem [13]) and Great-Britain (Wright et al. [29]) indicate that fertility decisions also influence the labor force participation. Hence, Schröder and Brüderl [25] conclude that the relationship between employment and labor force participation is in fact quite complex. In this context they also graphically illustrate how, beside the employment status, also attitudes, moral concepts and long-term plans on the one hand, but

also opportunities and restrictions on the other hand could have effects on the fertility.

But for the present analysis the relationship between fertility decisions and the preceding employment status is of most interest. In this context, one problem is that the exact time of a fertility decision cannot be observed and usually birth is used as a simple indicator. Hence, neither the influence of the preceding employment status on the fertility decision nor a possible influence of a fertility decision on the subsequent employment period can be analyzed in a reasonable way. For this reason, Schröder and Brüderl [25] also mention that it is possible that the effect of the current labor force participation on the fertility, to which most of the studies mentioned in Section 1 refer, in fact is an effect of the anticipated fertility on the employment status. Furthermore, they point out that for an optimal analysis of the influence of the employment status on the fertility a data set would be required, which contains the fertility intentions as a time-dependent covariate with the same temporal preciseness as the employment variable. For this purpose a panel with rather short interview intervals would be required. Unfortunately, such data are currently not available, neither for our analysis nor in Schröder and Brüderl [25].

Another important aspect in this context is the problem of so-called *unobserved heterogeneity*, also known as self-selection or spurious correlation. Even if the fertility intentions could be observed at any time and an effect of the preceding employment status on the fertility would be discovered, statements concerning the causality of this effect can only be made, if one can control for all factors which may have an influence on both the employment status and the fertility decision. If instead some of these factors are unobservable, then the relationship between fertility and labor force participation is (at least partly) a spurious correlation, i.e. non-working women would possibly be more likely to get children than employed women anyway (also without a causal effect of the employment status on the fertility), simply because they differ with respect to some unobserved factors relevant for the fertility decision. Hence, the effect of the employment status on the fertility would (at least partly) reflect this unobserved[1], compare Schröder and Brüderl [25].

Regarding these theoretical considerations, a major task is now to find a suitable method, which allows to empirically test the causality of the employment effect. Ideally, panel data containing the fertility intentions as a time-dependent covariate with the same temporal preciseness as the employment variable would be available, but as mentioned above such data are not (yet) on hand. Hence, Schröder and Brüderl [25] propose two indirect[2] causality tests.

[1]Possible candidates for such unobserved factors are the family, employment and career orientation or the fertility intentions. In this context Schröder and Brüderl [25] mention several research results, which indicate that such unobserved factors might be relevant. For example, Stolzenberg and Waite [26] found a negative relationship between (long-term) fertility intentions and employment plans and Cramer [9] and Budig [6] show that fertility intentions actually have an effect on the fertility.

[2]Schröder and Brüderl [25] call these tests *indirect*, because they base on additional assumptions, which cannot be checked on the basis of their data. Nevertheless, the

The first test assumes that women have different family orientations and can be divided into different (observable) groups according to their family orientation. It analyzes the progress of the effect of employment on the fertility over the cohorts and is based on the assumption that the differences with regard to family orientation between employed on non-working women have increased over the cohorts[3]. However, in the following analysis we abstain from performing this test for two reasons. First, pairfam's youngest cohort covers people born in the years 1971-1973, so even women from the youngest cohort already benefit from modern opportunities and working time organization models increasing the compatibility of family and work, such as e.g. public financial support, part-time work, trust-based working etc., when they reach their reproductive age. Second, in total pairfam contains only three different cohorts and people from the third cohort (1991-1993) are still in their teens at the time of the third interview wave (2010/2011). So, our data basis contains basically women belonging to only two different cohorts and hence, the corresponding indirect causality test would not be very meaningful.

With their second indirect causality test Schröder and Brüderl [25] want to check if the effect of the current employment status on the fertility in fact results from a reverse effect of an anticipated fertility decision on the employment status. The idea is that if some women would determine their employment status due to a preceding fertility decision, then one could expect among the group of women, who change from employment to unemployment and vice versa, a high percentage of such women. For this reason women are divided into the following four different groups: (a) mainly employed, (b) mainly non-working, (c) changers from employment to unemployment and (d) changers from unemployment to employment. For women belonging to group (c) one would expect very high transition rates for the transition into motherhood, while on the contrary for women belonging to group (d), very low transition rates are expected. Finally, for the other two groups (a) and (b) one would expect moderate transition rates lying in between. If instead only the current employment status causally affects the transition rate into motherhood, one would expect that the transition rate of currently employed women is much lower than the one of currently non-working women, independent of the former employment history. Following Schröder and Brüderl [25], we hope that if we regard a survival model with a single categorical covariate for these four groups, this allows us some conclusions about the causality of the effect of employment on the fertility or whether the effect in fact results from a reverse effect of an

tests are quite transparent, compare e.g. Brüderl et al. [5] or Beck and Hartmann [2] for similar test applications.

[3]The idea behind this assumption is that while in the 1950s and 1960s the bigger part of the female population was extensively restricting their labor force participation when getting their children, nowadays women have many possibilities and alternatives to combine their professional career with their family life, with the consequence that today only women with a very strong child-orientation are supposed to decide themselves against labor force participation. Hence, an increasing effect of employment over the cohorts would indicate self-selection as described in Section 3.

anticipated fertility decision on the employment status. We present the results of the corresponding indirect causality test in Section 4.

4 Data, methods and results

In this section we first illustrate the data and shortly comment on operationalization. Furthermore, we explain the used methods and finally present the results.

4.1 Data

Germany's current panel analysis of intimate relationships and family dynamics (pairfam, release 4.0; Nauck et al. [24]), started in 2008 and contains about 12,000 randomly chosen respondents, belonging to the birth cohorts 1971-73, 1981-83 and 1991-93. Pairfam follows the cohort approach, i.e. the main focus is on anchor persons of certain birth cohorts, who provide in yearly conducted interviews detailed information, orientations and attitudes (mainly concerning the family situation) of themselves and their partners. A detailed description of the study is found in Huinink et al. [16].

Here, for a subsample of 2,289 women the retention time (in days) until the birth of the first child is considered as the dependent variable, starting at their 14th birthdays. In order to ensure that the independent time-varying covariates are temporally preceding the events, the duration until conception (and not birth) is considered, i.e. the time of event is determined by subtracting 7.5 months from the date of birth, which is when women usually notice pregnancy. For each woman the employment status is given as a time-varying categorical covariate with eight categories, compare Table 3. Note that due to gaps in the women's employment histories a category called "no info" is introduced. As in the study of Schröder and Brüderl [25], for women who belong to this category for longer than 24 months it is set to "unemployed". Besides, several other time-varying and time-constant control variables are considered. Tables 2-4 give an overview of all considered variables together with their proportions in the sample. An extraction of the data set is shown in Table 1.

Id	start	stop	birth	employment	education level	relationship status	cohort	# siblings	education level of parents
111000	0	730	0	school	apprenticeship	single	1	1	traineeship
111000	730	1434	0	no info	apprenticeship	single	1	1	traineeship
111000	1434	1891	0	no info	apprenticeship	cohab	1	1	traineeship
111000	1891	1939	1	full-time	apprenticeship	cohab	1	1	traineeship
907000	0	365	0	school	secondary educ.	single	2	0	traineeship
907000	365	2438	0	no info	secondary educ.	single	2	0	traineeship
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Structure of the data

For the indirect causality test we extract a second, smaller data set, called *event.data.test*, with the employment status as the only covariate of interest. Observations in the categories "school", "education" or "no info" are dropped. As in Schröder and Brüderl [25], we construct the time-varying covariate *employ.test* with four categories: (a) mainly employed, (b) mainly non-working,

(c) changers from employment to unemployment, (d) changers from unemployment to employment. Each category is computed proportionally on the preceding intervals (threshold: $> 50\%$) and also accounts for the current employment status. E.g., if a woman has been employed for more than 50 % of her employment biography and is currently unemployed, then she is currently in status (c).

One can observe that most of the variables have similar proportions compared to the West-German Family Survey 2000 , with the major difference that for the variable *employment status* we found higher proportions in the categories "school" and "no info" and consequently lower proportions in the categories "full-time employed", "part-time employed" and "education", see Table 3.

4.2 Methods

In the following we use a semi-parametric approach, which is suitable for the estimation of the influence of specific covariates on the survival time of certain statistical objects. The most common class of models used in the literature is the class of hazard rate models, in particular the so-called proportional hazards rate (PH-)model. This model belongs to the class of semi-parametric regression models, as for the baseline hazard function no specific form needs to be assumed.

	proportion
Birth cohort	
1971-1973	0.49
1981-1983	0.41
1991-1993	0.10
# siblings	
no siblings	0.20
one sibling	0.44
two siblings	0.21
three or more siblings	0.14
Education level of parents	
university with PhD	0.015
university without PhD	0.095
A levels	0.003
college of higher education	0.138
apprenticeship	0.103
traineeship	0.440
general secondary education	0.005
secondary education	0.024
no graduation	0.007
other graduation	0.001
no info	0.169
Number of women	2,289
Number of events	1,371

Table 2: Distribution of the time-constant covariates in the sample

	# days	proportion
Employment status		
full-time employed	3,089,174	0.274
self-employed	85,560	0.007
part-time employed	252,396	0.022
marginally employed	107,087	0.009
education	165,165	0.015
school	2,634,246	0.233
unempl./job-seeking/housewife	216,639	0.019
no info	4,737,190	0.420
Education level		
university with PhD	483,529	0.043
university without PhD	1,669,741	0.148
A levels	396,253	0.035
college of higher education	1,764,788	0.156
apprenticeship	2,226,048	0.197
traineeship	4,004,395	0.355
general secondary education	298,837	0.026
secondary education	299,438	0.027
no graduation	45,206	0.004
no info	99,222	0.009
Relationship status		
single	5,471,726	0.485
partner	3,310,963	0.293
cohabitation	1,904,906	0.169
married	599,862	0.053
Number of women	2,289	
Number of events	1,371	
Number of days	11,287,457	

Table 3: Distribution of the time-varying covariates in the sample

	# days	proportion
Combination employment history/ current employment status		
continuously unemployed	150,340	0.040 (0.013)
change from employed to unemployed	66,299	0.018 (0.006)
change from unemployed to employed	85,717	0.023 (0.008)
continuously employed	3,448,500	0.919 (0.306)
Number of women	1,705	
Number of events	863	
Number of days	3,750,856	

Table 4: Distribution of the four groups that are considered in the indirect causality test; in brackets: proportion with respect to the main data set

The influence of explanatory variables is modeled parametrically, assuming that these covariates directly influence an individual's hazard rate. The hazard

rate has the following well-known form:

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}^t \boldsymbol{\beta}) = \lambda_0(t) \exp(x_1 \beta_1) \cdot \ldots \cdot \exp(x_p \beta_p),$$

with baseline-hazard $\lambda_0(t)$ and linear predictor $\mathbf{x}^t \boldsymbol{\beta}$ (usually containing no intercept β_0, as it is already covered by $\lambda_0(t)$). The hazard rate is defined as follows:

$$\lambda(t, \mathbf{x}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t, \mathbf{x})}{\Delta t},$$

representing the instantaneous risk of a transition at time t (here: a transition into motherhood), given that the transition did not yet occur. Characteristic property is the proportionality of the hazard rates: for two arbitrary individuals with corresponding vectors of covariates $\mathbf{x}_i, \mathbf{x}_j$ we get

$$\frac{\lambda(t, \mathbf{x}_i)}{\lambda(t, \mathbf{x}_j)} = \frac{\lambda_0(t) \exp(\mathbf{x}_i^t \boldsymbol{\beta})}{\lambda_0(t) \exp(\mathbf{x}_j^t \boldsymbol{\beta})} = \exp((\mathbf{x}_i - \mathbf{x}_j)^t \boldsymbol{\beta}),$$

i.e. the proportion of the hazard rates of woman i and j at time t is not depending on time, but solely on their covariate realizations; major objective is the estimation of the covariate effects $\boldsymbol{\beta}$.

4.3 Results

In the following we consider two rather similar PH-models, the famous Cox-model (Cox [7]) and the so-called piece-wise constant (PWC-)model (e.g. Blossfeld et al. [4]). In the PWC-model the basic assumption is that the baseline hazard can change on predefined intervals, but remains constant within these intervals. In contrast, the Cox-model uses the so-called Nelson-Aalen estimator (Aalen [1]) for the baseline hazard. The corresponding cumulative baseline hazard functions are illustrated in Figure 1, showing that the PWC cumulative hazard is coarser, but has the same general course as the Cox estimate. Exemplarily, a Cox model incorporating all covariates from Section 4.1 can be fitted in R using the package `survival` (Therneau and Grambsch [27]) by the call:

```
>cox.obj <- coxph(Surv(start,stop,birth) ~ employment + education
+ relationship+ siblings + edu.parents + cohort + cluster(id),
data=event.data, method="breslow"),
```

presuming that all categorical covariates are already transformed into factors[4]. Similarly, a PWC-model can be fitted using the `phreg` function from the R package `eha`. Figure 1 also shows the effect of the *employment status* on the cumulative baseline hazard functions for both approaches: women, who are still at school (blue), have the lowest transition rate into motherhood, whereas women in the reference category (represented by the baseline hazard; black), i.e. who are unemployed, job-seeking or housewives have the highest transition rate. As the Cox estimates are smoother, exhibit no big jumps and hence

[4]The `cluster(id)` term in the formula implies that robust variance estimators are used. The `method` argument specifies the method for tie handling.

more adequately model the data structure, in the following we focus on the Cox model when comparing our results with those obtained in Schröder and Brüderl [25].

Figure 2 shows the estimated fixed effects and 95%-confidence intervals corresponding to the German Family Survey 2000 (Schröder and Brüderl [25]; dashed lines) and the pairfam data (solid lines). As not all covariates exhibit

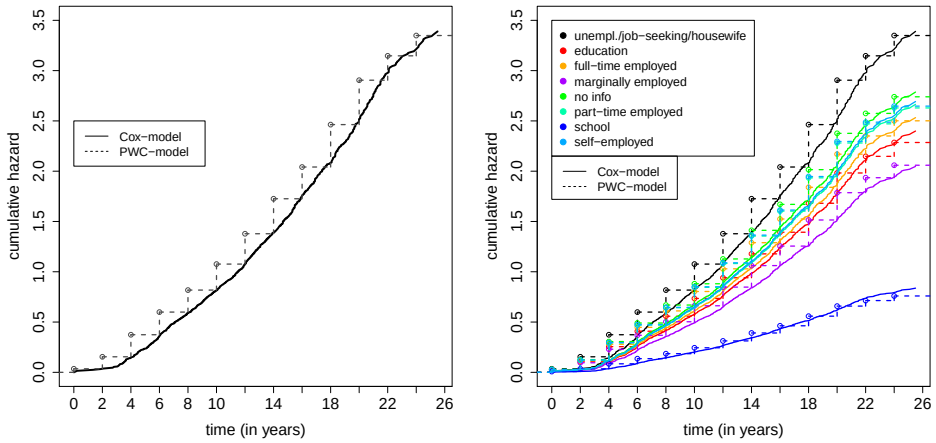


Fig. 1: Left: comparison of the cumulative baseline hazard functions, PWC- and Cox-model; right: effect of the *employment status* on the cumulative baseline hazard functions, PWC- and Cox-model

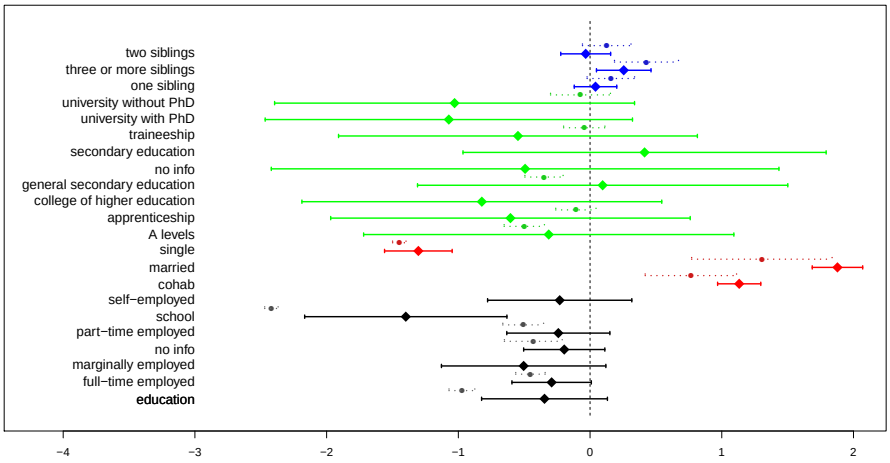


Fig. 2: Comparison of the fixed effects corresponding to the German Family Survey 2000 data (Schröder and Brüderl [25]; dashed lines) and the pairfam data (solid lines)

exactly the same categories for both studies, only the effects of those covariates are shown where a comparison is (at least approximately) possible. Note that

the effects of the *parents' education level* are not shown here, as in the pairfam study it is measured in more detailed levels compared to the German family survey. First, it turns out that the estimated effects for the Cox model based on pairfam are surprisingly consistent with those from Schröder and Brüderl [25]. Second, standard errors and confidence intervals are larger for the pairfam data, which is partly due to the used special variance-robustness method. All estimated (exponential) regression coefficients together with standard errors are presented in Table 6 in the Appendix.

In detail, we get the following results. Similar to Schröder and Brüderl [25], we find a strong negative, significant effect when women still go to school. Besides, the categories "part-time employed" and especially "full-time employed" have negative effects on the transition into motherhood compared to unemployed women, the first effect being close to significance and the latter being significant. Hence, our results confirm a negative effect of female labor force participation on the fertility for whole Germany. Later, we focus on the investigation of the causality of this effect.

With respect to the other control variables we find that the degree of institutionalization of the relationship shows the expected effects: married women have the highest transition rate into motherhood, followed by (unmarried) women who live together with their partner and women who live (apart) together with a partner; single women have the lowest transition rates. While the birth cohort has no influence on the hazard rate, women who grow up with many siblings have significantly higher transition rates. Besides, it is seen that in comparison to the reference category "no graduation" higher educational levels, except for the two types of secondary education, have negative effects, with similar trends as in Schröder and Brüderl [25], though without being significant. Similar tendencies, but with significance, are observed for the *parents' level of education* see Table 6. Next, we consider several goodness-of-fit criteria for the fitted model.

Goodness-of-fit

First, we check the proportional hazards (PH-)assumption for the hazard function. Grambsch and Therneau [12] propose a test on the validity of the PH-assumption against the alternative of time-varying coefficients. While Table 6 in the Appendix shows that the global test rejects the PH-assumption, also tests for single covariates should be considered, in particular those corresponding to key variables. A closer examination of the single tests shows that for the variables *education level* and *relationship status* the PH-assumption is generally violated ($\alpha = 0.05$), as for at least one category the null hypothesis is significantly rejected. In contrast, for the variables *employment status*, *cohort*, *number of siblings* and *parents' education level* the PH-assumption is not rejected.

The model's overall performance can be graphically assessed by investigating the Cox-Snell residuals (Cox and Snell [8]), i.e. by comparing empirical and theoretical cumulative hazard functions of the residuals. If the true underlying model is close to the specified one, the estimated cumulative hazard rate of the Cox-Snell residuals is close to the bisecting line, which is generally fulfilled here, see Figure 3 in the Appendix. Besides, similar to the residuals of an or-

dinary least-squares-estimator in linear regression, the Cox-deviance residuals can be regarded, separately for each covariate. They should vary symmetrically around zero and are also suitable to detect outliers. Figure 4 in the Appendix shows the Cox-deviance residuals, exemplarily for the covariates *employment status* and *relationship status*, which manifest a slight negative trend, i.e. survival times are slightly over-estimated by the model. Consequently, some model assumptions might be violated. Nevertheless, all in all the fitted model seems appropriate and provides an adequate fit.

Indirect causality test
To check if the effect of the current employment status on the fertility in fact results from a reverse effect of an anticipated fertility decision on the employment status, we fit the following model:

```
>cox.obj2 <- coxph(Surv(start,stop,birth) ~ employ.test + cluster(id),
 data=event.data.test, method="breslow"),
```

which is based on the smaller data set *event.data.test* and on the constructed time-varying covariate *employ.test*, introduced in Section 4.1. Even though the fitted effects in Table 5 show the same trend as in Schröder and Brüderl [25], they are far from significance. Hence, our test does not directly indicate that the estimated negative effect of female labor force participation is not causal.

	$exp(\beta)_{SB}$	$exp(\beta)_{pairfam}$
Combination employment history/ current employment status		
continuously unemployed	1	1
change from employment to unemployment	$1.822***$	1.014
changers from unemployment to employment	$0.449*$	0.653
continuously employed	0.862	0.776
individuals	2,093	1,705
number of events	1,447	863

Table 5: Comparison of the indirect causality test results for the German Family Survey 2000 data (Schröder and Brüderl [25]; $exp(\beta)_{SB}$) and the pairfam data ($exp(\beta)_{pairfam}$)

5 Conclusion

In this work the relationship of employment and fertility in reunified Germany is analyzed on basis of the pairfam data, also regarding causality. We find that the estimated effects for a Cox proportional hazards model based on the pairfam data are surprisingly consistent with the results of an earlier study from Schröder and Brüderl [25], which is based on the West-German Family Survey 2000. However, a corresponding indirect causality test cannot confirm the opposite direction, namely that self-selection in terms of anticipated fertility decisions also affects employment. We conclude that with respect to causality a more sophisticated analysis seems necessary.

Acknowledgment: This article uses data from the German family panel pairfam, coordinated by Josef Brüderl, Johannes Huinink, Bernhard Nauck, and Sabine Walper. Pairfam is funded as long-term project by the German Research Foundation (DFG).

Appendix: Estimation, goodness-of-fit and PH-test results

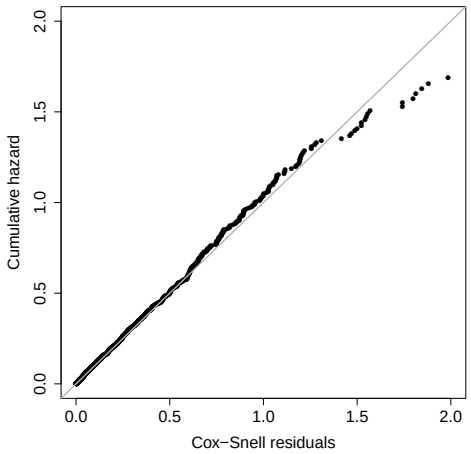


Fig. 3: Cox-Snell residuals for the Cox-model

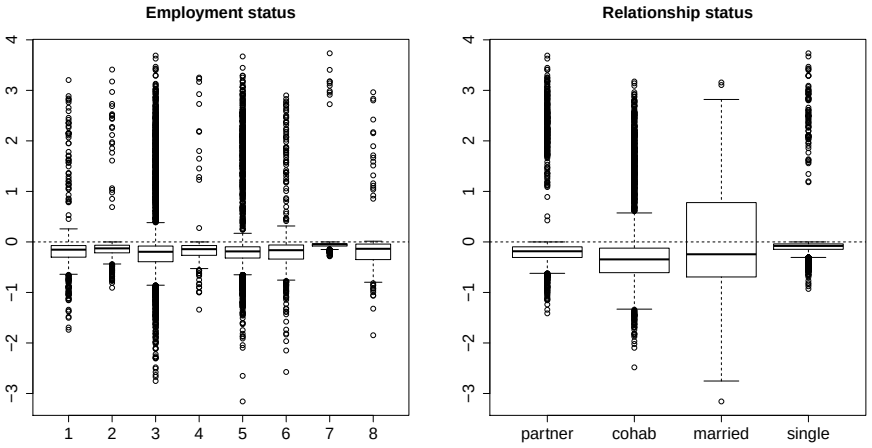


Fig. 4: Cox-deviance residuals for the Cox-model

	$\exp(\beta)$	$\text{se}(\beta)$	ρ	χ^2	$P(\cdot > \chi^2)$
Employment status (Ref.: unempl./job-seeking/housewife)					
education	0.708	0.244	0.006	0.045	.832
full-time employed	$0.747^{\bullet}$	0.154	0.046	3.572	.059
marginally employed	0.604	0.318	0.006	0.051	.822
no info	0.882	0.157	0.037	2.444	.118
part-time employed	0.786	0.200	0.024	1.031	.310
school	0.247^{***}	0.392	-0.014	0.292	.589
self-employed	0.794	0.279	0.048	3.679	.055
Cohort (Ref.: cohort 1)					
cohort 2	1.049	0.065	0.006	0.059	.809
cohort 3	0.884	0.348	0.016	0.392	.531
Relationship status (Ref.: partner)					
cohabitation	3.103^{***}	0.084	0.008	0.125	.724
married	6.543^{***}	0.098	-0.085	14.208	$< .001$
single	0.272^{***}	0.131	-0.042	3.027	.082
Education level (Ref.: no graduation)					
A levels	0.730	0.717	0.040	9.826	.002
apprenticeship	0.546	0.696	0.040	10.362	.001
college of higher education	0.440	0.697	0.045	13.069	$< .001$
general secondary education	1.100	0.717	0.034	7.321	.007
no info	0.611	0.983	0.014	0.802	.370
secondary education	1.513	0.703	0.034	7.425	.006
traineeship	0.579	0.695	0.038	9.284	.002
university with PhD	0.342	0.711	0.046	13.158	$< .001$
university without PhD	0.358	0.697	0.049	15.783	$< .001$
# siblings (Ref.: no siblings)					
one sibling	1.042	0.082	0.045	0.09	7.68e-01
two siblings	0.967	0.097	0.036	2.59	1.08e-01
three or more siblings	1.291^{*}	0.106	-0.004	0.03	8.54e-01
Education level parents (Ref.: no graduation)					
A levels	0.430	0.516	-0.018	0.345	.557
apprenticeship	0.492^{*}	0.296	-0.056	3.557	.059
college of higher education	0.526^{*}	0.293	-0.052	3.084	.079
general secondary education	1.156	0.553	-0.007	0.068	.795
no info	0.725	0.291	-0.076	6.448	.112
secondary education	0.578	0.343	-0.045	3.104	.078
traineeship	$0.573^{\bullet}$	0.286	-0.070	5.385	.020
other	0.134^{***}	0.330	0.013	0.185	.667
university with PhD	0.429^{*}	0.395	-0.043	2.347	.125
university without PhD	$0.603^{\bullet}$	0.298	-0.054	3.202	.074
Global test				158.451	$< .001$

Table 6: Estimated (exponential) regression coefficients together with robust standard errors (left) and test on the PH-assumption (right) for the Cox-model on the pairfam-data; $^{\bullet}p < 0.1$;* $p < 0.05$;** $p < 0.01$;*** $p < 0.001$.

References

[1] O. O. Aalen. A linear regression model for the analysis of life-times. *Statistics in Medicine*, 8:907–925, 1989.

[2] N. Beck and J. Hartmann. Die Wechselwirkung zwischen Erwerbstätigkeit der Ehefrau und Ehestabilität unter der Berücksichtigung des sozialen

Wandels. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 51:655–680, 1999.

[3] Gary S. Becker. An economic analysis of fertility. In National Bureau of Economic Research, editor, *Demographic and economic change in developed countries: a conference on the Universities-National Bureau Committee for Economic Research*, pages 209–231. Princton University Press, Princton, 1960.

[4] H.P. Blossfeld, G. Rower, and K. Golsch. *Event history analysis with Stata*. NJ: Erlbaum, Mahwa, 2007.

[5] J. Brüderl, A. Diekmann, and H. Engelhardt. Erhöht eine Probeehe das Scheidungsrisiko? Eine empirische Untersuchung mit dem Familiensurvey. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 49:205–222, 1997.

[6] M. J. Budig. Are women's employment and fertility histories independent?an examination of causal order using event history analysis. *Social Science Research*, 32:376–401, 2003.

[7] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, B 34:187–220, 1972.

[8] D. R. Cox and E. J. Snell. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society series B*, 30:248–275, 1968.

[9] J. C. Cramer. Fertility and female employment - problems of causal direction. *American Sociological Review*, 45:167–190, 1980.

[10] J.-M. Dornseiff and R. Sackmann. Familien-, Erwerbs- und Fertilitätsdynamiken in Ost- und Westdeutschland. In W. Bien and J. H. Marbach, editors, *Partnerschaft und Familiengründung, Ergebnisse der dritten Welle des Familien-Survey*. Leske+Budrich, Opladen, 2003.

[11] D. H. Felmlee. The dynamic interdependence of women's employment and fertility. *Social Science Research*, 22:333–360, 1993.

[12] P. Grambsch and T. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

[13] B. Hoem and J. M. Hoem. The impact of women's employment on 2nd and 3rd births in modern Sweden. *Population Studies*, 43:47–67, 1989.

[14] V.J. Hotz, J. A. Klerman, and R. J. Willis. The economics of fertility in developed countries. In M. Rosenzweig and O. Stark, editors, *Handbook of Popoulation and Family Economics*, pages 275–347. 1997.

[15] J. Huinink and D. Konietzka. *Familiensoziologie. Eine Einföhrung*. Campus Verlag GmbH, Frankfurt am Main, 2007.

[16] J. Huinink, J. Brüderl, B. Nauck, S. Walper, L. Castiglioni, and M. Feldhaus. Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Journal of Family Research*, 23: 77–101, 2011.

[17] A. Kohlmann and J. Kopp. Verhandlungstheoretische modellierung des Übergangs zu verschiedenen Kinderzahlen. *Zeitschrift für Soziologie*, 26: 258–274, 1997.

[18] M. Kreyenfeld. *Employmend and Fertility - East Germany in the 1990s*. University Rostock, Rostock, 2001.

[19] M. Kreyenfeld. Fertility decision in the FRG and GDR: An analysis with data from the German fertility and family survey. *Demographic Research Special Collection*, 3:275–318, 2004.

[20] C. Lauer and A. M. Weber. Employment and mothers after childbirth: a French-German comparison. *ZWE Discussion Paper*, 03-50, 2003.

[21] H. Leibenstein. *Economic Backwardness and Economic Growth.* John Wiley & Sons, New York, 1957.

[22] Harvey Leibenstein. An interpretation of the economic theory of fertility: Promising path or blind alley? *Journal of Economic Literature*, 12:457–479, 1974.

[23] A.C. Liefbroer and M. Corijn. Who, what, where, and when? Specifying the impact of educational attainment and labour force participation on family formation. *European Journal of Population*, 15:45–75, 1999.

[24] B. Nauck, J. Brüderl, J. Huinink, and S. Walper. The german family panel (pairfam). *GESIS Data Archive, Cologne*, 2013. ZA5678 Data file Version 4.0.0.

[25] J. Schröder and J. Brüderl. Der Effekt der Erwerbsttigkeit von Frauen auf die Fertilitt: Kausalitt oder Selbstselektion? *Zeitschrift fr Soziologie*, 37:2:117–136, 2008.

[26] R. M. Stolzenberg and L. J. Waite. Age, fertility expectations and plans for employment. *American Sociological Review*, 42:769–783, 1977.

[27] T. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, 2000.

[28] R. J. Willis. New approach to economic theory of fertility behavior. *Journal of Political Economy*, 81:14–64, 1973.

[29] R. E. Wright, J. F. Ermisch, and P. R. A. Hinde. The 3rd birth in great-britain. *Journal of Biosocial Science*, 20:489–496, 1988.

A survey on the isometry of certain orthogonal polynomial systems in martingale spaces

Edmundo J. Huertas[1] and Nuria Torrado[2]

[1] Centre for Mathematics, University of Coimbra (CMUC), Largo D. Dinis, Apartado 3008, EC Santa Cruz, 3001-501 Coimbra, Portugal
(E-mail: ehuertasce@mat.uc.pt, ehuertasce@gmail.com)
[2] Centre for Mathematics, University of Coimbra (CMUC), Largo D. Dinis, Apartado 3008, EC Santa Cruz, 3001-501 Coimbra, Portugal
(E-mail: nuria.torrado@mat.uc.pt, nuria.torrado@gmail.com)

Abstract. In this paper we survey how an inner product derived from an Uvarov transformation of the Laguerre weight function is used in the orthogonalization procedure of a sequence of martingales related to a Lévy process. The orthogonalization is done by isometry and it is based in previous works of Nualart and Schoutens (see [18] and [19]), where the resulting set of pairwise strongly orthogonal martingales involved are used as integrators in the so-called chaotic representation property. Finally, we give an idea of how to generalize the above works.
Keywords: Orthogonal polynomials; Laguerre-type polynomials; Krall-Laguerre polynomials; Inner products; Lévy processes; Stochastic processes.

1 Introduction

The Laguerre orthogonal polynomials are defined as the polynomials orthogonal with respect to the Gamma distribution. Therefore, they are orthogonal with respect to the inner product in the linear space $\mathbb{P}$ of polynomials with real coefficients (see [2])

$$\langle p, q\rangle_{\alpha} = \int_0^{\infty} pqx^{\alpha}e^{-x}dx, \quad \alpha > -1,\ p, q \in \mathbb{P}. \tag{1}$$

From now on, $\{\widehat{L}_n^{\alpha}(x)\}_{n\geq 0}$ stands for the sequence of monic Laguerre polynomials orthogonal with respect to (1). From the above inner product, let us introduce the modified inner product

$$\langle p, q\rangle = \int_0^{\infty} pqx^{\alpha}e^{-x}dx + \sigma^2 p(c)q(c), \quad \alpha > -1,\ p, q \in \mathbb{P} \tag{2}$$

where $\sigma^2 \in \mathbb{R}_+$, and $c \in (-\infty, 0]$. Notice that $\langle p, q\rangle = \langle p, q\rangle_{\alpha} + \sigma^2 p(c)q(c)$, so therefore (2) can be interpreted as a modification (or perturbation) of the

3^{rd} *SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal*
C. H. Skiadas (Ed)

Laguerre measure $d\mu_\alpha(x) = x^\alpha e^{-x}dx$ with a discrete measure given by a mass point at $x = c$,

$$d\tilde{\mu}_\alpha(x) = x^\alpha e^{-x}dx + \sigma^2\delta(x-c),$$

where $\delta(x-c)$ is the Dirac delta at $x = c$. This perturbation is known as the Uvarov perturbation of the measure $d\mu_\alpha(x)$ (see [5], [6], [7] and the references given there). The case $c = 0$ has been deeply studied in the literature (see [3], [4], [11] among others). These polynomials are called either Laguerre-type polynomials (see, for instance, [3] and [14]) or Krall-Laguerre polynomials (in [8]). They were also obtained by T.H. Koornwinder [12] as a special limit case of the Jacobi-Koornwinder (Jacobi type) orthogonal polynomials, and they are also known as Laguerre-Koornwinder polynomials.

A Lévy process is a stochastic process with independent and stationary increments which consists of three basic stochastically independent parts: a deterministic part, a pure jump part and a Brownian motion. Lévy processes play an important role in many fields of science. For example, in engineering, they are used for the study of networks; in the actuarial science, for the calculation of insurance and re-insurance risk and, in economics, for continuous time-series models. In the last decades, the study of the relation between orthogonal polynomials and Lévy processes have become increasing, see [17], [18], [21] and [22].

Consider a Lévy process and let σ^2 the constant of the Brownian motion part and ν its Lévy measure. Nualart and Schoutens [17] and Schoutens [19] have shown that there exists an isometry between the space of orthogonal polynomials with respect to the inner product (2) when $c = 0$ and the space of strongly orthogonal martingales which are the building blocks of a kind of chaotic representation of the square functionals of the Lévy process. The chaotic representation property (CRP) says that any square integrable random variable measurable with respect to normal martingales X can be expressed as an orthogonal sum of multiple stochastic integrals with respect to X. Based in the aforementioned previous works of Nualart and Schoutens, we are interested in finding new isometries that encompass the above cases using new families of orthogonal polynomials that generalize the Laguerre-type orthogonal polynomials used by Schoutens in [18]. Recently, several authors have begun to study the case when c is a negative number, i.e., the mass point is located outside the support of the Laguerre measure. The study of their asymptotic and analytic properties can be founded in [5], [6] or [7].

Our main goal is to consider a natural generalization of the work done in [18]. In other words, meanwhile Schoutens analyzed the connection between Lévy processes and Laguerre-type orthogonal polynomials with respect to the inner product (2) for the particular case $c = 0$, we study a more general case where $c \in (-\infty, 0)$. In our opinion, the differences between these two cases are sufficient to justify a new study of the isometry of these polynomials with certain sets of martingales.

The structure of the manuscript is as follows. In Section 2, we summarize some properties of Laguerre-type orthogonal polynomials to be used in the sequel. We briefly review the concept of Lévy process and study the orthogonalization procedure for a sequence of martingales related to the powers of the

jumps of this stochastic process in Section 3. As an original contribution that have not been published elsewhere, in Section 4 we provide the explicit coefficients of the generalize Laguerre-type orthogonal polynomials which eventually can lead to the desired isometry with some new space of Teugel martingales. We also discuss the work done so far and stress the problems founded to figure out the desired isometry, and give some few ideas that may help to find isometries for these new families of modified (standard and non standard) sequences of orthogonal polynomials.

2 Laguerre-type orthogonal polynomials

First, we review some basic properties of classical Laguerre polynomials $\widehat{L}_n^\alpha(x)$ useful in the sequel. Their corresponding norm is given by $||\widehat{L}_n^\alpha||_\alpha^2 = n!\Gamma(n+\alpha+1)$. Dealing with Laguerre polynomials, it is customary to use the normalization such that the leading coefficient of the n-th degree classical Laguerre polynomial (denoted by $L_n^{(\alpha)}(x)$) equals $\frac{(-1)^n}{n!}$, i.e., $L_n^{(\alpha)}(x) = \frac{(-1)^n}{n!}x^n +$ *lower degree terms,* and therefore

$$L_n^{(\alpha)}(x) = \frac{(-1)^n}{n!}\widehat{L}_n^\alpha(x).$$

It is very well known that these polynomials satisfy the following three term recurrence relation

$$x\widehat{L}_n^\alpha(x) = \widehat{L}_{n+1}^\alpha(x) + \beta_n\widehat{L}_n^\alpha(x) + \gamma_n\widehat{L}_{n-1}^\alpha(x), \quad n \geq 1, \tag{3}$$

with initial conditions $\widehat{L}_0^\alpha(x) = 1$, $\widehat{L}_1^\alpha(x) = x - (\alpha+1)$, and recurrence coefficients $\beta_n = 2n+\alpha+1$, $\gamma_n = n(n+\alpha)$ for every $n \geq 1$ (see [16], [23] among others). They constitute a family of classical orthogonal polynomials (see [13] and [16]), and they are the eigenfunctions of a second order linear differential operator with polynomial coefficients. The kernel polynomials (see [2, Ch.I, Â§7]) associated with Laguerre polynomials will play a key role in order to obtain some conclusions of the manuscript. Let

$$K_n(x,y) = \sum_{k=0}^{n} \frac{\widehat{L}_k^\alpha(x)\widehat{L}_k^\alpha(y)}{||\widehat{L}_k^\alpha||_\alpha^2}$$

denotes the n-th kernel polynomial associated with the Laguerre orthogonal polynomials. Thus, according to the Christoffel-Darboux formula, for every $n \in \mathbb{N}$ we get the alternative expression

$$K_n(x,y) = \frac{\widehat{L}_{n+1}^\alpha(x)\widehat{L}_n^\alpha(y) - \widehat{L}_{n+1}^\alpha(y)\widehat{L}_n^\alpha(x)}{x-y}\frac{1}{||\widehat{L}_n^\alpha||_\alpha^2}.$$

The limit when $y \to x$ is known as the *confluent form* of the n-th kernel, and it reads

$$K_n(x,x) = \sum_{k=0}^{n} \frac{[\widehat{L}_k^\alpha(x)]^2}{||\widehat{L}_k^\alpha||_\alpha^2} = \frac{[\widehat{L}_{n+1}^\alpha(x)]'\,\widehat{L}_n^\alpha(x) - [\widehat{L}_n^\alpha(x)]'\,\widehat{L}_{n+1}^\alpha(x)}{||\widehat{L}_n^\alpha||_\alpha^2}.$$

From now on, $\{\widehat{L}_n^{\alpha,c,\sigma^2}(x)\}_{n\geq 0}$ denotes the sequence of monic polynomials orthogonal with respect to (2) when $c \in (-\infty, 0)$, and $\{\widehat{L}_n^{\alpha,\sigma^2}(x)\}_{n\geq 0}$ stands for the monic Laguerre-type orthogonal polynomials with $c = 0$.

We next present some specific properties of $\widehat{L}_n^{\alpha,c,\sigma^2}(x)$, showing the differences which appear when $c = 0$ or $c \in (-\infty, 0)$. The first remarkable fact is that the position of the first (or least) zero of the Laguerre-type polynomials, strongly depends on the value of the real and positive parameter σ^2, and the position of the mass point c (see [7] for detailed study). Obviously, if $\sigma^2 = 0$, the zeros of the Laguerre-type polynomials trivially reduces to the zeros of the classical Laguerre polynomials. Moreover, if $\sigma^2 > 0$ and $c \in (-\infty, 0)$, then one can find values of σ^2 for which the least zero of $\widehat{L}_n^{\alpha,c,\sigma^2}(x)$, $n \geq 1$, is located in the interval $(c, 0)$, i.e., outside of the support of the classic Laguerre measure, whereas if $\sigma^2 > 0$ and $c = 0$ this phenomenon does not occur at all, and all the zeros of $\widehat{L}_n^{\alpha,\sigma^2}(x)$, $n \geq 1$, are located inside the interval $(0, +\infty)$ for any value of σ^2. For every $n = 1, 2, \ldots$, the polynomials $\widehat{L}_n^{\alpha,c,\sigma^2}(x)$ satisfy as well a three term recurrence relation

$$x\widehat{L}_n^{\alpha,c,\sigma^2}(x) = \widehat{L}_{n+1}^{\alpha,c,\sigma^2}(x) + \tilde{\beta}_n \widehat{L}_n^{\alpha,c,\sigma^2}(x) + \tilde{\gamma}_n \widehat{L}_{n-1}^{\alpha,c,\sigma^2}(x), \quad n \geq 1,$$

with recurrence coefficients

$$\tilde{\beta}_n = \beta_n + \frac{\widehat{L}_{n+1}^{\alpha}(c)}{\widehat{L}_n^{\alpha}(c)}\left(1 - \frac{1 + \sigma^2 K_{n-1}(c,c)}{1 + \sigma^2 K_n(c,c)}\right) - \frac{\widehat{L}_n^{\alpha}(c)}{\widehat{L}_{n-1}^{\alpha}(c)}\left(1 - \frac{1 + \sigma^2 K_{n-2}(c,c)}{1 + \sigma^2 K_{n-1}(c,c)}\right),$$

$$\tilde{\gamma}_n = \frac{\left(1 + \sigma^2 K_n(c,c)\right)\left(1 + \sigma^2 K_{n-2}(c,c)\right)}{\left(1 + \sigma^2 K_{n-1}(c,c)\right)^2}\gamma_n.$$

where β_n and γ_n are the recurrence coefficients in (3) for the classical Laguerre polynomials. Notice that $\widehat{L}_n^{\alpha}(c) \neq 0$ for every $n = 0, 1, 2, \ldots$, because c does not belong to the support of the classical Laguerre measure. Finally, a very remarkable difference appear when we express the aforementioned families in terms of Gauss hypergeometric functions. The classical Laguerre polynomials $\widehat{L}_n^{\alpha}(x)$ can be expressed as hypergeometric functions of type ${}_1F_1$ (see [9], [23] among others). The addition of a mass point at $c = 0$ implies that the Laguerre-type polynomials $\widehat{L}_n^{\alpha,\sigma^2}(x)$, which are used in [18], are expressed in terms of ${}_2F_2$ hypergeometric functions (see, for example [10]), meanwhile moving the mass point to $c < 0$, as in our case, implies that the Laguerre-type polynomials $\widehat{L}_n^{\alpha,c,\sigma^2}(x)$ turn to be expressed in terms of ${}_3F_3$ hypergeometric functions (see [5]).

3 Lévy processes and Teugels martingales

Let $X = \{X_t, t \geq 0\}$ be a Lévy process (meaning that X has stationary and independent increments and is continuous in probability and that $X_0 = 0$), cadlag and centered, with moments of all orders. Let us remind that a stochastic process is cadlag if its sample paths are right continuous and have left-hand limits. Denote by σ^2 the variance of the Gaussian part of X and by ν its Lévy

measure. The existence of moments of all orders of X_t implies that the Lévy measure ν has moments of all orders ≥ 2. Write

$$m_n = \int_{\mathbb{R}} x^n \nu(dx) \quad \text{for} \quad n \geq 2.$$

For background on all these notions, we refer to Bertoin [1] and Sato [20]. Following Nualart and Schoutens [17], we introduce the square-integrable martingales (and Lévy processes) called Teugels martingales, related to the powers of the jumps of the process:

$$\begin{aligned} Y_t^{(1)} &= X_t, \\ Y_t^{(n)} &= \sum_{0<s\leq t} (\Delta X_t)^n - m_n t, \quad n \geq 2, \end{aligned}$$

where $\Delta X_t = X_t - X_{t^-}$ is the jump size at time t and

$$X_{t^-} = \lim_{s<t, s\to t} X_s, \ \ t > 0$$

is the left limit process. The compensated power jump process $Y^{(n)}$ of order n is a normal martingale.

An important question is the orthogonalization of the set $\left\{Y^{(n)},\, n = 1, 2, \ldots\right\}$ of martingales, called *Teugels Martingales*, as stochastic integrators of a kind of chaotic representation as we briefly discuss in the Introduction. Specifically, let $\mathcal{M}^2$ be the space of square-integrable martingales M such that $\sup_t E(M_t^2) < \infty$ and $M_0 = 0$ a.s. We recall that two martingales $M, N \in \mathcal{M}^2$ are strongly orthogonal if and only if their product MN is a uniform integrable martingale. Nualart and Schoutens [17] showed that every random variable F in $L^2(\Omega, \mathcal{F})$ has a representation of the form

$$F = E\left[F\right] +$$

$$\sum_{j=1}^{\infty} \sum_{(i_1,\ldots,i_j)\in N^j} \int_0^{\infty} \int_0^{t_1-} \cdots \int_0^{t_{j-1}-} f_{(i_1,\ldots,i_j)}(t_1, \ldots, t_j) dH_{t_j}^{(i_j)} \ldots dH_{t_2}^{(i_2)} dH_{t_1}^{(i_1)}$$

where $f_{(i_1,\ldots,i_j)}$'s are real deterministic functions, $N = \{1, 2, 3, \ldots\}$ and $\left\{H^{(i)}, i = 1, 2, \ldots\right\}$ is an orthogonalized set of martingales. A direct consequence is the weaker predictable representation property (PRP) with respect to the same set of orthogonalized martingales, saying that every random variable F in $L^2(\Omega, \mathcal{F})$ has a representation of the form

$$F = E\left[F\right] + \sum_{j=1}^{\infty} \int_0^{\infty} \Phi_s^{(i)} dH_s^{(i)},$$

where $\Phi_s^{(i)}$ is predictable (see [17] for more details).

Nualart and Schoutens[17], by using the measure $d\mu = x^r \nu(dx) + \sigma^2 \delta_0(dx)$, where δ_0 denotes the Dirac measure at point 0, σ^2 is the variance of the Gaussian part and ν the Lévy measure of a Lévy process, showed that the mapping

$x^{n-1} \longleftrightarrow Y^{(n)}$ defines an isometry between the space of polynomials $\mathbb{P} \subseteq L^2(\mu)$ and $\mathrm{Span}(\{Y^{(1)}, Y^{(2)}, Y^{(3)}, \ldots\}) \subseteq \mathcal{H}^2([0,1])$, where $\mathcal{H}^2([0,1])$ is the space of square integrable martingales on the time interval $[0,1]$. This isometry is given by the equality

$$\langle x^{i-1}, x^{j-1} \rangle_1 = m_{i+j} + \sigma^2 1_{\{i=j=1\}} = \langle Y^{(i)}, Y^{(j)} \rangle_2, \quad \text{for} i, j \geq 1.$$

Here the scalar product $\langle \cdot, \cdot \rangle_1$ in $L^2(\mu)$ is given by

$$\langle p(x), q(x) \rangle_1 = \int_{-\infty}^{+\infty} p(x)q(x)x^2\nu(dx) + \sigma^2 p(0)q(0),$$

whereas $\langle \cdot, \cdot \rangle_2$ denotes the scalar product in $\mathcal{H}^2([0,1])$. We observe that, because of the special structure of the Teugels martingales, for $M, N \in \mathrm{Span}(\{Y^{(1)}, Y^{(2)}, Y^{(3)}, \ldots\})$ the relation $\langle M, N \rangle_2 = 0$ is equivalent to the property that the martingales M and N are strongly orthogonal.

Nualart and Schoutens[17] and Schoutens[18] used this isometry for an orthogonalization procedure of the Teugels martingales: If the polynomials $P_0, P_1, \ldots$ are an orthogonalization of the monomials $\{1, x, x^2, \ldots\}$ in $L^2(\mu)$ and if these monomials $\{1, x, x^2, \ldots\}$ in $L^2(\mu)$ are substituted by $\{Y^{(1)}, Y^{(2)}, Y^{(3)}, \ldots\}$, then we obtain a system of strongly orthogonal martingales $\{H^{(1)}, H^{(2)}, H^{(3)}, \ldots\}$ with $\mathrm{Span}(\{H^{(1)}, H^{(2)}, H^{(3)}, \ldots\}) = \mathrm{Span}(\{Y^{(1)}, Y^{(2)}, Y^{(3)}, \ldots\})$. In particular, if the Lévy process is a Gamma process, then the obtained polynomials are just the Laguerre polynomials with parameter $\alpha = 1$. This makes it possible to find the coefficients for the above linear martingale transformation explicitly.

Here we consider a modified measure $d\mu^c = x^2\nu(dx) + \sigma^2\delta_c(dx)$, where δ_c denotes the Dirac measure at point $c < 0$, σ^2 is the variance of the Gaussian part and ν the Lévy measure of a Lévy process. In the concrete case of a Laguerre weight function this is related with the notion of the Uvarov transformation. In the case of the Gamma process this would lead to Laguerre-type polynomials also called Krall-Laguerre polynomials defined in Section 2.

Thus, we consider a new first space S_1^c as the space of all real polynomials on the positive real line endowed with the scalar product $\langle \cdot, \cdot \rangle_1$ given by

$$\langle p(x), q(x) \rangle_1 = \int_{-\infty}^{+\infty} p(x)q(x)x^2\nu(dx) + \sigma^2 p(c)q(c)$$

and a second space

$$S_2 = \{a_1 Y^{(1)} + a_2 Y^{(2)} + \cdots + a_n Y^{(n)} : n \in \{1, 2, \ldots\}, a_i \in \mathbb{R}, i = 1, \ldots, n\}$$

endowed with the scalar product

$$\langle X, Y \rangle_2 = E([X, Y]_1).$$

The elements of the space S_2 are linear combinations of Teugels martingales and the orthogonalization procedure produces a set of strongly pairwise orthogonal martingales

$$\{H^{(j)} = a_{1,j} Y^{(1)} + \cdots + a_{j,j} Y^{(j)}, \quad j = 1, 2, \ldots\}$$

that can be used in the *chaotic representation property* defined above.

4 Coefficients in the orthogonalization procedure

As detailed in [18], the coefficients of the Laguerre-type polynomials when $c = 0$ are used in the orthogonalization process of the Teugels martingales. In this section, as an original contribution we give the coefficients of the modified Laguerre-type polynomials for $c < 0$ which, to the best of our knowledge, they have not been previously computed or published elsewhere. Our guess is that these coefficients will eventually be needed in new attempts to find the desired new isometries.

In order to find the coefficients $\{b_{k,n}\}_{k=0}^{n}$ of the Laguerre-type polynomials $L_n^{\alpha,c,\sigma^2}(x)$ when $c \in (-\infty, 0)$, we introduce the notation

$$\lambda_n^{\alpha,c} = \frac{L_{n+1}^{(\alpha)}(c)}{L_n^{(\alpha)}(c)}, \text{ and } \kappa_n^{\alpha,c} = 1 + \sigma^2 K_n(c,c).$$

Notice that $L_n^{(\alpha)}(c) \neq 0$ for every $n = 0, 1, 2, \ldots$, because c does not belong to the support of the classical Laguerre measure. In [5, Th. 1] the authors obtained a connection formula between the monic Laguerre-type orthogonal polynomials and the classical monic Laguerre orthogonal polynomials. Following [18], we will consider the alternative normalization of Laguerre-type polynomials with leading coefficient $\frac{(-1)^n}{n!}\kappa_{n-1}^{\alpha,c}$, and we will denote them by $L_n^{\alpha,c,\sigma^2}(x)$ when $c \in (-\infty, 0)$, and by $L_n^{\alpha,\sigma^2}(x)$ when $c = 0$. Using this normalization, the connection formula [5, Th. 1] between the Laguerre-type and the classical Laguerre polynomials reads

$$(x - c)\, L_n^{\alpha,c,\sigma^2}(x) = A_n L_{n+1}^{(\alpha)}(x) + B_n L_n^{(\alpha)}(x) + C_n L_{n-1}^{(\alpha)}(x), \tag{4}$$

where

$$\begin{aligned} A_n &= -(n+1)\kappa_{n-1}^{\alpha,c}, \\ B_n &= (n+1)\kappa_{n-1}^{\alpha,c}\lambda_n^{\alpha,c} + (n+\alpha)\frac{\kappa_n^{\alpha,c}}{\lambda_{n-1}^{\alpha,c}}, \\ C_n &= -(n+\alpha)\kappa_n^{\alpha,c}. \end{aligned}$$

Next, we would like to obtain the coefficients $\{b_{k,n}\}_{k=0}^{n}$ such that

$$L_n^{\alpha,c,\sigma^2}(x) = b_{n,n}x^n + b_{n-1,n}x^{n-1} + \cdots + b_{1,n}x + b_{0,n}.$$

Proposition 1. *Let*

$$L_n^{\alpha,c,\sigma^2}(x) = \sum_{k=0}^{n} b_{k,n}x^k$$

be the Laguerre-type polynomials, orthogonal with respect to the inner product (2). Then, the sequence $\{b_{k,n}\}_{k=0}^{n}$ is given by

$$\begin{cases} b_{n,n} = \dfrac{(-1)^n}{n!}\kappa_{n-1}^{\alpha,c}, \\ b_{k-1,n} = t_{k,n+1} + c b_{k,n}, \quad k = n, n-1, \ldots, 1, \end{cases}$$

where

$$t_{k,n+1} = \begin{cases} \dfrac{(-1)^n}{n!}\kappa_{n-1}^{\alpha,c}, & \text{for } k = n+1, \\ \dfrac{u_{k,n}(\alpha,c)}{n!k!}(-n)_k(\alpha+k+1)_{n-k}, & \text{for } 0 \le k \le n, \end{cases}$$

$$u_{k,n}(\alpha,c) = (n+1)\left(\lambda_n^{\alpha,c} + \frac{\alpha+(n+1)}{k-(n+1)}\right)\kappa_{n-1}^{\alpha,c} + \left((k-n) + \frac{(\alpha+n)}{\lambda_{n-1}^{\alpha,c}}\right)\kappa_n^{\alpha,c}$$

and

$$\sum_{k=0}^{n+1} t_{k,n+1}c^k = 0.$$

Proof. The proof will be divided into 2 steps. First, from (4) we will obtain the coefficients $\{t_{k,n+1}\}_{k=0}^{n+1}$ of the polynomial

$$T_{n+1}(x) = (x-c)\,L_n^{\alpha,c,\sigma^2}(x) = \sum_{k=0}^{n+1} t_{k,n+1}x^k, \tag{5}$$

which is obviously related with the Laguerre-type polynomials $L_n^{\alpha,c,\sigma^2}(x)$. Second, we obtain the desired coefficients $\{b_{k,n}\}_{k=0}^{n}$ from $\{t_{k,n+1}\}_{k=0}^{n+1}$.

From (4), and the explicit coefficients for the classical Laguerre polynomials with leading coefficient $\frac{(-1)^n}{n!}$, (see [18])

$$L_n^{(\alpha)}(x) = \frac{1}{n!}\sum_{k=0}^{n}(-n)_k(\alpha+k+1)_{n-k}\frac{x^k}{k!}, \quad \alpha > -1,$$

we get

$$(x-c)\,L_n^{\alpha,c,\sigma^2}(x) = \frac{(-1)^n}{n!}\kappa_{n-1}^{\alpha,c}x^{n+1} + \frac{1}{n!}\sum_{k=0}^{n} u_{k,n}(\alpha,c)(-n)_k(\alpha+k+1)_{n-k}\frac{x^k}{k!},$$

where

$$u_{k,n}(\alpha,c) = (n+1)\left(\lambda_n^{\alpha,c} + \frac{\alpha+(n+1)}{k-(n+1)}\right)\kappa_{n-1}^{\alpha,c} + \left((k-n) + \frac{(\alpha+n)}{\lambda_{n-1}^{\alpha,c}}\right)\kappa_n^{\alpha,c}.$$

Thus,

$$t_{k,n+1} = \begin{cases} \dfrac{(-1)^n}{n!}\kappa_{n-1}^{\alpha,c}, & \text{for } k = n+1, \\ \dfrac{u_{k,n}(\alpha,c)}{n!k!}(-n)_k(\alpha+k+1)_{n-k}, & \text{for } 0 \le k \le n. \end{cases} \tag{6}$$

Next, we deduce the sequence $\{b_{k,n}\}_{k=0}^{n}$ in terms of $\{t_{k,n+1}\}_{k=0}^{n+1}$. (5) makes it obvious that, for every $n \ge 0$

$$T_{n+1}(x) = (x-c)\,L_n^{\alpha,c,\sigma^2}(x),$$

$$t_{n+1,n+1}x^{n+1} + \sum_{k=1}^{n} t_{k,n+1}x^k + t_{0,n+1} = b_{n,n}x^{n+1} + \sum_{k=1}^{n}(b_{k-1,n} - cb_k)x^k - cb_{0,n},$$

being c a root of $T_{n+1}(x)$, i.e.

$$\sum_{k=0}^{n+1} t_{k,n+1} c^k = 0.$$

Hence, the following relations matching the coefficients of $T_{n+1}(x)$ and $L_n^{\alpha,c,\sigma^2}(x)$ hold

$$\begin{cases} t_{n+1,n+1} = b_{n,n}, \\ t_{k,n+1} = b_{k-1,n} - c b_{k,n}, \quad 1 \leq k \leq n, \\ t_{0,n+1} = -c b_{0,n}. \end{cases}$$

The above provide a simple recursive rule to obtain the n coefficients of $L_n^{\alpha,c,\sigma^2}(x)$, as follows

$$\begin{cases} b_{n,n} = t_{n+1,n+1}, \\ b_{k-1,n} = t_{k,n+1} + c b_{k,n}, \quad k = n, n-1, \ldots, 1. \end{cases}$$

From (6) the statement holds.

To give an idea of the work done so far, we mention our first attempt to get a new isometry for values of $c < 0$, we tried to construct a new family of martingales

$$\widetilde{Y}^{(i)} = \sum_{\ell=0}^{i-1} \binom{i-1}{\ell} \frac{1}{c^\ell} Y^{(\ell+1)}$$

that are suitable linear combinations of Teugels martingales. If we consider the basis

$$\{(\frac{x}{c} - 1)^n\}_{n \geq 0}$$

then we though that it could be possible to find that $(\frac{x}{c} - 1)^{n-1} \longleftrightarrow \widetilde{Y}^{(n)}$ works as an isometry between S_1 and S_2, but when one computes $\langle (\frac{x}{c} - 1)^{i-1}, (\frac{x}{c} - 1)^{j-1} \rangle_1$ and $\langle \widetilde{Y}^{(i)}, \widetilde{Y}^{(j)} \rangle_2$ one clearly sees that the above process fails. Work is currently underway on a new isometry and we hope to report these findings in a future paper.

By way of conclusion, we would like to remark that the substantial and recent advances (see the nice survey [15]) in these kind of modified inner products, such as the ones studied here, open the door to a vast mine of beautiful, interesting and accessible open problems.

Acknowledgements

The authors are thankful to professor Fabrizio Leisen at the School of Mathematics, Statistics and Actuarial Sciences of University of Kent (England) for useful discussions. This work was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC) under the project PEst-C/MAT/UI0324/2013. The research of N. Torrado was supported under the grant SFRH/BPD/91832/2012 and the research of E.J. Huertas under the grant SFRH/BPD/91841/2012 both by the Portuguese Government through the Fundação para a Ciência e a Tecnologia (FCT).

References

1. J. Bertoin, Lévy Processes. Cambridge University Press, Cambridge, 1996.
2. T.S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York. (1978).
3. T.S. Chihara, *Orthogonal polynomials and measures with end point masses*, Rocky Mountain J. Math. **15** (1985), 705–719.
4. H. Dueñas and F. Marcellán, *Laguerre-Type orthogonal polynomials. Electrostatic interpretation*, Int. J. Pure and Appl. Math. **38** (2007), 345-358.
5. H. Dueñas, E.J. Huertas and F. Marcellán, *Analytic Properties of Laguerre-type Orthogonal Polynomials*, Integral Transforms Spec. Funct. **22** (2011), 107–122.
6. B.Xh. Fejzullahu and R.Xh. Zejnullahu, *Orthogonal polynomials with respect to the Laguerre measure perturbed by the canonical transformations*, Integral Transforms Spec. Funct. **17** (2010), 569–580.
7. E.J. Huertas, F. Marcellán, and F.R. Rafaeli, *Zeros of orthogonal polynomials generated by canonical perturbations of measures*, Appl. Math. Comput. **218** (2012), 7109–7127.
8. F.A. Grünbaum, L. Haine and E. Horozov, *Some functions that generalize the Krall-Laguerre polynomials*, J. Comput. Appl. Math. **106** (1999), 271–297.
9. M.E.H. Ismail, *Classical and Quantum Orthogonal Polynomials in one variable*, Encyclopedia of Mathematics and its Applications Vol **98**, Cambridge University Press, Cambridge, UK. (2005).
10. R. Koekoek, *Generalizations of classical Laguerre polynomials and some q-analogues*. Doctoral Dissertation, Technical University Delft.(1990)
11. J. Koekoek and R. Koekoek, *Differential equation for Koornwinder's generalized Laguerre polynomials*. Proc. Amer. Math. Soc. **112** (1991), 1045–1054.
12. T.H. Koornwinder, *Orthogonal polynomials with weight function* $(1-x)^{\alpha}(1+x)^{\beta}+M\delta(x+1)+N\delta(x-1)$, Canad. Math. Bull. **27** (1984), 205–214.
13. F. Marcellán, A. Branquinho, and J. C. Petronilho, *Classical Orthogonal Polynomials: A Functional Approach*, Acta Appl. Math. **34** (1994), 283–303.
14. F. Marcellán and A. Ronveaux, *Differential equations for classical type orthogonal polynomials*, Canad. Math. Bull. **32** (1989), 404-411.
15. F. Marcellán and Y. Xu, *On Sobolev orthogonal polynomials*, arXiv preprint arXiv:1403.6249 (2014).
16. A.F. Nikiforov and V.B. Uvarov, *Special Functions of Mathematical Physics: An Unified Approach*, Birkhauser Verlag, Basel. (1988).
17. D. Nualart and W. Schoutens, *Chaotic and predictable representations for Lévy processes*, Stochastic processes and their applications 90 (2000) 109–122.
18. W. Schoutens, *An application in stochastics of the Laguerre-type polynomials*, J. Comput. Appl. Math., **133**, Issues 1–2, (2001), 593–600.
19. W. Schoutens, *Stochastic processes and orthogonal polynomials*, in Lecture Notes in Statist., vol **146**, Springer-Verlag, New York, (2000).
20. K. Sato, Lévy processes and infinitely divisible distributions. Cambridge University Studies in Advanced Mathematics, Vol. 68. Cambridge University Press, Cambridge (1999).
21. J.L. Solé and F. Utzet, *Time-Space harmonic polynomials relative to a Lévy processes*, Bernoulli 14 (2008), 1–13.
22. J.L. Solé and F. Utzet, *On the orthogonal polynomials associated with a Lévy processes*, The Annals of Probability 36 (2008), 765–795.
23. G. Szegő, *Orthogonal Polynomials*, 4^{th} ed., Amer. Math. Soc. Colloq. Publ. Series, vol **23**, Amer. Math. Soc., Providence, RI. (1975).

Modelling relations between returns of financial investments using perturbed copulas

Jozef Komorník[1], Magda Komorníková[2], and Jana Kalická[2]

[1] Faculty of Management, Comenius University, Odbojárov 10, P.O.BOX 95, 820 05 Bratislava, Slovakia
(E-mail: Jozef.Komornik@fm.uniba.sk)
[2] Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 813 68 Bratislava, Slovakia
(E-mail: Magdalena.Komornikova@stuba.sk, Jana.Kalicka@stuba.sk)

Abstract. We have investigated the relations between 4 selected countries' (USA, Australia, Japan and UK) daily returns of the REIT (Real Estate Investment Trust) indexes in different time periods, determined by the recent global financial markets crises (July 1, 2008 – April 30, 2009). We have applied the fitting by copulas to the residuals of ARMA–GARCH filters. We considered models from strict Archimedean copulas (Joe, Frank, Clayton and Gumbel) families and their mixtures with corresponding survival copulas as well as their perturbation. For selecting the optimal models we have applied the Kolmogorov – Smirnov – Anderson – Darling (KSAD) test statistic (for which we also constructed a GoF simulation based test). We observed that for all 3 considered time periods, the minimal (considerably reduced) value of KSAD were received for perturbed copulas.
Keywords: Copula, Perturbed copulas, Real Estate Investment Trust (REIT) index, returns of REIT indexes.

1 Introduction

We have investigated the relations between 4 selected countries' (USA, Australia, Japan and UK) daily returns of the REIT (Real Estate Investment Trust) indexes in different time periods, determined by the recent global financial markets crises (July 1, 2008 – April 30, 2009). We have applied the fitting by copulas to the residuals of ARMA–GARCH filters. We fitted these residuals by suitable marginal distributions in one of the Normal, Logistic and Laplace classes of distributions. Next, for each considered time period and all six possible couples of filtered residuals, we investigated models from strict Archimedean copulas (Joe, Frank, Clayton and Gumbel) families and their mixtures with corresponding survival copulas (that have been applied e.g. in Patton's paper [15]) as well as their perturbations given in our paper Mesiar *et al.*[12]. We observed that for all three considered time periods, the minimal (considerably reduced) value of Kolmogorov–Smirnov–Anderson–Darling test statistic were received for perturbed model.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

The paper is organized as follows. In the second section we present selected models of univariate (marginal) distributions as well as a brief overview of the theory of copulas including the methodology of their fitting to two–dimensional time series. The third section is devoted to perturbations of bivariate copulas. The fourth section, contains application to real data modelling. First we filter the considered group of REIT indexes (separately in the individual time sub–periods) by ARMA–GARCH models (in order to avoid a possible violation of the i.i.d. property). Then we fit the resulting time series of residuals by suitable marginal distributions and apply non–parametric correlation analyses (based on the Kendall coefficients) to all possible couples of the residual time series (for the individual time sub–periods). Next we provide an overview of the best copula models for different time sub–periods and selected significantly correlated pairs of returns of REIT indexes. Finally, some conclusions are presented.

2 Fitting univariate and bivariate distributions

2.1 Selected classes of univariate distributions

Recall that a *Logistic* distribution (see [5]) is determined by two parameters $(\mu,\ \beta)$ and its probability density function is given by

$$f(x,\mu,\beta) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(e^{-\frac{x-\mu}{\beta}}+1\right)^2}.$$

Moreover, its theoretical parameters satisfy the relations

$$E[X]=\mu, \quad D[X]=\frac{\pi^2\beta^2}{3}, \quad \text{Skewness } =0, \quad \text{Kurtosis } =4.2.$$

Similarly a *Laplace* distribution (see [10]) is determined by two parameters $(\mu,\ \beta)$ and its probability density function is given by

$$f(x,\mu,\beta) == \begin{cases} \frac{e^{-\frac{x-\mu}{\beta}}}{2\beta} & x \geq \mu \\ \\ \frac{e^{-\frac{\mu-x}{\beta}}}{2\beta} & \text{otherwise} \end{cases}.$$

Its theoretical parameters satisfy the relations

$$E[X]=\mu, \quad D[X]=2\beta^2, \quad \text{Skewness } =0, \quad \text{Kurtosis } =6.$$

For all GoF tests with Logistic and Laplace distributions we applied the Anderson–Darling GoF test (see e.g. Anderson and Darling [3], Anderson [2]) that effectively uses a test statistic based on

$$A_n^2 = n\int_{-\infty}^{\infty} \frac{\left(\hat{F}(x)-F(x)\right)^2}{F(x))(1-F(x)}\, d\,F(x), \tag{1}$$

where n is the sample size, $\hat{F}(x)$ is the empirical distribution function and $F(x)$ is the specified distribution function. It was shown in Anderson and Darling [3] that (1) can be written as

$$A_n^2 = -n - \frac{1}{n}\sum_{k=1}^{n}(2k-1)\left(\log\left(1-u_{(-k+n+1)}\right)+\log u_{(k)}\right),$$

where $u_{(k)} = F(x_{(k)})$ and $x_{(1)} < ... < x_{(n)}$ is the ordered sample. The p-values of Anderson–Darling GoF test were calculated by the software Mathematica, version 9.

2.2 Copulas

Copula represents a multivariate distribution that capture the dependence structure among random variables. It is a great tool for building flexible multivariate stochastic models. Copula offers the choice of an appropriate model for the dependence between random variables independently from the selection of marginal distributions. This concept was introduced in the early 50's and became popular in several fields beyond statistics and probability theory, such as finance, actuarial science, fuzzy set theory, hydrology, etc.

Definition 1. A function $C : [0,1]^2 \to [0,1]$ is called a (bivariate) copula whenever it is

i) 2–increasing, i.e.,

$$V_C\left([u_1,u_2]\times[v_1,v_2]\right) = C(u_1,v_1)+C(u_2,v_2)-C(u_1,v_2)-C(u_2,v_1) \geq 0$$

for all $0 \leq u_1 \leq u_2 \leq 1,\ 0 \leq v_1 \leq v_2 \leq 1$ (recall that $V_C\left([u_1,u_2]\times[v_1,v_2]\right)$ is the C–volume of the rectangle $[u_1,u_2]\times[v_1,v_2]$);

ii) grounded, i.e., $C(u,0) = C(0,v) = 0$ for all $u,v \in [0,1]$;

iii) it has a neutral element $e = 1$, i.e., $C(u,1) = u$ and $C(1,v) = v$ for all $u,v \in [0,1]$.

Sklar[16] proved in 1959 that $H(x,y) = C(F(x),G(y))$, where H is the joint distribution function of a random vector (X,Y) with marginal distribution functions F and G. If the marginals are continuous, the copula is unique. Thus, the copula function has other important interpretation as the joint distribution function.

For more details, examples and applications we recommend monographs Joe[9] and Nelsen[13]. The Table 1 provides a summary of some selected basic facts that are related to some classes of Archimedean copulas that we utilize in our analyses.

We follow the approach of Patton[15] and consider a so–called *survival copula* derived from a given copula $C(u,v)$ corresponding to the couple (X,Y) by

$$SC(u,v) = u+v-1+C(1-u,1-v) \tag{2}$$

which is the copula related to the couple $(-X,-Y)$ with the marginal distribution functions

$$F_{-X}(x) = 1-F_X(-x^+) \text{ and } F_{-Y}(y) = 1-F_Y(-y^+). \tag{3}$$

Family of copulas	Parameter	Bivariate copula $C(u,v)$
Gumbel	$\theta \geq 1$	$e^{-[(-ln(u))^\theta + (-ln(v))^\theta]^{\frac{1}{\theta}}}$
Clayton (strict)	$\theta > 0$	$(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$
Frank	$\theta \in \Re$	$-\frac{1}{\theta}\ln(1 + \frac{(e^{-\theta u}-1)\,(e^{-\theta v}-1)}{(e^{-\theta}-1)})$
Joe	$\theta \in [1,\infty)$	$1 - \left((1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right)^{1/\theta}$
Ali-Mikhail-Haq	$\theta \in [-1,1]$	$\frac{uv}{1-\theta(1-u)(1-v)}$

Table 1. Some Archimedean copulas

2.3 Fitting of copulas

In practical fitting of the data we have utilized the *maximum pseudolikelihood method* (MPL) of parameter estimation with initial parameters estimates received by the minimalization of the mean square distance to the empirical copula C_n presented e.g. in Genest and Favre[7]

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v\right)$$

where n is the sample size, R_i stands for the rank of X_i among $X_1,\ldots,X_n$, S_i stands for the rank of Y_i among $Y_1,\ldots,Y_n$ and $\mathbf{1}(\Omega)$ denoting the indicator function of set Ω. It requires that the copula $C_\theta(u,v)$ is absolutely continuous with density $c_\theta(u,v) = \frac{\partial^2}{\partial u\,\partial v}C_\theta(u,v)$. This method (described e.g. in Genest and Favre[7]) involves maximizing a rank-based log-likelihood of the form

$$L(\theta) = \sum_{i=1}^{n} \ln\left(c_\theta\left(\frac{R_i}{n+1}; \frac{S_i}{n+1}\right)\right).$$

where θ is vector of parameters in the model. Note that arguments $\frac{R_i}{n+1}$, $\frac{S_i}{n+1}$ equal to the corresponding values of the empirical marginal distributional functions of random variables X and Y.

For selecting the optimal models we applied the Kolmogorov – Smirnov – Anderson – Darling (KSAD, for which we use the abbreviation AD) test statistic defined e.g. in Berg and Bakken[6]

$$AD(\theta) = \sqrt{n}\ \max\left|\frac{C_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) - C_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right)}{C_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) * \left(1 - C_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right)\right)}\right| \tag{4}$$

for which we also constructed a GoF simulation based test, when comparing models with their submodels and different families of models.

3 Perturbation of bivariate copulas

Fitting of an appropriate copula to real data is one of major tasks in application of copulas. For this purpose, a large buffer of potential copulas is

necessary, preferably parametric families of copulas. Once we know approximately a copula C appropriate to model the observed data, we look for a minor perturbation of C which fit better then C itself. This is, e.g., the case of Farlie–Gumbel–Morgenstern (FGM) class of copulas, all of them being a perturbation of the independence copula Π, $\Pi(u,v) = u\,v$. Recall that FGM family $\left(C_\alpha^{FGM}\right)_{\alpha\in[-1,1]}$ of copulas is given by

$$C_\alpha^{FGM}(u,v) = u\,v + \alpha\,u\,(1-u)\,v\,(1-v). \tag{5}$$

Several generalizations of FGM approach to perturb the product copula Π can be found in literature, see, for example Amblard and Girard[1], Bairamov and Kotz [4], Rodríguez–Lallena and Úbeda–Flores[14].

For a given copula $C : [0,1]^2 \to [0,1]$, we will look for constraints on the noise $H : [0,1]^2 \to \Re$ so that the function $C_H : [0,1]^2 \to [0,1]$ given by

$$C_H(u,v) = \max\left(0,\ C(u,v) + H(u,v)\right) \tag{6}$$

is also a copula. Obviously FGM copulas given by (5) are linked to $C = \Pi$ and $H_\alpha(u,v) = \alpha\ u\ (1-u)\ v\ (1-v)$ (observe that in this case, no truncation is necessary).

For a fixed copula $C : [0,1]^2 \to [0,1]$, consider the function C_H given by (6). To satisfy the groundedness condition of copulas by C_H, necessarily $H(u,0) \le 0$ and $H(0,v) \le 0$ for all $u, v \in [0,1]$. Similarly, $e = 1$ is a neutral element of C_H only if $H(u,1) = H(1,v) = 0$ for all $u, v \in [0,1]$. The main problem to ensure that C_H is a copula is to guarantee the 2–increasingness of C_H, which depends both on C and H.

In general, if C is a singular copula, the function $H \neq 0$ cannot be absolutely continuous. Similarly, if C is an absolutely continuous copula, H cannot be singular. Therefore, as a special case of the perturbation (7), one can deal with perturbation related to functions $f, g : [0,1] \to [0,1]$ and constant $\lambda \in \Re$ in the form already discussed in Mesiar *et al.*[11], namely

$$C_{\lambda,f,g}(u,v) = \max\left(0,\ C(u,v) + \lambda\,C(f(u),f(v))\right). \tag{7}$$

Obviously, FGM family given in (5) can be seen as a special case of construction (6), considering $C = \Pi$, $\lambda \in [-1,1]$ and $f = g$ given by $f(x) = x - x^2$. Note that as a necessary condition to ensure that $e = 1$ is a neutral element of $C_{\lambda,f,g}$, one should consider $f(1) = g(1) = 0$. On the other hand, if $\lambda \le 0$, then $C_{\lambda,f,g}$ is always grounded. However, if $\lambda > 0$, then one should consider $C(f(0), g(0)) = 0$ (which is trivially satisfied for any copula C if $f(0) = g(0) = 0$).

In the case when no truncation is necessary, we have two general results.

Proposition 1. *Let $C : [0,1]^2 \to [0,1]$ be a copula and $H : [0,1]^2 \to [0,1]$ be a function so that $C + H \ge 0$ and C_H is a copula, i.e., $C_H = C + H$ is a copula. Then also $C_{\lambda\,H} = C + \lambda\,H$ is a copula for each $\lambda \in [0,1]$.*

Proposition 2. *Under the constraints of Proposition 1, the function $\hat{C}_{\bar{H}}$ is a copula, where $\hat{C} : [0,1]^2 \to [0,1]$ is the survival copula related to C,*

$$\hat{C}(u,v) = u + v - 1 + C(1-u, 1-v),$$

and $\bar{H} : [0,1]^2 \to [0,1]$ *is given by*

$$\bar{H}(u,v) = H(1-u, 1-v).$$

We have a next perturbation method valid for any copula C.

Theorem 1. *Let* $C : [0,1]^2 \to [0,1]$ *be a copula and define* $H_\lambda^C : [0,1]^2 \to [0,1], \lambda \in [0,1]$ *by*

$$H_\lambda^C(u,v) = \lambda\,(u - C(u,v))\,(v - C(u,v)).$$

Then $C_{H_\lambda^C} : [0,1]^2 \to [0,1]$ *given by*

$$C_{H_\lambda^C}(u,v) = C(u,v) + H_\lambda^C(u,v) \tag{8}$$

is a copula for each $\lambda \in [0,1]$ *and any copula* C.

4 Application to real data modelling

4.1 Real Estate Investment Trust

A REIT (**R**eal **E**state **I**nvestment **T**rust) is a company that mainly owns, and in most cases, operates income–producing real estate such as apartments, shopping centers, offices, hotels and warehouses. Some REITs also engage in financing real estate. The shares of many REITs are traded on major stock exchanges.

REIT Index Series is designed to present investors with a comprehensive family of REIT performance indexes that spans the commercial real estate space across the economy of the country. The index series provides investors with exposure to all investment and property sectors. In addition, the more narrowly focused property sector and sub–sector indexes provide the facility to concentrate commercial real estate exposure in more selected markets.

We have investigated the relations between 4 selected countries' (USA, Australia, Japan and UK) daily returns of the REIT (Real Estate Investment Trust) indexes in different time periods, determined by the recent global financial markets crises (July 1, 2008 – April 30, 2009) that can be also clearly identified from next Figure 1, presenting the parallel development of the considered REIT indexes.

We have performed filtering of the returns of all individual REIT indexes (in order to avoid a possible violation of the i.i.d. property) by ARMA–GARCH models (separately for the individual considered time sub-periods). Some basic descriptive statistical characteristics of 12 resulting time series are presented in Tables 2, 3 and 4 (mean values equal to 0 are in accordance to the expected properties of residuals of filtering).

The values of kurtosis for all 12 considered time series suggest that only the UK data during and after the crisis can be well fitted by Normal distributions

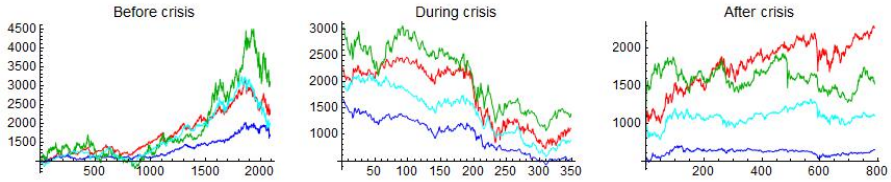


Fig. 1. Real Estate Investment Trust indexes in different time periods (USA = red, Australia = blue, Japan = green, UK = cyan)

Country	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
USA	0.00	0.71	0.07	4.31	-3.85	3.31
Australia	0.00	0.39	0.14	4.83	-2.12	1.82
Japan	0.00	0.03	0.18	3.59	-0.09	0.11
U.K.	0.00	0.02	0.14	4.95	-0.07	0.08

Table 2. Descriptive statistics for filtered returns of the REIT indexes before crisis

Country	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
USA	0.00	0.05	0.08	5.29	-0.23	0.24
Australia	0.00	0.08	-0.55	5.01	-0.49	0.26
Japan	0.00	0.05	-0.23	5.08	-0.29	0.16
U.K.	0.00	0.03	0.05	3.27	-0.10	0.11

Table 3. Descriptive statistics for filtered returns of the REIT indexes during crisis

Country	Mean	Standard deviation	Skewness	Kurtosis	Min	Max
USA	0.00	0.02	0.98	7.18	-0.11	0.44
Australia	0.00	0.01	0.90	6.97	-0.04	0.21
Japan	0.00	0.02	0.31	7.23	-0.12	0.29
U.K.	0.00	0.02	0.01	3.25	-0.08	0.08

Table 4. Descriptive statistics for filtered returns of the REIT indexes after crisis

(with the theoretical value of kurtosis equal to 3). This intuitive guess has been justified by the results of Jarque-Bera GoF test (see e.g. [8]) applied to all 12 time series. For fitting the remaining 10 time series, we utilized the Logistic and Laplace classes of distributions that have larger theoretical values of kurtosis.

Resulting types of distributions and their parameters for all 12 time series together with p-values are shown in Table 5, 6 and 7.

For all three time sub-periods and all couples of (filtered) returns of the REIT indexes we have performed the non–parametric correlation analyses based on the Kendall coefficients (see Table 8, 9 and 10). We have observed that the values of the correlation coefficients have dropped substantially between the first and the second considered time sub-periods and even more dramatically

Country	Type of distribution	Parameters	p–value
USA	Logistic	$\mu = 0,\ \beta = 0.3908$	0.97
Australia	Logistic	$\mu = 0,\ \beta = 0.2123$	0.49
Japan	Logistic	$\mu = 0,\ \beta = 0.0155$	0.21
U.K.	Logistic	$\mu = 0,\ \beta = 0.0082$	0.18

Table 5. Marginal distributions for filtered returns of the REIT indexes before crisis

Country	Type of distribution	Parameters	p–value
USA	Logistic	$\mu = 0,\ \beta = 0.0288$	0.09
Australia	Logistic	$\mu = 0,\ \beta = 0.0427$	0.11
Japan	Logistic	$\mu = 0,\ \beta = 0.0297$	0.33
U.K.	Normal	$\mu = 0,\ \sigma = 0.0344$	0.45

Table 6. Marginal distributions for filtered returns of the REIT indexes during crisis

Country	Type of distribution	Parameters	p–value
USA	Laplace	$\mu = 0,\ \beta = 0.0142$	0.12
Australia	Laplace	$\mu = 0,\ \beta = 0.0097$	0.16
Japan	Laplace	$\mu = 0,\ \beta = 0.0120$	0.15
U.K.	Normal	$\mu = 0,\ \sigma = 0.0247$	0.44

Table 7. Marginal distributions for filtered returns of the REIT indexes after crisis

for the third sub-period. These changes are illustrated in the scatter plots (see Figure 2 – Figure 7).

before crisis	USA	Australia	Japan	UK
USA	1	0.994	0.731	0.737
Australia	0.994	1	0.727	0.738
Japan	0.731	0.727	1	0.609
UK	0.737	0.738	0.609	1

Table 8. The values of the Kendall's correlation coefficient τ for the pre–crisis period

during crisis	USA	Australia	Japan	UK
USA	1	0.301	0.267	0.306
Australia	0.301	1	0.535	0.397
Japan	0.267	0.535	1	0.378
UK	0.306	0.397	0.378	1

Table 9. The values of the Kendall's correlation coefficient τ for the crisis period

after crisis	USA	Australia	Japan	UK
USA	1	0.111	0.061	0.221
Australia	0.111	1	0.222	0.087
Japan	0.061	0.222	1	0.073
UK	0.221	0.087	0.073	1

Table 10. The values of the Kendall's correlation coefficient τ for the post–crisis period

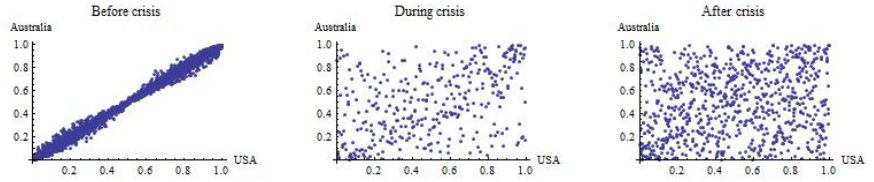


Fig. 2. USA & Australia

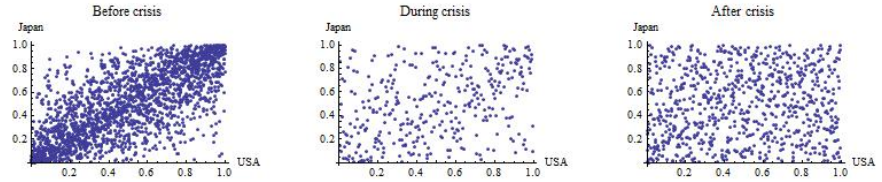


Fig. 3. USA & Japan

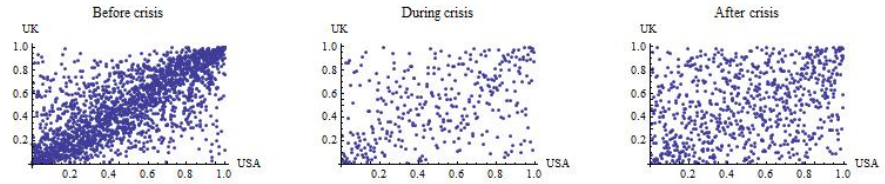


Fig. 4. USA & U.K.

We have applied the fitting by copulas to the residuals of ARMA–GARCH filters. We considered models from strict Archimedean copulas (Joe C^J, Frank C^F, Clayton C^{Cl} and Gumbel C^G) families and their mixtures with corresponding survival copulas $\hat{C}$ as well as their perturbations given by (8). We also tried the Farlie–Gumbel–Morgenstern (FGM) and Ali–Mikhail–Haq (AMH)

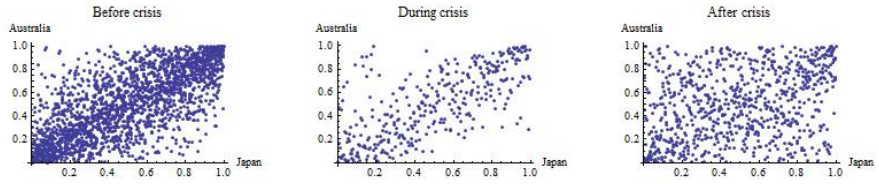


Fig. 5. Japan & Australia

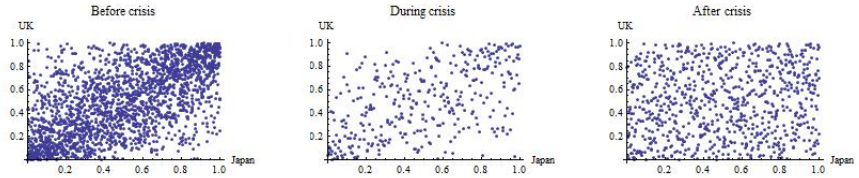


Fig. 6. Japan & U.K.

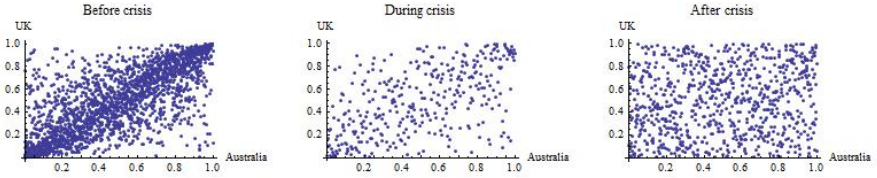


Fig. 7. Australia & U.K.

copulas, but these had the greatest values of the AD for all the pairs and time periods.

For estimation of parameters for each type of models we have used the maximum pseudo–likelihood method. For selecting the optimal models we have applied the Kolmogorov – Smirnov – Anderson – Darling (for which we have used the abbreviation AD) test statistic (4). For all of them, the simulation based GoF test yielded p–value > 0.1. Overview of optimal types of copulas for all couples and all time sub–periods of the filtered returns of REIT indexes is in Table 11 – Table 16.

5 Concluding remarks

We have found suitable marginal distribution models for all considered time series of filtered returns of REIT indexes and considered time period in one of Normal, Logistic or Laplace classes of distributions. We can observe that (although there is no clear relation between the pseudo likelihood functions and AD criterion) for all considered 6 couples of (filtered) returns of REIT indexes in 3 time periods and for all 6 couples of considered models the best perturbed models have lower values of AD criterion than the best models in the corresponding non–perturbed model classes. Moreover, for a great majority (16/18) of the considered 18 couples of (filtered) returns of REIT indexes, the

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009–08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	1.99	4.42	x	1.31	2.56	x	1.05	8.81
$C^G + H_\lambda^{C^G}$	0.98	1.80	4.23	0.63	1.18	2.44	0.16	1.03	8.43
$0.5 * (C^G + \hat{C}^G)$	x	2.19	1.72	x	1.80	1.27	x	1.65	1.82
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.79	1.98	1.55	0.46	1.24	**1.03**	0.05	1.08	1.74
C^{Cl}	x	1.48	4.96	x	0.48	1.85	x	0.17	2.26
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.97	1.18	3.72	0.76	0.24	1.45	0.04	0.17	2.19
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	1.56	1.98	x	0.21	1.42	x	0.43	3.49
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.98	1.71	**1.52**	0.53	0.42	1.69	0.03	0.16	3.27
C^J	x	2.25	12.09	x	1.40	4.09	x	1.04	10.32
$C^J + H_\lambda^{C^J}$	0.97	1.94	9.98	0.83	1.15	3.03	0.22	1.02	9.45
$0.5 * (C^J + \hat{C}^J)$	x	2.72	2.07	x	2.56	1.63	x	4.01	1.48
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.99	2.56	1.97	0.65	1.30	1.27	0.23	1.13	**1.34**

Table 11. The overview of optimal types of copulas for the couple USA & Japan of the (filtered) returns of REIT indexes

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009 – 08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	2.21	3.75	x	1.42	3.54	x	1.27	6.05
$C^G + H_\lambda^{C^G}$	0.45	2.13	3.29	0.25	1.38	3.32	0.36	1.20	5.83
$0.5 * (C^G + \hat{C}^G)$	x	1.90	1.60	x	1.05	1.68	x	1.13	1.79
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.07	2.31	**1.47**	0.08	1.47	1.38	0.06	1.29	**1.58**
C^{Cl}	x	1.62	5.27	x	0.76	1.37	x	0.43	2.27
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.98	1.32	4.22	0.41	0.65	**1.22**	0.62	0.24	1.93
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	2.18	2.23	x	1.41	1.97	x	0.54	1.78
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.54	2.17	2.12	0.02	0.87	1.71	0.05	0.59	1.62
C^J	x	2.60	10.00	x	1.51	6.46	x	1.34	12.05
$C^J + H_\lambda^{C^J}$	0.97	2.29	7.75	0.77	1.31	5.33	0.68	1.18	9.21
$0.5 * (C^J + \hat{C}^J)$	x	2.91	2.25	x	1.40	1.91	x	1.16	1.77
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.61	2.96	2.16	0.09	1.72	1.81	0.14	1.45	1.65

Table 12. The overview of optimal types of copulas for the couple USA & UK of the (filtered) returns of REIT indexes

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009–08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	4.74	2.11	x	1.42	2.81	x	1.12	5.92
$C^G + H_\lambda^{C^G}$	0.95	4.18	1.92	0.08	1.41	2.49	0.06	1.11	5.58
$0.5 * (C^G + \hat{C}^G)$	x	7.03	1.67	x	1.27	1.57	x	1.04	2.05
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.97	6.64	**1.46**	0.05	1.44	**1.54**	0.05	1.14	1.98
C^{Cl}	x	5.94	4.07	x	0.59	1.95	x	0.22	1.53
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.98	5.63	3.94	0.68	0.40	1.56	0.10	0.19	**1.27**
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	11.08	1.58	x	0.56	1.88	x	0.80	2.19
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.98	11.96	1.54	0.21	0.71	1.73	0.04	0.28	2.16
C^J	x	7.92	4.15	x	1.15	6.18	x	1.55	4.65
$C^J + H_\lambda^{C^J}$	0.95	7.07	3.78	0.26	1.09	5.51	0.59	1.39	3.75
$0.5 * (C^J + \hat{C}^J)$	x	13.81	1.58	x	2.15	1.88	x	1.39	1.80
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.99	13.65	1.54	0.03	1.24	1.58	0.30	1.59	1.57

Table 13. The overview of optimal types of copulas for the couple USA & Australia of the (filtered) returns of REIT indexes

non–perturbed models corresponding to the optimal perturbed ones attain the minimal values of the AD criterion among all considered non–perturbed classes of models for the (filtered) returns of REIT indexes (for 2 remaining couples

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009–08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	1.97	3.77	x	1.97	3.70	x	1.28	4.54
$C^G + H_\lambda^{C^G}$	0.97	1.77	3.68	0.93	1.78	3.14	0.08	1.27	4.42
$0.5 * (C^G + \hat{C}^G)$	x	2.29	1.76	x	2.39	1.45	x	1.13	1.71
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.89	1.91	**1.52**	0.44	2.01	**1.32**	0.05	1.31	1.61
C^{Cl}	x	1.41	4.95	x	1.34	2.31	x	0.46	2.20
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.95	1.11	3.65	0.98	1.04	1.93	0.45	0.33	1.76
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	1.32	1.85	x	1.56	1.55	x	1.27	1.77
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.97	1.65	1.76	0.76	1.75	1.47	0.05	0.60	1.65
C^J	x	2.21	11.53	x	2.25	12.41	x	1.36	7.35
$C^J + H_\lambda^{C^J}$	0.98	1.90	10.05	0.96	1.94	10.84	0.53	1.22	6.28
$0.5 * (C^J + \hat{C}^J)$	x	3.06	1.81	x	2.73	1.55	x	1.34	1.35
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.98	2.51	1.74	0.81	2.60	1.48	0.03	1.51	**1.18**

Table 14. The overview of optimal types of copulas for the couple Japan & Australia of the (filtered) returns of REIT indexes

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009–08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	1.66	5.01	x	1.53	2.67	x	1.07	4.27
$C^G + H_\lambda^{C^G}$	0.98	1.46	4.66	0.85	1.36	2.57	0.23	1.0.	4.14
$0.5 * (C^G + \hat{C}^G)$	x	1.64	1.54	x	1.49	1.35	x	1.08	1.86
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.55	1.64	**1.37**	0.48	1.49	**1.23**	0.05	1.09	1.79
C^{Cl}	x	1.01	4.74	x	0.88	1.86	x	0.16	1.40
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.97	0.71	3.07	0.93	0.61	1.24	0.02	0.15	**1.35**
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	1.64	1.74	x	1.49	1.62	x	1.09	2.01
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.71	1.16	1.58	0.79	0.76	1.54	0.03	0.18	1.94
C^J	x	1.83	5.99	x	1.66	4.50	x	1.07	5.10
$C^J + H_\lambda^{C^J}$	0.97	1.53	2.81	0.99	1.37	4.09	0.30	1.02	4.52
$0.5 * (C^J + \hat{C}^J)$	x	1.16	1.83	x	0.76	1.53	x	0.18	1.56
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.83	1.97	1.63	0.88	1.61	1.38	0.02	1.14	1.47

Table 15. The overview of optimal types of copulas for the couple Japan & U.K. of the (filtered) returns of REIT indexes

Type of copula	03.01.2000–31.07.2008			01.08.2008–30.04.2009			01.05.2009–08.05.2012		
	λ	θ	**AD**	λ	θ	**AD**	λ	θ	**AD**
C^G	x	2.26	3.94	x	1.59	2.78	x	1.08	2.98
$C^G + H_\lambda^{C^G}$	0.24	2.22	3.57	0.77	1.43	2.07	0.23	1.04	2.72
$0.5 * (C^G + \hat{C}^G)$	x	2.35	1.71	x	1.62	1.05	x	1.09	1.68
$0.5 * (C^G + \hat{C}^G) + H_\lambda^{0.5*(C^G+\hat{C}^G)}$	0.02	2.33	**1.67**	0.17	1.64	**0.93**	0.02	1.09	1.61
C^{Cl}	x	1.68	5.39	x	0.94	1.97	x	0.15	1.88
$C^{Cl} + H_\lambda^{C^{Cl}}$	0.98	1.38	4.44	0.98	0.66	1.65	0.24	0.08	**1.29**
$0.5 * (C^{Cl} + \hat{C}^{Cl})$	x	2.35	2.75	x	1.62	1.31	x	1.09	1.73
$0.5 * (C^{Cl} + \hat{C}^{Cl}) + H_\lambda^{0.5*(C^{Cl}+\hat{C}^{Cl})}$	0.42	2.26	2.63	0.72	0.89	1.08	0.02	0.21	1.57
C^J	x	2.68	5.22	x	1.73	6.08	x	1.10	3.69
$C^J + H_\lambda^{C^J}$	0.98	2.37	5.00	0.96	1.44	5.04	0.32	1.04	3.21
$0.5 * (C^J + \hat{C}^J)$	x	2.26	2.82	x	0.89	1.15	x	0.20	1.70
$0.5 * (C^J + \hat{C}^J) + H_\lambda^{0.5*(C^J+\hat{C}^J)}$	0.49	3.04	2.65	0.79	1.75	1.09	0.15	1.11	1.42

Table 16. The overview of optimal types of copulas for the couple U.K. & Australia of the (filtered) returns of REIT indexes

they narrowly exceed the minimum values). Moreover, for 17 of 18 considered couples of (filtered) returns of REIT indexes the θ coefficients of the optimal models do not considerably differ from the values of optimal models in the

corresponding classes of non–perturbed models. The only exception to this phenomenon could be attributed to very flat shapes of the respective pseudo likelihood functions around they minimum values.

Acknowledgement The support of the grants APVV–0073–10 and VEGA 1/0143/11 is kindly announced.

References

1. Amblard, C., Girard, S.: A new symmetric extension of FGM copulas. *Metrika* **70**, 1–17 (2009)
2. Anderson, T. W.: Anderson-Darling tests of goodness-of-fit. *International Encyclopedia of Statistical Science*, Springer (2011)
3. Anderson, T. W., Darling, D. A.: A Test of Goodness of Fit. *Journal of the American Statistical Association* Vol. **49**, No. 268, 765–769 (1954)
4. Bairamov, I., Kotz, S.: Dependence structure and symmetry of Huang-Kotz FGM distributions and their extensions. *Metrika* **56**, 55–72 (2002)
5. Balakrishnan, N.: *Handbook of the Logistic Distribution*. Marcel Dekker, New York (1992)
6. Berg, D., Bakken, H.: Copula Goodness-of-fit Tests: A Comparative Study. In: Working paper, University of Oslo and Norwegian Computing Center (2006)
7. Genest, C., Favre, A.C.: Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12**, 347–368 (2007)
8. Jarque, C.M., Bera, A.K.: A test for normality of observations and regression residuals. *International Statistical Review* **55** (2), 163-172 (1987)
9. Joe, H.: *Multivariate models and dependence concepts*. London : Chapman and Hall (1997)
10. Kotz, S., Kozubowski, T. J., Podgórski, K.: *The Laplace distribution and generalizations: a revisit with applications to Communications, Economics, Engineering and Finance*. Birkhauser. pp. 23 (Proposition 2.2.2, Equation 2.2.8) (2001)
11. Mesiar, R., Komorník, J., Komorníková, M.: On some construction methods for bivariate copulas. *Advances in Intelligent Systems and Computing*, **228**. Aggregation Functions in Theory and in Practise : Proceedings of the 7th International Summer School on Aggregation Operators at the Public University of Navarra, Pamplona, Spain, 16.-20.6.2013, 39–46 (2013)
12. Mesiar, R., Komorník, J., Komorníková, M.: Modification of bivariate copulas, *Fuzzy Sets and Systems* (2014), http://dx.doi.org/10.1016/j.fss.2014.04.016
13. Nelsen, R.B.: *An introduction to copulas*. Second Edition. Springer Series in Statistics, Springer-Verlag, New York (2006)
14. Rodríguez–Lallena, J. A., Úbeda–Flores, M.: A new class of bivariate copulas. *Statistics & Probability Letters* **66**, 315–325 (2004)
15. Patton, A.J.: Modelling Asymmetric Exchange Rate Dependence. *International Economic Review*, 47, **2**, 527–556 (2006)
16. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–231 (1959)

Two-Stage Kalman Filtering for Discrete Systems Using Nonparametric Algorithms[1]

Gennady M. Koshkin and Valery I. Smagin

Tomsk State University, Tomsk, Russia
Email: kgm@mail.tsu.ru
Email: vsm@mail.tsu.ru

Abstract: The paper addressed the filtering problem with using nonparametric algorithms for discrete stochastic systems with unknown input. The two-stage algorithm on the base of the Kalman filtering and nonparametric estimator for systems with unknown input is designed and explored. Examples are given to illustrate the usefulness of the proposed results in comparison with the known algorithms.
Keywords: Kalman filter, Unknown input, Two-stage filtering algorithm, Nonparametric estimator.

1 Introduction

An important issue of the Kalman filtering [1] is construction of algorithms for the class of systems with unknown additive perturbations. Such systems are used as the models of real physical systems, as the models of objects with unknown errors, and in control problems for economic systems. The known methods to calculate estimates of a state vector are based on the algorithms of estimation of an unknown perturbation [2-9].

In this paper, for discrete systems with unknown perturbations the two-stage optimal filtering with use of nonparametric estimators for unknown input are proposed. Examples are given to illustrate the properties of the proposed procedures in comparison with the known algorithms.

2 The problem statement

Consider the mathematical model of the linear discrete-time stochastic system with unknown input in the form:

$$x(k+1) = Ax(k) + Br(k) + q(k),\ \ x(0) = x_0, \qquad (1)$$

[1]Supported by Russian Foundation for Basic Research, projects 13-08-00744, 13-08-01015A, Tomsk State University Competitiveness Improvement Program, and the project "Goszadanie Minobrrnauki Rossii"

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

$$y(k) = Hx(k) + v(k)\,, \tag{2}$$

where $x(k)$ is the state of the object, $r(k)$ is an unknown input, $y(k)$ is the measurement, A, B, and H are matrices of the appropriate dimensions. It is assumed random perturbations $q(k)$ and the noise measurements $v(k)$ are not correlated between themselves and are subject to the Gaussian distribution with zero mean and the corresponding covariance: $\mathrm{E}[q(k)q(k)^{\mathrm{T}}] = Q\delta(k,t)$, $\mathrm{E}[v(k)v(k)^{\mathrm{T}}] = V\delta(k,t)$, where $\delta(k,t)$ is the Kronecker symbol, i.e., $\delta(k,t) = 1$ if $k = t$, and $\delta(k,t) = 0$ when $k \neq t$. It is also proposed that the vector of initial conditions is uncorrelated with values $q(k)$ and $v(k)$. This vector is defined by the following characteristics:

$$\mathrm{E}[x(0)] = \bar{x}_0\,,\ \mathrm{E}[(x(0) - \bar{x}_0)(x(0) - \bar{x}_0)^{\mathrm{T}}] = P_0\,.$$

3 The estimation algorithm of an unknown input and state space vector

In this paper, the optimal filter is defined by the following full-order Kalman filter. Filter equations have the form:

$$\hat{x}(k+1) = A\hat{x}(k) + B\hat{r}(k) + K(k)[y(k+1) - H(A\hat{x}(k) + B\hat{r}(k))]\,,\ \hat{x}(0) = \bar{x}_0\,, \tag{3}$$

$$P(k+1/k) = AP(k)A^{\mathrm{T}} + Q\,, \tag{4}$$

$$K(k) = P(k+1/k)H^{\mathrm{T}}[HP(k)H^{\mathrm{T}} + V]^{-1}\,, \tag{5}$$

$$P(k+1) = (I - K(k)H)P(k+1/k)\,,\ P(0) = P_0\,, \tag{6}$$

where $\hat{x}(k)$ and $\hat{r}(k)$ are estimators, $P(k) = \mathrm{E}[(x(k) - \hat{x}(k))(x(k) - \hat{x}(k))^{\mathrm{T}}]$.

However, formulas (3)–(6) can not be applied immediately because $\hat{r}(k)$ is unknown. Obtain estimator $\hat{r}(k)$ by making use of the following criteria:

$$J(r(k-1)) = \mathrm{E}\left[\sum_{i=1}^{k} \|u(i)\|_C^2 + \|r(i-1)\|_D^2\right], \tag{7}$$

where $u(i) = y(i) - H\hat{x}(i)$ is the innovation process, $\|\cdot\|_C^2$ is the Euclidian norm, C and D are symmetric positive definite weight matrices.

Optimal estimator of the unknown input at moment $k = 1$ is found by minimization of the criteria:

$$J(r(0)) = \min_{r(0)} \mathrm{E}\left[\|y(1) - H\hat{x}(1)\|_C^2 + \|r(0)\|_D^2\right]. \tag{8}$$

Substituting $\hat{x}(1) = A\hat{x}(0) + Br(0)$ into (8), we have

$$J(r(0)) = \min_{r(0)} \mathrm{E}\left[\left\| y(1) - HA\hat{x}(0) - HBr(0) \right\|_C^2 + \left\| r(0) \right\|_D^2 \right]. \quad (9)$$

Transform the norms in (9) and obtain

$$J(r(1)) = \min_{r(0)} \mathrm{E}\left[\alpha_0 - 2r(0)^{\mathrm{T}} B^{\mathrm{T}} H^{\mathrm{T}} V (y(1) - HA\hat{x}(0)) + \left\| r(0) \right\|_{B^{\mathrm{T}} H^{\mathrm{T}} CHB + D}^2 \right]. \quad (10)$$

Here, the parameter α_0 does not depend on $r(0)$. First, differentiate (10) w.r.t. $r(0)$, and then find the optimal estimator of the unknown input from the equation

$$\frac{dJ(r(0))}{dr(0)} = 2(B^{\mathrm{T}} H^{\mathrm{T}} CHB + D) r(0) - 2B^{\mathrm{T}} H^{\mathrm{T}} C \, \mathrm{E}[y(1) - HA\hat{x}(0)] = 0. \quad (11)$$

So, at the moment $k = 1$, we obtain the optimal estimator of the unknown input:

$$\hat{r}(0) = S\mathrm{E}[y(1) - HA\hat{x}(0)], \quad (12)$$

where

$$S = (B^{\mathrm{T}} H^{\mathrm{T}} CHB + D)^{-1} B^{\mathrm{T}} H^{\mathrm{T}} C. \quad (13)$$

Analogously, at the moment $k = 2$, the optimal estimator of the unknown input is found from the following criteria:

$$J(r(1)) = \min_{r(1)} \mathrm{E}\left[\left\| y(2) - H\hat{x}(2) \right\|_C^2 + \left\| r(1) \right\|_D^2 \right] + J(\hat{r}(0)). \quad (14)$$

Taking into account (14) and the expression $\hat{x}(2) = A\hat{x}(1) + Br(1)$ at the moment $k = 2$, we have

$$J(r(1)) = \min_{r(1)} \mathrm{E}\left[\left\| y(2) - HA\hat{x}(1) - HBr(1) \right\|_C^2 + \left\| r(1) \right\|_D^2 \right] + J(\hat{r}(0)).$$

As in the case of (10)

$$J(r(1)) = \min_{r(1)} \mathrm{E}\left[\alpha_1 - 2r(1)^{\mathrm{T}} B^{\mathrm{T}} H^{\mathrm{T}} C (y(2) - HA\hat{x}(1)) + \left\| r(1) \right\|_{B^{\mathrm{T}} H^{\mathrm{T}} CHB + D}^2 \right], \quad (15)$$

where the value α_1 does not depend on $r(1)$. Differentiating (15) w.r.t. $r(1)$, as in the first step, we obtain the optimal estimator:

$$\hat{r}(1) = S\mathrm{E}[y(2) - HA\hat{x}(1)]. \quad (16)$$

Using the mathematical induction, for the next steps

$$\hat{r}(k) = S\mathrm{E}[w(k)]. \quad (17)$$

Here, the matrix S is given by the formula (13), and $w(k) = y(k) - HA\hat{x}(k-1)$.

Now, let us calculate value $\mathrm{E}[w(k)]$ using nonparametric estimators [10]. Applying the well known kernel estimates, we obtain

$$\hat{r}(k) = S\hat{w}_{np}(k), \tag{18}$$

where the j component of the vector takes the form:

$$\hat{w}_{np,j}(k) = \frac{\sum_{i=1}^{k} w_j(i) K_j\left(\frac{k-i+1}{h_{i,j}}\right)}{\sum_{i=1}^{k} K_j\left(\frac{k-i+1}{h_{i,j}}\right)}. \tag{19}$$

In formula (19), $K_j(\cdot)$ is a kernel function, $h_{i,j}$ is a bandwidth parameter. We use the Gaussian kernels, and the bandwidths calculated by the cross-validation method [11].

4. Simulations

Apply the filtering algorithm using nonparametric estimates, i.e., (3)–(6) and (18), to the model of the second order (1) and to the observations (2) with the parameters:

$$A = \begin{pmatrix} 0 & 1 \\ 0.05 & 0.9 \end{pmatrix},\ B = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix},\ r = \begin{pmatrix} 5 \\ 2 \end{pmatrix},\ Q = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.02 \end{pmatrix},$$
$$V = \begin{pmatrix} 0.8 & 0 \\ 0 & 1.2 \end{pmatrix},\ H = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix},\ P_0 = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix},$$
$$C = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix},\ D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ \overline{x}_0 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}.$$

By the simulations, the proposed algorithms are compared with the algorithms using the LSM estimates from [3, 4]. These comparisons are given in Figures 1–4:

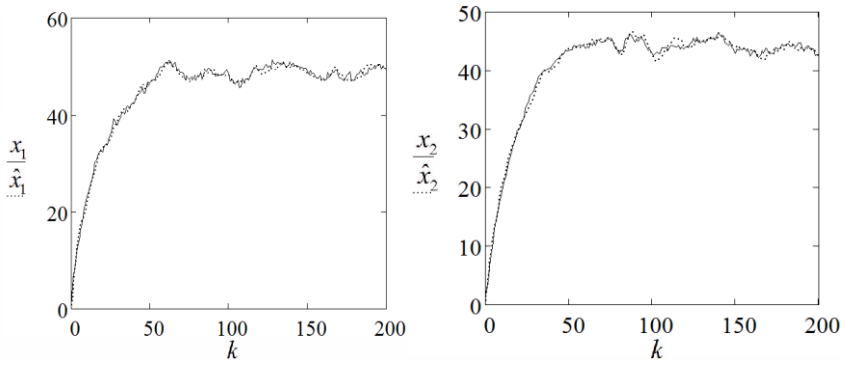


Fig. 1. The dependence on the components of state vectors and the nonparametric estimates of these components (3)–(6), (18)

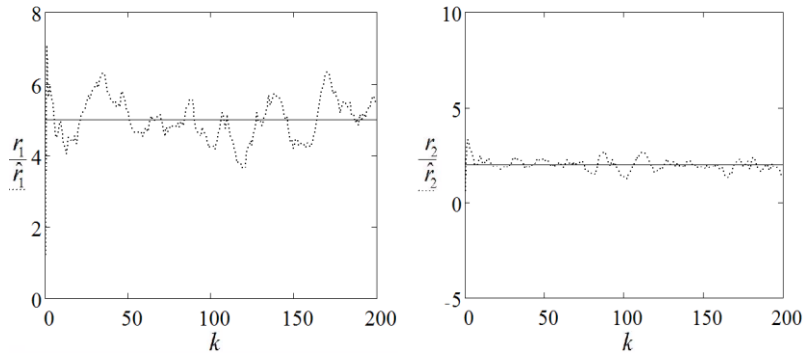


Fig. 2. The estimation of the unknown inputs by nonparametric algorithms (3)–(6), (18)

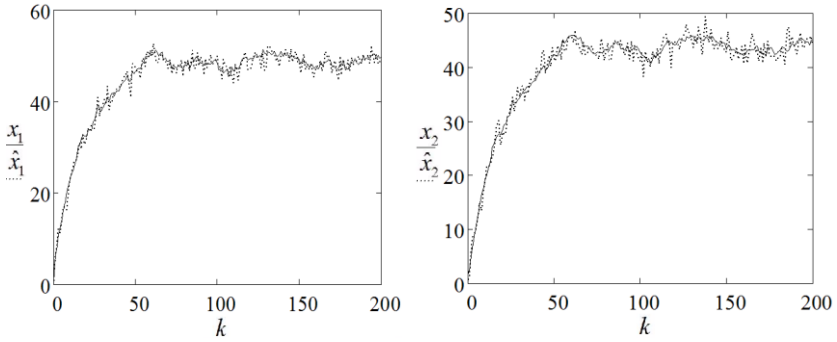


Fig. 3. The dependence on the components of state vectors and the LSM estimates of these components from the papers [3, 4]

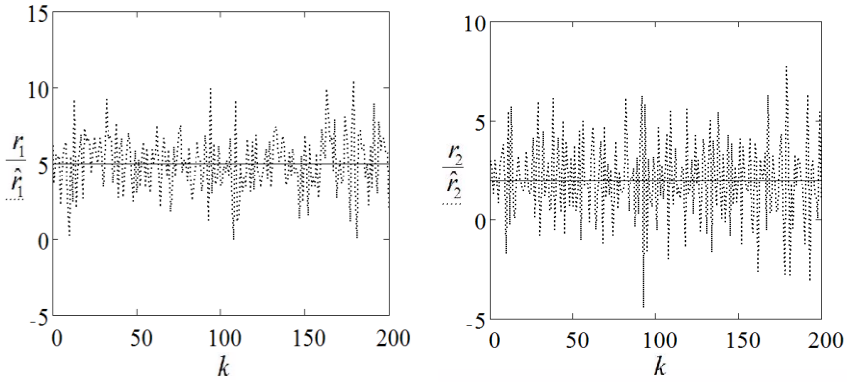


Fig. 4. The estimation of the unknown inputs by the LSM estimates from the papers [3, 4]

In Tables 1 and 2, the standard errors of estimation

$$\sigma_{x,i} = \frac{\sqrt{\sum_{k=1}^{N} (x_i(k) - \hat{x}_i(k))^2}}{N-1},$$

$$\sigma_{r,i} = \frac{\sqrt{\sum_{k=1}^{N} (r_i(k) - \hat{r}_i(k))^2}}{N-1}, \; i = \overline{1,2},$$

are given for two filtering algorithms ($N = 200$) by averaging 50 realizations.

Table 1. Standart Errors for Filtering Algorithm with Using Nonparametric Estimates

$\sigma_{x,1}$	$\sigma_{x,2}$	$\sigma_{r,1}$	$\sigma_{r,2}$
0.885	0.945	0.751	0.449

Table 2. Standart Errors for Filtering Algorithms with Using the LSM-estimates

$\sigma_{x,1}$	$\sigma_{x,2}$	$\sigma_{r,1}$	$\sigma_{r,2}$
1.348	1.514	2.014	2.082

5 Conclusion

In this paper, we deal with two-step algorithm of the Kalman filtering for systems with unknown input. The proposed method has been verified by the simulations study. Figures show that the filtering procedures, using nonparametric estimates, have the advantages in the accuracy in comparison with the known algorithms using LSM-estimates (cf. Fig. 1 and 3, Fig. 2 and 4, Table 1 and 2).

References

1. Kalman, R.E. and Busy, R. A new results in linear filtering and prediction theory. *Trans. ASME J. Basic Engr.* 83, 95–108 (1961)
2. Astrom, K. and Eykhoff, P. System identification – A survey. *Automatica*. **7**, 123-162 (1971)
3. Janczak, D. and Grishin, Y. State Estimation of Linear Dynamic System with Unknown Input and Uncertain Observation Using Dynamic Programming. *Control and Cibernetics*. **35**, 4, 851–862 (2006)

4. Gillijns, S. and Moor, B. Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica*. **43**, 111–116 (2007)
5. Hsien, C.-S. On the Optimal of Two-Stage Kalman Filter for Systems whis Unknown Input. *Asian Journal of Control*, **12**, 4, 510–523 (2010)
6. Darouach, M., Zasadzinski, M. and Xu S. J. Full-order observers for linear systems with unknown inputs. *IEEE Trans. on Automat. Contr*. **39**, 606–609 (1994)
7. Hou, M., Patton, R. Optimal filtering for systems with unknown inputs. *IEEE Trans. on Automat. Contr*. **43**, 445–449 (1998)
8. Smagin, S.V. Filtering in linear discrete systems with unknown perturbartion. *Optoelectronic, Instrumentation and Data Processing*. **45**, 6, 513–519 (2009)
9. Witczak, M. Fault diagnosis and fault-tolerant control strategies for non-linear systems. Chapter 2. Unknown input observers and filters. *Lecture Notes in Electrical Engineering. Springer International Publishing, Switzerland*, 9-56 (2014)
10. Dobrovidov, A., Koshkin, G., and Vasiliev, V. Non-Parametric State Space Models. *Heber, UT 84032, USA. Kendrick Press, Inc.* (2012)
11. Leung, D. Cross-validation in nonparametric regression with outliers. *Annals of Statistics.* **33,** 2291–2310 (2005)

Monte-Carlo Reliability Evaluation of the Ring Detector based on Heavily Masked Normalized Correlation

Samuel Kosolapov

Signal and Image Processing Laboratory, ORT Braude Academic College of Engineering, Karmiel, Israel
Email: ksamuel@braude.ac.il

Abstract: A number of general-purpose ring and circle detectors are known. In most cases, template matching and Hough transform are used to detect rings inside the image. However, ring detectors described in the literature were found impractical for the real-life implementation of the camera-based Instant Feedback System (IFS). Goal of the IFS is to collect answers of the students to the multiple-choice questions during the lecture. In the frames of the camera-based IFS, students answer to the specific multiple-choice question by presenting to the camera a specially designed IFS cards. Image of the class contains plurality of IFS cards in the different orientations and of different sizes, which makes recognition non-trivial. To simplify recognition, preferred design of IFS card contains bounding black ring and some other IFS specific elements positioned inside the bounding ring. IFS cards in the periphery of the real-life image are geometrically distorted, making standard template match approach too slow and non-reliable. To cope with this problem, standard Normalized Correlation template-matching algorithm was modified by adding the mask hiding the IFS elements inside the ring. In this case the number of templates needed to isolate IFS card is significantly smaller. In order to evaluate reliability of the proposed algorithm, special software Monte-Carlo simulator was created. Monte-Carlo simulation results show that in case of non-overlapped cards recognition error is less than 1%, which can be considered as adequate for the real-life camera-based IFS. Developed approach can be used to speed-up recognition in the other practically interesting cases, for example, for the traffic signs recognition.

Keywords: Image Processing, Ring Detector, Normalized Correlation, IFS, Monte-Carlo simulation

1. Introduction

In many practically important application there is a need to find rings (or circles) in the image. A number of general-purpose ring and circle detectors are known. In most cases modifications of template maching algorithms and Hough

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

Transform algorithms are used. Those approches are implemented in a number of popular libraries and software packages. For example, MATLAB contains *"imfindcircles"* function to automatically detect circles or circular objects in an image. This function requres a radius range in pixels to search for the circles and a number of "sensitivity" parameters. This function implements two different methods. Using "two-stage method" enables to detects parts of the circles, so that overlapping circular objects can be detected. Additional option in MAPTLAB is to use *"CircularHough_Grd"* based on "Circular Hough Transform" [1]. As in the previous case, a range of radii and other "sensitivity" parameters must be specified. In case ellipse must be found, modifications of the "Randomized Hough Transform" [2] based on original algorithm [3] can be used. Popular "OpenCV" library contains function *"HoughCircles"* [4]. This library can be used to create PC, Android and iPhone real-time application.

However, ring detectors described in the above examples were found impractical for the real-life implementation of the camera-based Instant Feedback System (IFS). Goal of the IFS is to collect answers of the students to the multiple-choice questions during the lecture. In the frames of the camera-based IFS, students answer to the specific multiple-choice question by presenting to the camera a specially designed IFS cards [5]. Photo of the class contains plurality of IFS cards images in the different orientations and of different sizes, which makes recognition and analysis non-trivial.

2. IFS Card Design

To simplify recognition and analysis of the IFS cards, preferred design of IFS card contains bounding black ring and some other IFS specific elements.

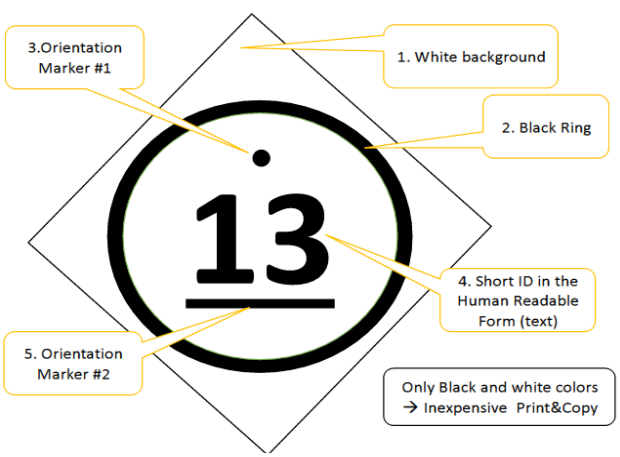


Fig. 1. IFS Card Design

Figure 1 presents IFS card design described in this work. IFS Card is printed on the thick white paper by using any available black and white printer. Background of the card is white (1). The card has a black ring (2), orientation markers (3 and 5) and two-digit number in the human-readable form (4). This number designates short ID of the specific student (for example, number of the student in the class list). Orientation of the IFS card specifies number of the selected answer. Orientation markers (3 and 5) prevents orientation ambiguity for the numbers like 66-99. Described design does not contains color element, so that set of cards can be copied using standard copy-machine.

3. Steps in the IFS Card recognition and analysis

Human observer analyzing the image of the class easily recognizes bounding black ring and the number inside the ring for any possible orientation. For the computer the problem of recognition of the plurality of the IFS cards is not trivial, because orientation and sizes of the bounding rings and digits are different. Direct template-matching approach is possible, but time consuming, because a very big number of templates (having different digits in the different sizes and in the different orientations) must be used. For this specific IFS design, the search can be executed faster. On the first step, only bounding rings (of different sizes) are to be found. Then sub images inside the bounding rings are to be scaled to the "standard size". On the second step markers inside sub-images are to be used to evaluate specific IFS card orientation and rotate its digits to the "standard position". On the third step OCR or direct template match algorithm can be used to recognize two digits of the "standard size" and in the "standard orientation".

4. Ring Detector based on Heavily Masked Normalized Correlation

Very popular and practical template matching algorithm widely used in the Image Processing is "2D Normalized Correlation":

$$R[row,col] = \frac{\sum_{y=0}^{m-1}\{\sum_{x=0}^{n-1}(T[row,col]-\bar{T})(I[row+y,col+x]-\bar{I})\}}{\sqrt{\sum_{y=0}^{m-1}\{\sum_{x=0}^{n-1}(T[row,col]-\bar{T})^2(I[row+y,col+x]-\bar{I})^2\}}}$$

High value of the R (close to 1.0) means that template T is found inside the image I starting from [row, col]. Normalization is needed to make recognition invariant to the brightness variations.

Unfortunately, direct implementation of this algorithm in our case is not practical, because of markers and digits inside the bounding ring. To cope with this problem, standard Normalized Correlation template-matching algorithm was modified by adding the mask hiding the markers and digits inside the ring.

Fig. 2. Mask used to find the ring

Figure 2 presents green region that is to be used to find bounding black ring. Pixels outside the green ring are to be excluded from the sums in the Normalized Correlation equation. Number of pixels to be used for the calculation of the sums is significantly smaller.

Despite idea of masking looks simple and nearly obvious, exact analog was not found in the literature.

5. Monte-Carlo Simulator

In order to evaluate reliability of the proposed algorithm, special software Monte-Carlo simulator was created as Windows Forms Desktop C# .NET application.

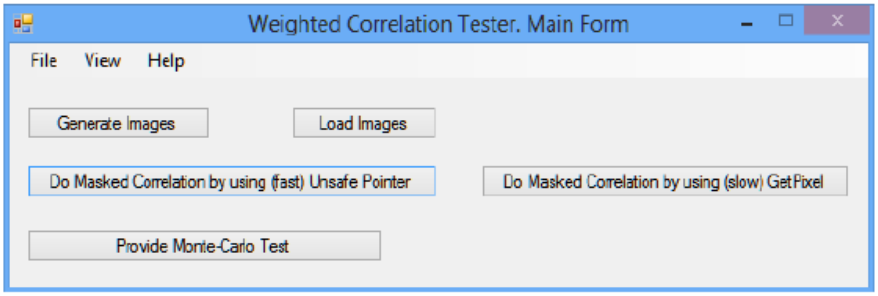


Fig. 3. Appearance of the Monte-Carlo software simulator

Figure 3 presents appearance of the simulator. Pressing the button “Generate Images” call “Pattern Generator Form” (see Figure 4). Operator can specify a background image (for example, image of the real class), number of IFS cards to place on this background image in the pseudo-random sizes, positions and orientations. Additional geometrical parameters of the IFS cards, noise level and some others (like level of geometrical distortions, level of cards overlap) can be specified.

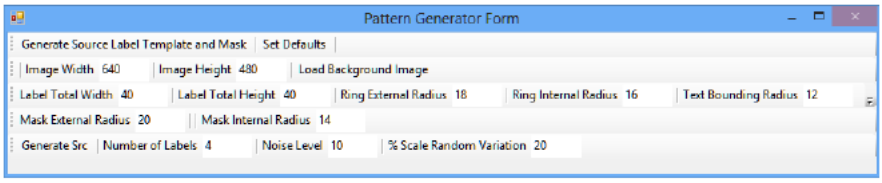


Fig. 4. Pattern Generator Form

Pressing button “Generate Src” creates a number of images: “Label”, “Template”, “Mask” (see Figure 5) and resulted synthetic image (see Figure 6)

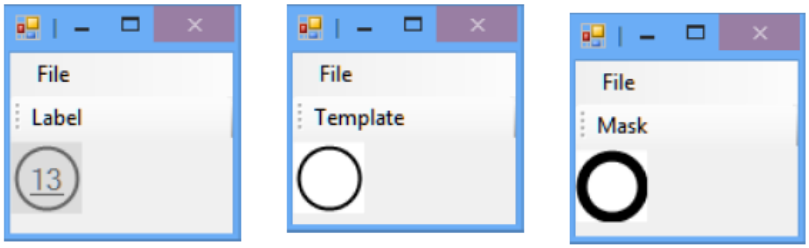


Fig. 5. Label, Template, Mask.

Label, Template and Mask can be stored to files and loaded from files.

Second part of the Monte-Carlo simulator attempts to recognize IFS cards in the image in test. Currently, two implementations of the Heavily Masked Normalized Correlations are supported: by using unsafe pointer and by using GetPixel function. Obviously, implementation with unsafe pointer is nearly 100 times faster: VGA size image was processed during 5 seconds. Typical results of the processing are presented in the Figure 7. Current utility deals with ring recognition only because other recognitions steps are trivial. Button “Provide Monte-Carlo Tests” provides long series of tests {create image – process image} while collecting recognition success rate.

Fig. 6. Resulted synthetic image

Fig. 7. Results of Heavily Masked Normalized Correlation algorithm: Region of the Higest Correlation value (~ 0.99) marked by red cross (Label #10).

Conclusions

Monte-Carlo simulation results provided on synthetic images show that in case of non-overlapped cards circles detection success is close to 99%, which can be considered as adequate for the real-life camera-based IFS, because, practically, processed image of the class, with recognized cards marked as "green", must be presented to the lecturer for the final approval. By visually inspecting image of the class, human observer (lecturer) will easily reveal non-recognized (and/or cards recognized in the wrong way) cards, and manually correct the final grades list. Despite the need of this manual inspection, time to get the final grades list still is fast enough (number of seconds) to consider all the process as Instant Feedback System. High detection rate can be achieved only in case that cards are not overlapped and in case that camera resolution is high enough to properly resolve elements of the IFS cards. Practically, for the camera with 16 Mpixel resolution, reliable IFS cards recognition is limited to the class of 20-30 students. In case of bigger class, a number of images must be obtained, which may be considered as not convenient or even not practical for the selected camera-based concept.

Future R&D and Applications

Current implementation was limited to the rings detection on synthetic images only. Next R&D will include evaluation of the IFS card orientation and OCR of the number (short ID) inside the ring as for the synthetic as for real class images. It can be expected, that Heavily Masked Normalized Correlation may be instrumental for the fast markers search inside the external ring. More, considering that practical number of short IDs is limited to 30, OCR algorithm can analyze only unique parts of the IDs by using properly selected masks. According to our preliminary evaluations, in this case, OCR speed can be increased at least by factor 3. Developed ring detection algorithm uses no third party libraries and thus can be ported to any platform (PC, Android, and iPhone) by using any modern programming language (C, C++, C#, Java, Python, etc.) Additionally, this algorithm can easily be implemented as web service or web/cloud application. In case of cloud implementation of the camera-cased IFS multiple-choice exam lecturer will grab the image of the class by using simple cloud application for the standard smartphone. Grabbed image (or images) will be immediately send to the cloud server for the proper image processing. Cloud image processing time may be very short. Additional advantage of the cloud approach is that no software installation is needed. Additionally, developed Heavily Masked Normalized Correlation approach can be used to speed-up recognition in the other practically interesting cases, for example, for the traffic signs recognition.

References

[1] Tao Peng. Detect circles with various radii in grayscale image via Hough Transform. . MATLAB Central, 2005

[2] McLaughlin and Robert. Randomized Hough Transform. Improved Ellipse Detection with Comparison. Technical Report JP98-01, 1998.

[3] Lei XU, Erkki OJA, Pekka Kultanena. A new curve detection method: Randomized Hough Transform. Pattern Recognition Letters, 11 331-338, 1990

[4] OpenCV Library http://opencv.org/

[5] S. Kosolapov, E. Gershikov and N. Sabag, "Feasibility of Camera-Based Instant Feedback System", *Think Mind*, Content 2013 Proceedings, 23-29, 2013

Discrete observation of a continuous time semi Markov model for HIV control

Zacharias Kyritsis[1] and Aleka Papadopoulou[2]

1.Department of Mathematics, Aristotle University of Thessaloniki, Greece (Email: zkyritsi@math.auth.gr)
2.Department of Mathematics, Aristotle University of Thessaloniki, Greece (Email: apapado@math.auth.gr)

The aim of the present paper is to review a continuous time semi Markov model for HIV control and to apply an algorithm for data simulation analysis which we run to provide the data for transitions and the sojourn times of the corresponding visited states. After the simulation process three different models are developed and validated for the discrete observation of the simulated continuous process and an estimation method is applied to get the respective distributions. Finally, the above results are illustrated numerically with synthesized data.
Keywords: *health care, semi Markov process, HIV*

1 Introduction

The definition of the non homogeneous semi Markov process is provided in Iosifescu-Manu [10] for the continuous time case, in Janssen and De Dominics [12] for the discrete case and in De Dominics and Manca [7]. Later on the definition of a non homogeneous semi Markov system in discrete time is provided in Vassiliou and Papadopoulou [27] and the asymptotic behavior of the same model is studied in Papadopoulou and Vassiliou [24]. Important theoretical results and applications for semi Markov models can be found in work of Cinlar [4], [5], [6], Teugels [26], Keilson [13], [14], McLean and Neuts [21], Howard [9], McCLean [17], [18], [19], [20], Limnios et al [15] and in Janssen [11].

In the present, we study the discrete observation of a continuous time semi Markov model for HIV control. In section 2, first we describe the semi Markov model for HIV control, review the continuous time case and then we provide the technique for the discrete observation of the reviewed model. In section 3, the technique described in Section 2 is illustrated with synthesized data derived by a simulation process. Last, conclusions from the previous results are provided.

2 Discrete observation of a continuous time semi Markov model for HIV control

The process of infection by HIV is characterized by two fundamental markers. The first is the viral load (VL) and the second CD4 lymphocyte. Hence, the history of the disease can be considered as a series of stages through which a

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

patient progresses. The first stage is called primary infection. And the corresponding symptoms vary to duration, that is twenty eight days in average and at least one week. At this stage there are no specific symptoms and often they are not recognized as signs of HIV infection. Even if a patient goes to a doctor or a hospital, he might be misdiagnosed. The second stage is called clinically asymptomatic stage lasts for an average of ten years and, as its name suggests, that is free from major symptoms. The HIV antibodies are detectable in the blood, so antibody tests will show a positive result. On the third stage called symptomatic HIV the lymph nodes and tissues are damaged because of the years of activity, HIV mutates and becomes more pathogenic, leading to more T helper cell destruction and the body fails to keep up with replacing the lost T helper cells. Antiretroviral treatment is usually started once an individual's CD4 index falls to a low level which is an indication that the immune system is deteriorating. Finally, on the fourth stage called progression for AIDS as the immune system becomes more and more damaged the individual may develop increasingly severe opportunistic infections and cancers, leading eventually to an AIDS diagnosis.

Using these markers and the above mentioned four stages we can describe the progress of HIV by a semi Markov model considering four health states (Tan [25]) as follows:

Primary infection → First stage: VL≤400 and CD4≤200
Asymptomatic stage → Second stage: VL≤400 and CD4>200
Symptomatic HIV → Third stage: VL>400 and CD4>200
Progression for AIDS → Fourth stage: VL>400 and CD4≤200

From the above, we can consider a non homogeneous semi Markov process with discrete and finite state space symbolized by S={1, 2, 3, 4}. The continuous time case for non homogeneous semi Markov systems is studied in Papadopoulou and Vassiliou [23]. The transition probability matrix of the embedded Markov chain is defined by $\mathbf{P}(s,t)=\{p_{ij}(s,t)\}_{i,j\in S}$, where $p_{ij}(s,t)$=prob{a patient selects state j for its next transition during (s,t) / entered state i at time s}and the holding time mass function matrix for the semi Markov process is defined by $\mathbf{H}(m)=\{h_{ij}(m)\}_{i,j\in S}$ where $h_{ij}(m)$=prob{a patient which entered state i at its last transition to hold for m time in state i before making its next transition given that state j has been selected}.

Also, we define by :

$$^{\leq}w_i(s,t) = \text{prob}\{\text{ a patient which entered state i at time s holds for time } \leq t \text{ in i before making its next transition }\}$$

$$= \sum_{k=1}^{4} \int_0^t p_{ik}(s, s+x)\, h_{ik}(x)dx$$

Also, we can define:

$$\varphi_{ij}(s,t) = \text{prob}\{\text{a patient which entered state i at time s is in state j at time } s+t\}$$

$$= \delta_{ij} {}^{>}w_i(s,t) + \sum_{k=1}^{4} \int_0^t p_{ik}(s, s+x)\, h_{ik}(x) \varphi_{kj}(s+x, t-x)dx \quad (1)$$

and if we define $c_{ij}(s,x)=p_{ij}(s,s+x)h_{ij}(x)$

$$\varphi_{ij}(s,t) = \delta_{ij} {}^{>}w_i(s,t) + \sum_{k=1}^{4} \int_0^t c_{ik}(s,x)\varphi_{kj}(s+x, t-x)\, dx \quad (2)$$

where:

$$\delta_{\iota j} = \begin{cases} 1, & if\ i = j \\ 0, & otherwise \end{cases}$$

The above equation can be written in matrices form as follows:

$$\mathbf{\Phi}(s,t) = {}^{>}\mathbf{W}(s,t) + \int_0^t \mathbf{C}(s,x)\, \mathbf{\Phi}(s+x, t-x)dx \quad (3)$$

where $\mathbf{C}(s, x)=\mathbf{P}(s, s+x)\lozenge\mathbf{H}(x)$ is the Hadamard product of the two matrices. The initial condition for equation (3) is $\mathbf{\Phi}(s,0)=\mathbf{I}$.

The matrix $\mathbf{\Phi}(s,t)$ defines the interval transition probabilities for the patients of the non homogeneous semi Markov chain which is imbedded in our system. This semi Markov chain is fully described by the probability $p_{ij}(s,t)$ and the probability density functions of the holding times $h_{ij}(x)$ as it is shown by equation (1). So, we have a non homogeneous semi Markov system in which the individual transitions take place according to a non homogeneous semi Marko chain. Using probabilistic argument it can be proved that the closed analytic form of probabilities $\varphi_{ij}(s,t)$ is:

$$\mathbf{\Phi}(s,t) = {}^{>}\mathbf{W}(s,t) + \int_0^t \mathbf{C}(s,x_1)\{{}^{>}\mathbf{W}(s+x_1, t-x_1) + \mathbf{C}(s+x_1, t-x_1)\}dx_1$$
$$+ \sum_{k\geq 2} \int_0^t \int_0^{t-x_1} \int_0^{t-x_1-x_2} \dots \int_0^{t-x_1-\cdots-x_{k-1}} \mathbf{C}(s,x_1)\mathbf{C}(s+x_1,x_2) \dots \mathbf{C}(s+\cdots+$$

$$x_{k-1}, x_k)\{{}^{>}\mathbf{W}(s+\cdots+x_k, t-x_1-\cdots x_k) + \mathbf{C}(s+\cdots+x_k, t-x_1-\cdots-x_k)\}$$
$$dx_k dx_{k-1} \dots dx_2 dx_1 \quad (4)$$

We assume that our system is a closed one i.e. the total patients' population is constant at any time. The previous hypothesis is not in conflict with real systems because the patients' population under treatment, in that kind of chronic diseases, is usually constant. Thus, the states sizes of the system at any time is described by the vector $\mathbf{N}(t)=[N_1(t), N_2(t), N_3(t), N_4(t)]$ where $N_i(t)$ is the expected number of patients in the i-th state at time t. It is proved that :

$$N_i(t) = \sum_{j=1}^{4} N_j(0)\varphi_{ij}(0,t). \quad (5)$$

Equation (5) in matrix form is as follows:

$$\mathbf{N}(t)=\mathbf{N}(0)\mathbf{\Phi}(0,t) \quad (6)$$

where $\mathbf{N}(0)$ is the initial population structure.

Relations (5) and (6) provide the expected population structure as a function of the basic sequences of the system.

The corresponding discrete non homogeneous semi Markov model was defined in Vassiliou and Papadopoulou [27]. The transition probability matrix of the embedded Markov chain is defined by $\mathbf{P}(t)=\{p_{ij}(t)\}_{i,j\in S}$, where $p_{ij}(t)$=prob{a patient which entered in state i at time t to move in the state j at its next transition} and the holding time mass function matrix for the semi Markov chain is defined by $\mathbf{H}(m)=\{h_{ij}(m)\}_{i,j\in S}$ where $h_{ij}(m)$=prob{a patient which entered state i at its last transition to hold for m time in state i before making its next transition given that state j has been selected}.

Also, we define as:

$$w_i(t,m) = \text{prob\{a patient which entered state i at time t to stay m time units in state i before its next transition}\}.$$

It's proved that:

$$w_i(t,m) = \sum_{j=1}^{4} p_{ij}(t)h_{ij}(m) \quad \text{for } i = 1,2,3,4 \text{ and } t, m = 0, 1, 2, \ldots$$

and $w_i(0,t)=0$ for every i, t.

Moreover, let us define:

$$\varphi_{ij}(t,m) = \text{prob}\{\text{ a patient which entered state i at time t to be in state j after m steps}\}.$$

It is also proved that:

$$\varphi_{ij}(t,m) = \delta_{ij}\sum_{s=m+1}^{\infty} w_i(s,t) + \sum_{k=1}^{4}\sum_{s=1}^{m} p_{ik}(t)h_{ik}(s)\varphi_{kj}(t+s,m-s) \quad (7)$$

for i, j=1, 2, 3, 4 and t, m =0, 1, 2, … .

Also, we define by $^{>}\mathbf{W}(t,m)$ the 4×4 matrix which has zeros everywhere apart from the diagonal which has in position i the element:

$$\sum_{s=m+1}^{\infty} w_i(t,s) = 1 - \sum_{s=1}^{m} w_i(t,s).$$

Then the equation (7) can be written in matrices as follows:

$$\mathbf{\Phi}(t,s) = {}^{>}\mathbf{W}(t,s) + \sum_{s=1}^{m}[\mathbf{P}(t) \lozenge \mathbf{H}(s)]\mathbf{\Phi}(t+s,m-s). \quad (8)$$

Obviously $\mathbf{\Phi}(t,0)=\mathbf{I}$.

The corresponding states' sizes description in the discrete time case is given by the vector $\mathbf{N}(t)=[N_1(t), N_2(t), N_3(t), N_4(t)]$ where $N_i(t)$ is the expected number of patients in the i-th state at time t, where

$$N_i(t) = \sum_{j=1}^{4} N_j(0)\varphi_{ij}(0,t). \quad (9)$$

The respective matrix form is:

$$\mathbf{N}(t)=\mathbf{N}(0)\mathbf{\Phi}(0,t) \quad (10)$$

where **N**(0)is the initial population structure.

The choice, in practice, between the discrete and continuous time versions of a model is partly a matter of realism and partly one of convenience. On grounds of realism, for example one would usually want to model the movement of people between occupations or regions in continuous time, but in practice the computational advantages of treating time as discrete have often led to the choice of a discrete time model (D.J. Bartholomew [2], page 85).

There are two main reasons that we can observe a continuous process in two or more specified intervals . The first reason is the observed computational advance when we treat the time as a discrete variable. The second is that practical difficulties often occur in considering continuous time models, which arise from the fact that is rarely possible to observe continuous data (D.J. Bartholomew [2]).

In what follows, we will describe the technique applied for the discrete observation of the continuous time model presented at first. A discrete model can be developed by discretization per unit time of the transition probability and the holding time mass function matrix. The discretization per unit time of the transition probability matrix is based on the relationship :

$$\mathbf{P}(t) =) \int_{bt}^{b(t+1)} \mathbf{P_j}(u)du \quad (11)$$

where b is the defined unit and $\mathbf{P_j}(u)$ is the transition probability matrix of the corresponding jump process.

In the following the above relationship is applied to the data of Mathieu et al [16] where $\mathbf{P_j}(u)$ is of linear form and for the cases i) b=1 and ii) b=0.5. The corresponding graphs are presented in Figures 2.1 and 2.2.

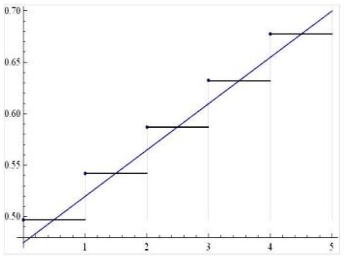


Figure 2.1: b=1

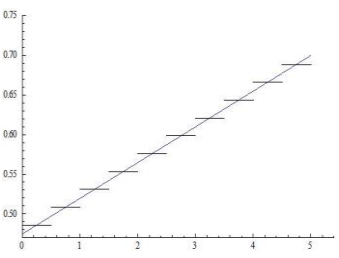


Figure 2.2: b=0.5

Concerning the holding time distributions, there are several ways to derive the corresponding discrete lifetime distribution from a continuous one. Two of the most usually applied are: a) consider a characteristic property of a continuous distribution and then build the similar property in discrete time and b) consider the discrete holding time as the integer part of the continuous holding time (Bracquemond and Gaudoin [3]).

By applying the second technique and if we consider the continuous time random variable T which describes the holding time in a state we can define the corresponding discrete random variable K and for time unit b=1 as:

$$K=[T]+1. \quad (12)$$

Also, the equivalent random variable L for time unit b=0.5 is :

$$L=[2T]+1 . \quad (13)$$

Let that F_K, R_K and F_T, R_T denote the cumulative distribution and reliability functions of the random variables K and T respectively. Then, the relationship between the probability function of K and the cummulative distribution function of T is:

$$\begin{aligned} p(k) &= Pr\{K = k\} = Pr\{k-1 \le T < k\} \\ &= F_T(k) - F_T(k-1), \ \forall k \in N \end{aligned} \quad (14)$$

Furthermore:

$$F_K(k) = Pr\{K \le k\} = Pr\{[T]+1 \le k\} = Pr\{T < k\} = F_T(k) \quad (15)$$

and

$$R_K(k) = R_T(k) \quad (16)$$

Finally and according to the previous, we discretize the holding times using the variables K and L and then, we can estimate the parameters of the holding time distributions to the states.

3 Illustration

In the present section, the discrete observation technique of a continuous semi Markov model is illustrated numerically with data from an HIV patients' population.

Firstly, a Monte Carlo simulation method is performed by using the data concerning the transition probabilities and the conditional distributions of the holding times as assessed by Mathieu et al [18]. The purpose of the simulation is to obtain the basic characteristics of the process in continuous time. The simulation is performed in ARENA 13.5 environment.

So, we consider a sampling path of a Semi-Markov chain over a period of time [0,C] where are observed M patients. Each patient starts the immunological and virological trajectory in any state, which is revealed by the first measurement at time t=0 and we assume that the k-th patient changes state n_k times in the instants $s_{k,1}<s_{k,2,}<\ldots<s_{k,n}<\ldots$ and occupies states $J_{k,1}<J_{k,2}<\ldots<J_{k,n}<\ldots$ and $J_{p,n} \neq J_{p,n+1}$ for every $n \in N$. In fact, such a path is a sequence H(C) of visited states and sojourn times :

$$H(C) = \{X_{0,} \ J_0 \ , X_1 \ , \ \ldots \ , J_{N(C)-1} \ , X_{N(C)} \ , J_{N(C)}\}$$

The above process is simulated by an algorithm of five steps as follows:

1. Set k=0, S_o=0 and sample J_0 from the initial distribution
2. Sample the random variable J~ $p(J_k,.)$ and set J_{k+1}=J
3. Sample the random variable X~ $H_{J_kJ_{k+1}}(.)$
4. Set $S_{k+1}= S_k + X$
5. If $S_{k+1}>M$ then end else set k=k+1 and continue to step 2

Thus, we can get the trajectory of a patient and if we do the same repeatedly for 1000 times we can get the trajectories of 1000 patients. The results of the simulation are presented in Table 1.

Table 1: Results of Simulation

Transitions	Number of transitions	Mean sojourn time (years)
1→2	934	1,59
1→3	195	1,77
1→4	473	1,69
2→1	807	0,15
2→3	2842	0,87
3→2	2759	1,08
3→4	506	0,8
4→1	358	1,16
4→2	164	1,58
4→3	469	1,74
Total	9507	

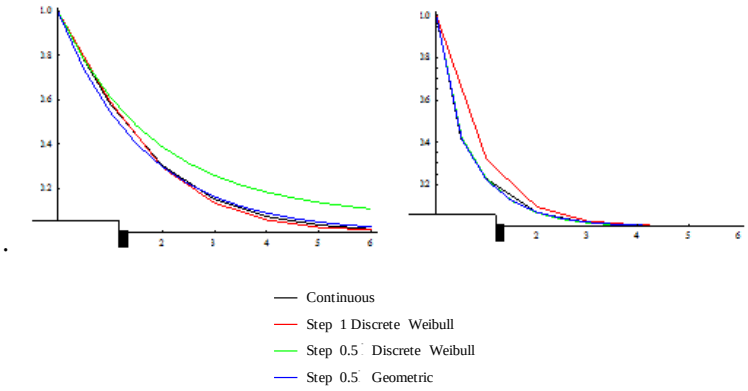


Figure 3.1 Survival functions for states 1 and 2

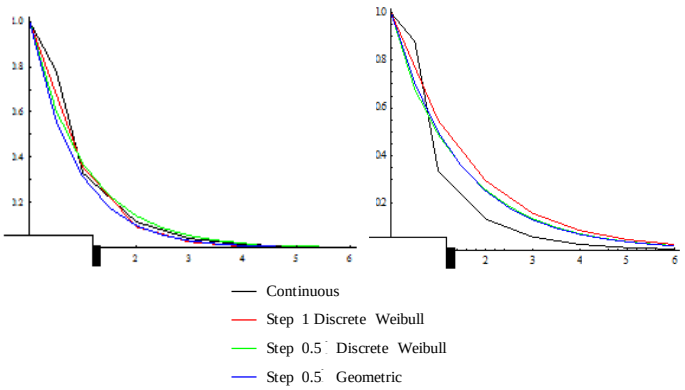


Figure 3.2 Survival functions for states 3 and 4

Using the results derived from the simulation we developed three discrete time models for HIV and we compared them with the continuous model. The transition probabilities and the holding times are derived by applying equations (11)-(14). We then estimated the parameters of the holding time distributions for the health states. For the holding times we used the type I discrete Weibull distribution (Nakagawa and Osaki [22]) and the geometric distribution. In the first and second model, the method of proportions (Ali Khan et al [1]) was used in order to estimate the parameters of the type I discrete Weibull distribution. In the third model we used the method of maximum likelihood to estimate the parameters of the geometric distribution. Finally, we evaluated the numerical results for the survival functions of every state and the population structures derived from the three models in comparison to the continuous one. The results are presented in Figures 3.1, 3.2, 3.3 and 3.4.

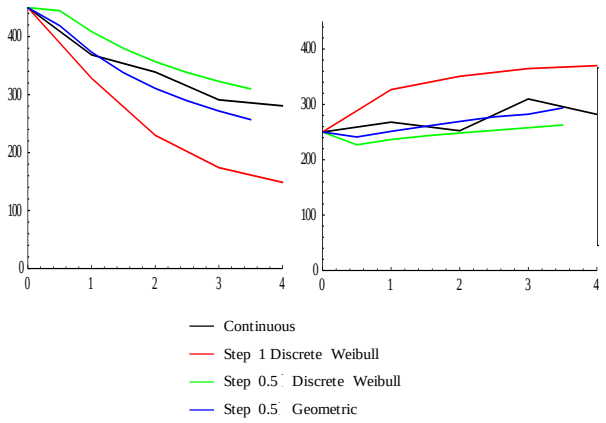


Figure 3.3 Population of states 1 and 2

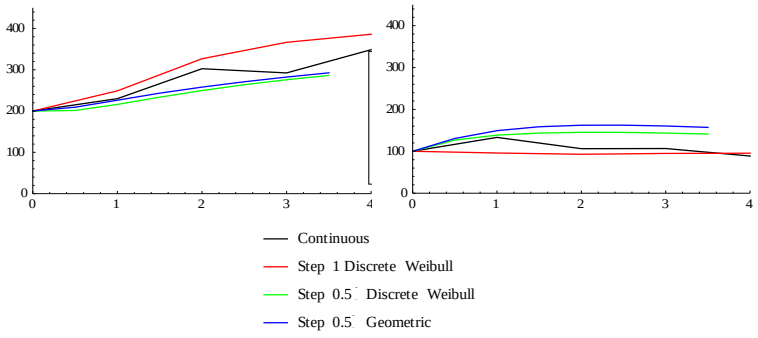


Figure 3.4 Population of states 3 and 4

Conclusions

In the present paper we reviewed a continuous time semi Markov model for HIV control and we provided a technique for the discrete observation of the continuous model. This technique was applied to the reviewed model. For this purpose, we used a simulation process to illustrate the discrete observation technique with synthesized data. Finally, we developed three discrete time models for HIV and we compared them with the continuous model. The derived results demonstrate the potential of the applied technique to provide a tool for discrete observation of continuous models.

References

1. Ali Khan, M., Khalique, A., & Abouammoh, A. (1989). On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability, Vol. 38, No. 3*, 348 -350.
2. Bartholomew, D.J. (1982). Stochastic Models for Social Process, 3rd edition, Wiley. Chichester.
3. Bracquemond, C., & Gaudoin, O. (2003). A survey on discrete lifetime distributions. *International Journal of Reliability, Quality and Safety Engineering ,10(1)*, 69-98.
4. Cinlar, E. (1969). Markov renewal theory. Adv. Appl. Prob. 1:123-187.
5. Cinlar, E. (1975a). Markov renewal theory: A survey. Manage. Sci, 21:727-752
6. Cinlar, E. (1975b). Introduction to Stochastic Processes. Englewood Cliffs, NJ: Prentice Hall.
7. De Dominics, R., Manca, R. (1985). Some new results on the transient behavior of semi Markov reward processes. Methods Oper. Res. Comput. 13:823-838.
8. Foucher, Y., Mathieu, E., Sain-Pierre, P., Durand, J., & Daures, J. (2005). A Semi-Markov Model Based on Generalized Weibull Distribution with an illustration for HIV Disease. Biometrical Journal,vol.47 , pages 825-833.
9. Howard, R. (1972). Dynamic Probabilistic Systems, vol. II. D Wiley.
10. Iosifescu-Manu, A. (1972). Non homogeneous semi Markov processes Studiisi Cercetuari Mathematice 24:529-533.
11. Janssen, J. (1986), Semi Markov Models: Theory and Applications. Janssen, J., ed. New York: Plenum Press.
12. Janssen, J., De Dominics, R. (1984). Finite non homogeneous semi Markov processes: Theoretical and computational aspects. Insurance: Math. Econ. 3:157- 165.
13. Keilson, J. (1969). On the matrix renewal function for Markov renewal processes. Ann. Math. Statist., 40:1901-107.
14. Keilson, J. (1971). A process with chain dependent growth rate. Markov Part II: The ruin and ergodic problems. Adv. Appl. Prob. 3:315-338.
15. Limnios, N., & Oprisan, G. (1999). *'Semi Markov Processes and Reliability.* Boston: Birkhauser.
16. Mathieu, E., Loup, P., Dellamonica, P., & Duares, J. (2006). 'Markov modeling of immunogical and virological states in HIV-1 infected patients. *Biometrical Journal,vol.48* , pages. 834-846.
17. McCLean, S.I. (1976). The two stage model for personnel behavior. J.R.S.S. A1399:205-217
18. McCLean, S.I. (1978). Continuous time stochastic models for a multigrade population. J. Appl. Prob. 15: 26-32
19. McCLean, S.I. (1980). A semi-Markovian model for a multigrade population . J.

Appl. Prob. 17:846-852
20. McCLean, S.I. (1986). Semi-Markov models for manpower planning. In: Semi-Markov Models: Theory and applications. New York: Plenum Press, 283-300.
21. McLean, R.A., Neuts, M.F. (1967). The integral of a step function defined on a semi Markov process. Siam. J. Appl. Math. 15:726-737.
22. Nakagawa, T., & Osaki, S. (1975). The discrete Weibull distribution. *IEEE* Transactions on Reliability, Vol. 24, No.5 , pages 300-301.
23. Papadopoulou, A., & Vassiliou, P.-C. (1999). Continuous time non homogenous Semi Markov systems. Semi Markov Models and Applications, J.Janssen & N.Limnios (Eds), Kluwer Academic Publishers, Dordrecht , 15 , pages 241-251.
24. Papadopoulou, A. A., Vassiliou, P.-C. G. (1994). Asymptotic behavior of non homogeneous semi-Markov systems. Linear Algebra Appl. 210:153–198.
25. Tan, W.Y. (2000) Stochastic modeling of AIDS Epidemiology and HIV Pathogenesis. World Scientific: Singapore
26. Teugels, J.L. (1976). A bibliography on semi-Markov processes. J. Comp. Appl. Math. 2:125-144
27. Vassiliou. P.-C., Papadopoulou A. (1992). Non homogeneous semi Markov systems and maintainability of the state sizes. Journal of Applied Probability, 29, pages 519-534.

Multiobjective Optimization Approach to Solve a Maintenance Process Problem

Nouha Lahiani[1, 2,3], Yasmina Hani[1,2], Abderrahman El Mhamedi[1, 2] and Abdelfateh Triki[3]

[1] Laboratoire d'ingénierie des systèmes mécaniques et matériaux LISMMA. (E-mail: n.lahiani; y.hani; a.elmhamedi@iut.univ-paris8.fr)
[2] University Paris VIII, University Institute of Technology of Montreuil, 140 Rue de la nouvelle France, 93100 Montreuil, France.
[3] University of Tunis, Laboratory ARBRE–Institut Supérieur de Gestion. (E-mail: abdel.triki@gmail.com)

Abstract. In this paper, a multiobjective optimization study of maintenance process is developed. The considered problem is inspired by a real case. Given the complexity of the industrial case, we combined a discrete event simulation model with an optimization engine based on non-dominated sorting genetic algorithm II (NSGA-II). The coupling is used in order to optimize the performances of the simulation model by choosing the best queues' scheduling policy. The issue is to reorganize the maintenance process under operator's qualifications and interventions emergency degree. The NSGA-II engine and simulation model operate in parallel over time with interactions. After computation, we obtain high quality solutions in very short commuting time.

Keywords: Multiobjective Optimization, Maintenance Process, Case Study.

1 Introduction

In the current business environment, the manufacturing companies should be more reactive, flexible and competitive. These objectives can be realized by improving production and maintenance systems. Therefore, many manufacturing companies are changing their systems. Given the complexity of the industrial organizations, diagnosis systems and decision-making tools are becoming an important requirement in our days. This is especially true in the maintenance management process, characterized by expensive specialized equipment and stringent environmental considerations. In fact, maintenance represents a significant function within the overall production environment. Thus, a good overview of maintenance processes and achievements is needed to ensure a good performance of the production plant.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

In this study, we consider a maintenance planning problem inspired by a real case. The maintenance human resources assignment is specially considered in order to check the interventions request (IR). As a matter of fact, the bad management of the IRs can lead to greater costs.
To evaluate the industrial systems' performance, simulation is usually used. Huang[1] uses Flexsim simulation tool to establish and solve re-scheduling problems under a flow-shop mixed-line production planning. Da-wei and Jian-guo[2] use Flexsim for simulating the planning and construction of highway bus terminal program dynamically. For Sharma *et al.*[3], the use of the simulation technique has been an emerging trend which changed the maintenance view, given that it allows experimentation and a better understanding of complex systems like the maintenance one. Alrabghi and Tiwari[4] and Sharma *et al.*[3] provide a comprehensive view of maintenance optimization using discrete event simulation models and show the potential of this technique to solve complex system problems such as maintenance systems. On the other hand, Dekker[5] proposes a comprehensive view and analyses of maintenance optimization models. In his research, the author was interested in mathematical models only.
Comparing twenty-eight scientific published papers, Alrabghi and Tiwari[4] notice that research on planning maintenance simulation is steadily rising. They also notice that research on the combined use of simulation and optimization is limited.
In this paper, an original approach based on a multiobjective optimization of maintenance process problem inspired by a real case is proposed. For this purpose, a combination of a simulation model and a non-dominated sorting genetic algorithm II (NSGA-II) optimization engine is developed. In this study the qualifications of operators and their availability are considered.
The remainder of this paper is organized as follows: the next section describes the case study and the considered constraints. The third section presents the optimization study. Finally, in section 4 we analyze the optimization results and conclude the paper along with some suggestions for future work.

2 Methods

2.1. Case Study

This study is inspired by a real case. It concerns a company that produces tractor transmission parts, essentially gearboxes and rear axles as described in Lahiani *et al.* [6], [7]. For the sake of confidentiality, we named this company as "MNL-company".
The company work time is divided into 3 periods over the 24hours of the day. Machines run for a long period. Therefore, machines breakdown is a common problem for those running continually without preventive maintenance. It was observed that the company does not perform the maintenance planning effectively, which impacts work effectiveness, equipment reliability, equipment uptime, costs, etc. The MNL-company finds itself facing greater costs due, in

the most part, to the maintenance problems. This is the consequence of the bad management of intervention requests (IR). In 2011, the repairing time varied between 1 minute and more than 30 days.
The IRs were treated in first-in, first out (FIFO) policy, which meant the emergency of the intervention wasn't considered. Thus, the assignment of the human resources was due subject to only its availability.
Production and maintenance processes have to interact in time which makes the system even more complex. The issue lies in the reorganization of the maintenance system although it is difficult for the company to manage.
Given the problem complexity and in order to optimize the maintenance process, we adopt a simulation modeling to analyze and evaluate the industrial performance.

2.2. Simulation Model

In this paper, a discrete event simulation model has been built by using the software package Flexsim version 6. The simulation model's aim is to re-organize the maintenance system, especially to manage the request of intervention on machines. Many aliases can be integrated in our model. However, in this paper we focus on the operator's competence indicators and the intervention emergency degree. For more visibility of the system, we have classified the failures according to their type and human resources' competences (table 1).

Table 1. Intervention Requests' Type

IR group	**Problem** type
$TyIR_1$	Palette default
$TyIR_2$	Filters problems
$TyIR_3$	Tank default
$TyIR_4$	Watering problem
$TyIR_5$	Axis machine problems
$TyIR_6$	Manipulation problems

In case of problems, the production service sends an IR. In our model, when an IR arrives, the emergency degree is firstly verified. It is classified, in accordance with the person in charge of the maintenance, in three classes: very urgent, moderately urgent, and not urgent. As a result, considering the human resources' qualifications and the IR emergency degree, the assignment protocol of the HR is as described in figure 1.

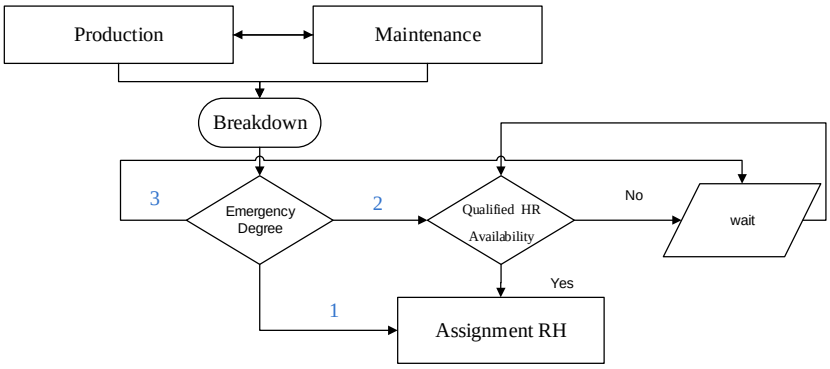


Figure 1. Assignment Protocol of Human Maintenance Resources

The discrete event simulation model objects consist of a shuttle, a source, a dispatcher to manage the operators, a sink, a queue and four identical machines. The simulation model runs at a horizon of one working year.
The source object sends IR into machines according to a known distribution function.

The model inputs are:

- The structure (buildings, locations, number of machines and characteristics...);
- Human resource number;
- Human resources qualifications,
- Operating time for every breakdown;
- The simulation horizon (one working year);
- The statistic rules representing time between two arrivals (ExpertFit^{TM} software was used to identify statistical rules)

The model outputs are:

- The stopping time of machines;
- Number of failures repaired;
- Number of intervention over one year
- The duration of immobilization of maintenance HR.

By dint of simulation model, we observe that the checked IR number is not important compared to the long machines stopping times. For this purpose, we decide to optimize the maintenance process in the last section.

3 Multiobjective Optimization

This study has three contractors' objectives: maximize the checked IR number, minimize stopping time of machines and minimize repairing time of machines.

For example to maximize the checked IR, there is a need to maximize stopping time. For this reason, Pareto approach is selected to optimize the maintenance process.
NSGA II is proposed among Pareto algorithms because it is an efficient recent method that was applied in different areas of research.

3.1.Optimization Algorithm

In this paper, we adopt a known metaheuristic: non dominated sorting genetic algorithm NSGA-II developed by Deb *et al.*[8]. It's a popular multiobjective evolutionary algorithm. Traditionally, it is chosen because of its three special characteristics as advocated by Yusoff *et al.* [9] and Raddaoui *et al.* [10]:
-Fast non-dominated sorting approach,
-Fast crowded distance estimation approach,
-Simple crowded comparison operator
The chosen algorithm converges to a global Pareto optimal front. For Deb *et al.*[8], the algorithm can maintain the diversity of population on the Pareto-optimal front. The major characteristic of NSGA-II lies in the concept of non-domination between two solutions.
In this algorithm, the population is firstly initialized as usual. After initializing, the population is sorted based on non-domination criteria into each front. The first front being completely a non-dominant set in the current population and the second front being dominated by the individuals in the first front only and the front goes on. Each individual in each front is assigned rank (fitness) values according to the front to which they belong.
In addition to fitness value, a new parameter (crowding distance) is calculated for each individual. The crowding distance is a measure of how close an individual is to its neighbors.
In the population, individuals are selected according to rank and crowding distance. Then, parents are selected from the population by using binary tournament selection based on the rank and crowding distance.
An individual is selected in the rank if it is less than the other or if crowding distance is greater than the other one.
The selected population generates off springs from crossover and mutation operators, which will be explained in detail in the next section. The population with the current population and current off springs is calculated again based on non-domination and only the right N individuals are chosen where N is the population size. The selection is based on rank and crowding distance on the last front.
Several researches use NSGA-II algorithm to optimize systems performances. Yusoff *et al.*[9] presents an overview on NSGA-II optimization method. His survey considers the machining process problems. The contribution of Raddaoui and Zidi[11] is limited to the use of the algorithm NSGAII for solving a dial ride problem.

3.2. Application

The NSGA-II algorithm is chosen in order to optimize the systems' performances (characterized by three objectives) by choosing the best queues' scheduling policy.
IR can be the result of several causes. For better visibility of system problems we have classify the IR on groups according to the failures types listed in Table 1. By means of this classification, the assignment of the HR is more controlled. Such, the RH' assignment is constraint to their availability and qualifications.

The algorithm steps are as following:

Step 0: *Initialization*: Generate an initial population.
Step 1: *Evaluation*: Evaluate the fitness using the DESM.
Step 2: *crowding distance*: for each couple of parents selected randomly, apply a crossover in two points from the selected parents and generate two children (*Qt).*
Step 3: use the non-dominated sorting procedure for separate fronts.
Step 4: Assign the best front solutions to construct the matrix P_{t+1} and use the crowding distance for the last selection front.
Step 5: the matrix P_{t+1} is considered as an initial population.
Step 6: return to step 1.
Step 7: Reiterate until the stop criterion.

In our problem, five sequencing rules are tested: FIFO (first in first out, LIFO (Last in first out), SPT (Short processing time), HPT (hight processing time), Task priority. The best priority rule must be chosen for each type of $TyIR_i$ in order to improving our system. For each policy, the three objectives are obtained. The machines' stopping time include waiting operators times, repairing times, and other alias.
The proposed approach is a combination of two processes. First, a set of configurations, named chromosomes, is selected to form the original population for NSGA-II. The algorithm provides the initial solutions which tested on simulation model. Thus, the population performances of each initial solution are identified. This process is repeated until the stopping criterion is satisfied. One of the methods that can be used as a stopping criterion is to fix a number of iterations arbitrations selected. In this paper, 1000 replications was tested.

4 Experimental Results

According to the parameters settings, we fixed the following algorithm entries:
- Initial population: 60
- Probability of mutation (P_m): 0.001%
- Probability of crowding (P_c): 100%
- Iterations number: 1000.

NSGA-II gets a wide variety of equivalent solutions. Solutions 111145, 454231, 151145 are examples of perform solutions (corresponding to the NSGA-II individuals, chromosomes). For the first chromosome, this code indicates that the queue priority rules is: FIFO, FIFO, FIFO, FIFO, HPT, priority.
The table 2 presents the average value of the optimization study and the best results for each objective.

Table 2. Experimental Results

	Real Value	Optimization (Average)		Optimization (Best)	
		Results (Average 10 itérations)	GAP %	Results	GAP %
Checked IR Number	1289	1139,00	-11.64	1376,00	6,75
Total stopping time	6900.00	6175,93	-10.49	5608,13	-18,72
Total reparing time	5627.10	5869,95	4.32	5100,51	-9,36

In this table, the total stopping time means the sum of stopping time of all machines. Repairing and stopping times are presented in hour.
By means of a NSGA-II all objectives are improved. The average value of total repairing time (5869,95h) is not improved comparing of the simulation average value (5630.09h). However, with the best value (5100,51h) there is a tremendous performance for 18.72% of times.
For these results, total computer time was 22.37 minutes, on a computer core i5, windows 7. It is relatively short.

5 Conclusions

In this paper, a simulation model was developed to analyze the performance of a maintenance process in an industrial case. Our contribution resides in the resolution of an existing industrial problem and optimizing the systems' maintenance process. A non dominated sorting genetic algorithm was used to optimize the system by testing different scheduling policies. In the proposed approach, the operators' competences, their availability and the emergency degrees of interventions are considered. This approach can be considered as a decision-making tool optimizing the number of the breakdown repaired, the stopping time of machines and the repairing time.
Applying this technique on an industrial case study, we show that it is more effective in detecting real faults than existing alternatives. The results showed remarkable improvements of the maintenance system performances. This

proposition can be extended to cover other domains and other types of simulation models.
Further studies may test other priority rules or other optimization algorithms.

References

[1] H.-H. Huang, W. Pei, H.-H. Wu, and M.-D. May, "A research on problems of mixed-line production and the re-scheduling," *Robot. Comput. Integr. Manuf.*, vol. 29, no. 3, pp. 64–72, Jun. 2013.

[2] H. Da-wei and X. Jian-guo, "Simulation system for bus station based on Flexsim," *J. Chang. an Univ. Nat. Sci. Ed.*, vol. 30, no. 2, pp. 89–95, 2010.

[3] A. Sharma, G. S. Yadava, and S. G. Deshmukh, "A literature review and future perspectives on maintenance optimization," *J. Qual. Maint. Eng.*, vol. 17, no. 1, pp. 5–25, 2011.

[4] A. Alrabghi and A. Tiwari, "A Review of Simulation-Based Optimisation in Maintenance Operations," in *UKSim15th International Conference on Computer Modelling and Simulation*, 2013, pp. 353–358.

[5] R. Dekker, "Applications of maintenance optimization models: a review and analysis," *Reliab. Eng. Syst. Saf.*, vol. 51, no. 3, pp. 229–240, Mar. 1996.

[6] N. Lahiani, Y. Hani, and A. El-mhamedi, "Modélisation et simulation des flux de maintenance dans un système de production," in *12éme Congrès annuel de la société Française de recherche opérationnelle et d'aide à la décision.*, 2012, pp. 1–2.

[7] N. Lahiani, Y. Hani, A. Triki, and A. El-Mhamedi, "Application de l ' Algorithme NSGA-II pour l ' Optimisation Multiobjectif dans un Atelier de Production de Biens .," in *15e congres de la Société Française de la Recherche Opérationnelle et d'Aide a la Décision ROADEF*, 2014.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm :NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[9] Y. Yusoff, M. S. Ngadiman, and A. M. Zain, "Overview of NSGA-II for Optimizing Machining Process Parameters," *Procedia Eng.*, vol. 15, pp. 3978–3983, Jan. 2011.

[10] A. Raddaoui, I. Zidi, K. Zidi, and K. Ghedira, "Distributed Approch Using NSGAII Algorithm to Solve the Dynamic Dial a Ride Problem," in *the World Congress on Engineering and Computer Science (Vol. 1)*, 2013, vol. I, pp. 23–25.

[11] A. Raddaoui and K. Zidi, "Distributed Genetic Algorithm NSGA-II for solving the DRP," in *4th International Computing, Conference on Metaheuristics and Nature Inspired*, 2012.

Life Expectancy and Modal Age at Death in Selected European Countries in the Years 1950-2012

Jana Langhamrová[1], Kornélia Cséfalvaiová[2] and Jitka Langhamrová[3]

[1]University of Economics, Prague, Czech Republic
Email: jana.langhamrova@vse.cz
[2]University of Economics, Prague, Czech Republic
Email: xcsek00@vse.cz
[3]University of Economics, Prague, Czech Republic
Email: langhamj@vse.cz

Abstract: At present, the majority of developed countries deal with the phenomenon of population ageing. This ongoing process is primarily caused by the increasing life expectancy at birth. Length of life is usually expressed by the indicator of life expectancy at age *x*. The values of life expectancy and modal age at death are different from the view of time evolution. This is particularly because life expectancy is the average age of deceased persons in the stationary population, whereas modal age at death is the most common age at death. The aim of this article is to analyse the trends in the development of the death rates in selected European countries using various methods for the death rates compensation. By means of data from the Human mortality database life expectancies and modal ages at death in selected countries will be calculated and compared. The purpose of this paper is to highlight the changes in the trend and dynamics of the life expectancy at birth and to compare its progress with the trend and dynamics of the modal age at death. By comparing the evolution of life expectancy at birth, life expectancy at age 65 and modal age at death, it is visible that modal age at death is not increasing as rapidly as life expectancy at age *x*. It is necessary to compensate the values of modal age at death or to use a more accurate calculation applying the Gompertz-Makeham function. Furthermore, there is a noticeable difference in the development of Western and Eastern Europe.

Keywords: Population ageing, Life expectancy at birth, Modal age at death

1 Introduction

21th century is from the demographic point of view mainly associated with the issue of population aging. This process is accompanied by an increasing rate of elderly persons, especially in economically developed countries. Mortality rates are improving and people live longer. The development of mortality in developed countries was neither linear nor logistic. Periods of faster or slower

3rd SMTDA Conference and Demographics Workshop Proceedings
11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

decline varied in time and some trends in mortality were unexpected. Currently, decrease in mortality in older age groups is mentioned as a major factor in the aging population. Trend of continuous aging of the European population will continue in the next period. Demographic aging is defined as a shift of the age structure to older ages (Gavrilov - Heuveline, 2003).

The question of longevity - the individual's ability to survive and the average length of human life have always been the object of interest of human populations. Longevity is often incorrectly defined as the presence of a larger group of old aged people at some territory in a certain population. This term does not refer to the upper limit of human life, which we can reach. It reports that natural life expectancy is increasing for which we consider modal age at death (Pavlík, 2009). In studies of longevity the most often used indicators are life expectancy and modal age at death.

Life expectancy is an indicator like average. It represents the average age of deaths in a stationary population. Modal age at death is an indicator like mode. It is the age at which adults most frequently die and from this perspective it better captures the length of human life than the life expectancy (Demopædia). Modal age at death is kind of a typical age, which most people in the population are expected to live.

From a long-term view, life expectancy and modal age at death are extending in most of the selected countries. This ongoing growth has a different intensity for each country. In the Czech Republic, and even in the former socialist countries, growth in life expectancy was noticed only in the 1990s of the 20th century. In some of the Western European countries, for example in Austria, life expectancy has been growing steadily since the 1970s of the 20th century. This has created significant differences in life expectancy and modal age at death between countries. These differences are slowly decreasing as the differences in mortality between women and men are descending. However, male excess mortality remains in all selected countries.

Since the beginning of the 20th century we have been watching a significant improvement in the development of life expectancy and modal age at death. The increase of life expectancy is generally seen as a positive process. Primary impulse of the increase of life expectancy was caused by decline of infant mortality and consequently by decrease of mortality in older age groups. Due to the improvement of mortality ratios, modal age at death is increasing and life expectancy is gradually approaching. While examining trends in mortality, life expectancy and modal age at death should be followed synchronously. As long as mortality rates of different age groups will tend to improve, life expectancy and modal age at death will increase as well. The speed of development will depend on the ages that contribute to the improved mortality rates (Wilmoth, 2000).

Unlike life expectancy at birth, modal age at death is substantially affected by mortality of adults and therefore reacts more sensitive to changes that occur among older aged population (Horiuchi 2008; Kannisto 2001). In countries with low mortality rates, where most of the deaths are recorded in old age, the indicator modal age at death becomes primary for monitoring the changes in the

age-at-death distribution (Ouellette - Bourbeau, 2011). After improved mortality in the first years of life, the current answer for extending life expectancy and modal age at death is associated with reducing mortality rates in old ages.
Life expectancy is extending due to low child mortality and modal age at death is increasing due to the decline in mortality rates at high ages (Canudas-Romo, 2010). The differences in the trends over time of these indicators of longevity well reflect their orientation to various aspects of mortality (Cheung – Robine, 2009). Life expectancy is currently the most commonly used mortality indicator despite the fact that it includes disadvantages of mean. Modal age at death has no disadvantages of average, it is the modal age at death among adults. For the purpose of this article life tables were calculated for selected countries. Life expectancy was calculated by (Fiala, 2005) using the Gompertz-Makeham function. In this article we present only the method for calculating the modal age at death.

2 Calculating Modal Age at Death

Modal age at death can be estimated as a rough estimate (age at last birthday with the maximum number of deaths). We are looking for an age when the number of deaths in mortality tables is the highest.
However, as already mentioned, this is only a rough estimate and more accurate results are obtained when using parameters of Gompertz-Makeham function. Calculation of modal age at death was performed by Fiala (2005). For this calculation it is necessary to know some source data as total deaths is specified age group ($M_{t,x}$) and mid-year population in the same age group ($\overline{S_{t,x}}$) or total population at the beginning of one year ($S_{t,x}$) or total population at the end of one year ($S_{t+1,x}$). Primary characteristics of mortality are age-specific death rates. Age-specific death rate for one calendar year is calculated using the formula

$$m_{t,x} = \frac{M_{t,x}}{\overline{S}_{t,x}}.$$

When we don´t know the number of mid-year population, but we have the number of total population at the beginning of the year *t* and *t+1*, we use the following formula

$$m_{t,x} = \frac{M_{t,x}}{\frac{S_{t,x} + S_{t+1,x}}{2}},$$

Where $M_{t,x}$ is the number of total deaths in completed years *x* and in calendar year *t*, $S_{t,x}$ is the total population aged *x* at the beginning of the year *t*. Compensation of age-specific death rates at age 60 and above can be calculated by Gompertz-Makeham equation

$$\widetilde{m}_x^{(GM)} = a + b \cdot c^{x+\frac{1}{2}}.$$

We select the beginning of the first $x_0 = 60$ and the length of intervals $k = 8$. We calculate the summation of empirical death rates by age in each interval and mark them as *G1, G2, G3*

$$G_1 = \sum_{x=60}^{67} m_x,$$
$$G_2 = \sum_{x=68}^{75} m_x,$$
$$G_3 = \sum_{x=76}^{83} m_x.$$

Now we can calculate the value of the parameter c of Gompertz-Makeham function whose eighth squared value can be expressed by using the sum of empirical death rates by age in each interval

$$c^8 = \frac{G_3 - G_2}{G_2 - G_1}.$$

Furthermore, it is necessary to calculate the value of the subexpression, by which we can express the remaining two parameters of the function

$$K_c = c^{60,5} \cdot (1 + c + ... + c^7) = c^{60,5} \cdot \frac{c^8 - 1}{c - 1}.$$

We can calculate the parameters b by using next expressions

$$b = \frac{G_2 - G_1}{K_c \cdot (c^8 - 1)},$$
$$a = \frac{G_1 - b \cdot K_c}{8}.$$

And according to these parameters of the Gompertz-Makeham function we can now calculate the modal age at death more precisely

$$\hat{y} = \frac{\ln \frac{\ln c - 2a + \sqrt{(\ln c - 4a) \cdot \ln c}}{2b}}{\ln c}.$$

3 Modal Age at Death and Life Expectancy in Selected European Countries

For reasons of comparison, countries representing the former socialist countries were selected (Bulgaria, Czech Republic, Slovakia, Ukraine) and the advanced Western countries (Austria, France, the Netherlands, Sweden). Choice of countries was also carried out according to the availability of required data. Not all countries are listed here because of the length of the article. Calculations are based on the data from "Human Mortality Database" for available years. Differences in modal age at death between women and men were smoothed by using the three-year moving average method to better reflect the trend over time. In this article this method was used only in case of differences in modal age at death between woman and men.

There are visible differences among the former socialist countries in the development of modal age at death, because modal age at death is developing more slowly compared to Western countries. At the beginning of the studied period differences in modal age at death between men and women were lower – 5 years for men and 3 years for women. Lowest values of modal age at death showed men in 1950 in the Czech Republic and Austria, women in the Czech Republic and Slovakia, highest values of modal age at death were for men and women in the Netherlands and Sweden. The highest values of modal age at death at the end of the studied period achieved France for both men and women. By contrast, the lowest values for both men and women were achieved by Ukraine (see figure 1 and figure 2). During the followed period differences between countries increased significantly, from a long-term view there is a divergence in the development of modal age at death. The difference between the countries at the end of the studied period was markedly higher – 14,5 years for men and 8,5 years for women.

From the perspective of the male excess mortality (see figure 1 and figure 2) it is visible how modal age at death for men and women differs in selected countries in different years. Again, differences between the former socialist countries and Western countries are noticeable. The smallest variance in modal age at death between men and women in 2010 was in Sweden and France and the highest variance was in Ukraine (see figure 3). According to figures 1 and 2 we can see that there was no growth in the decades at the beginning of the studied period, but rather a slight decrease of modal age at death. Only since 1970s there is an extending modal age at death in majority of the countries.

There was a considerable variability in the distribution of deaths by age in the previous period. At the present time, difference in the distribution of deaths by age has stabilized. What is more, the current model of mortality decrease where most of deaths are concentrated in a tight age range and where variability is low, could be kept for the future development (Wilmoth – Horiuchi, 1999).

Although mortality has been concentrated into shorter age intervals, it is impossible to tell for sure that it ever would be modified into one point in age. In human longevity, heterogeneity is a factor of individual variation (Vaupel 1979; Wilmoth – Horiuchi, 1998).

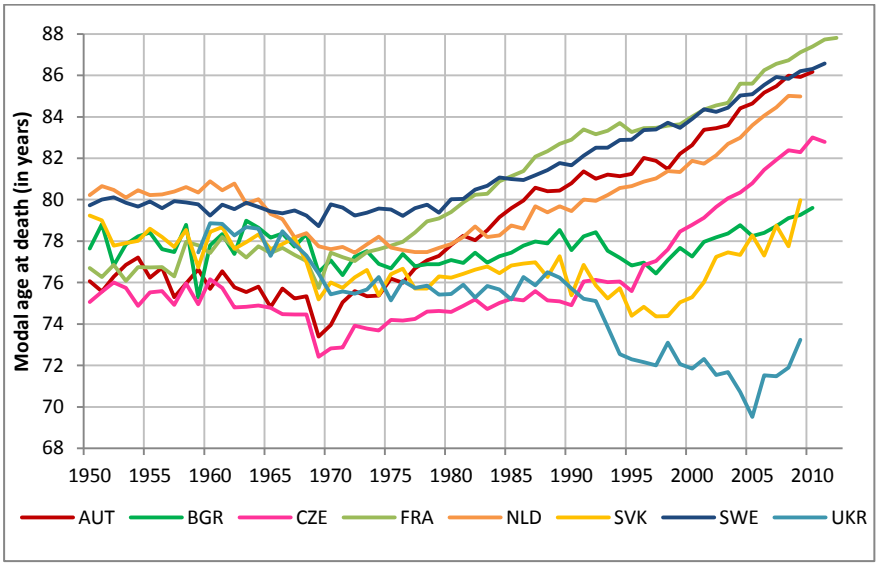


Fig. 1 Modal age at death for men in selected European countries in 1950-2012

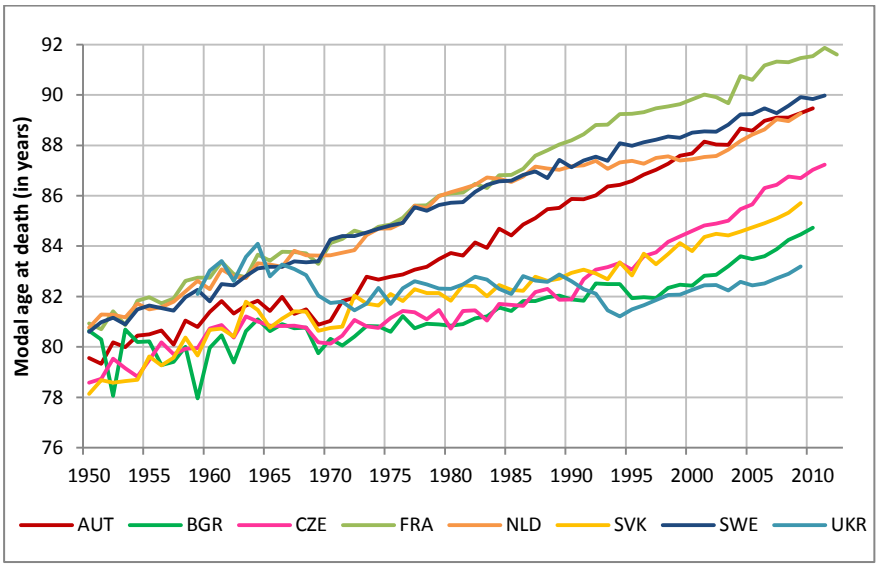


Fig. 2 Modal age at death for women in selected European countries in 1950-2012

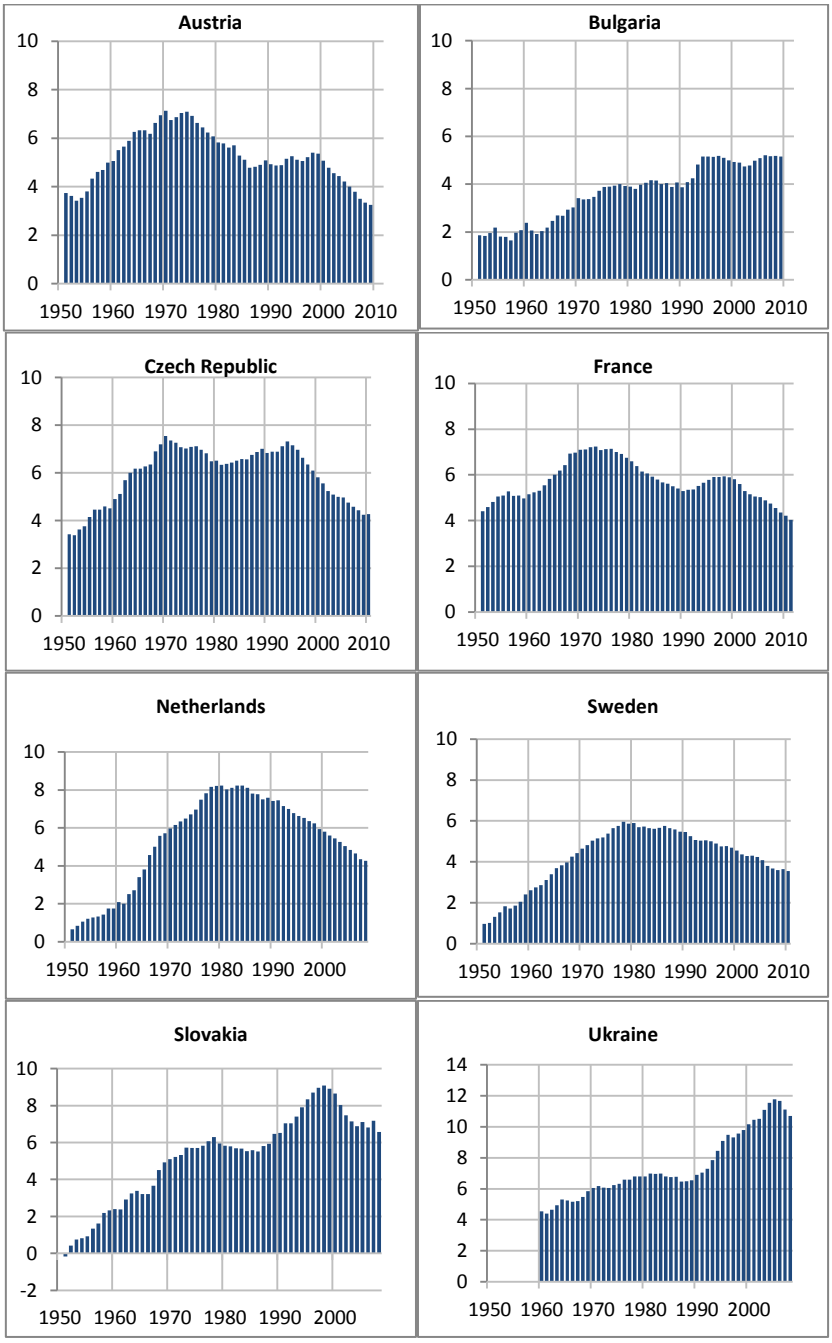


Fig. 3 Difference in modal age at death between women and men in selected European countries in 1950-2012

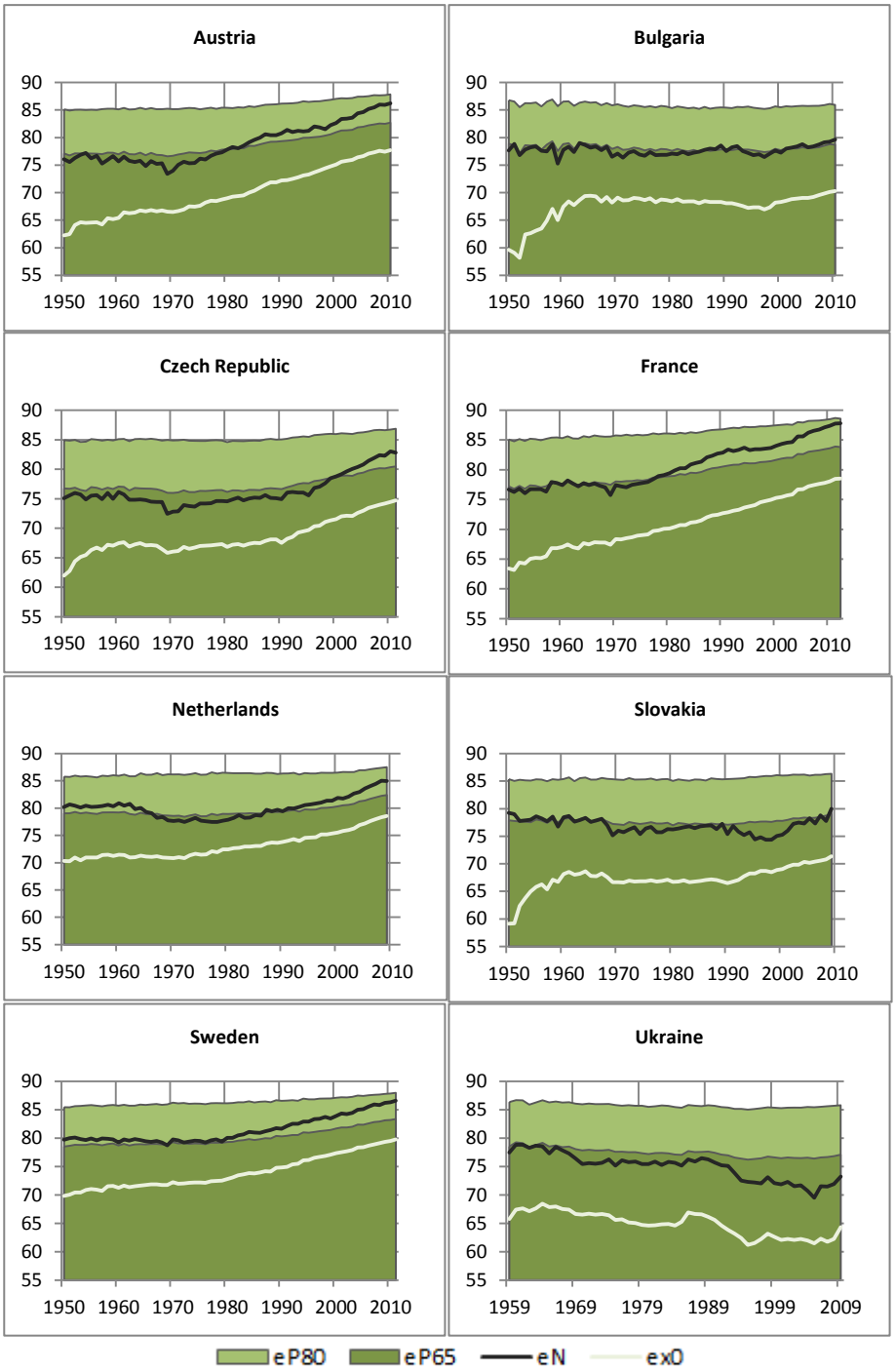


Fig. 4 Modal age at death, life expectancy at birth, probable age at death of 65-year-old and 80-year-old men in selected European countries in 1950-2012

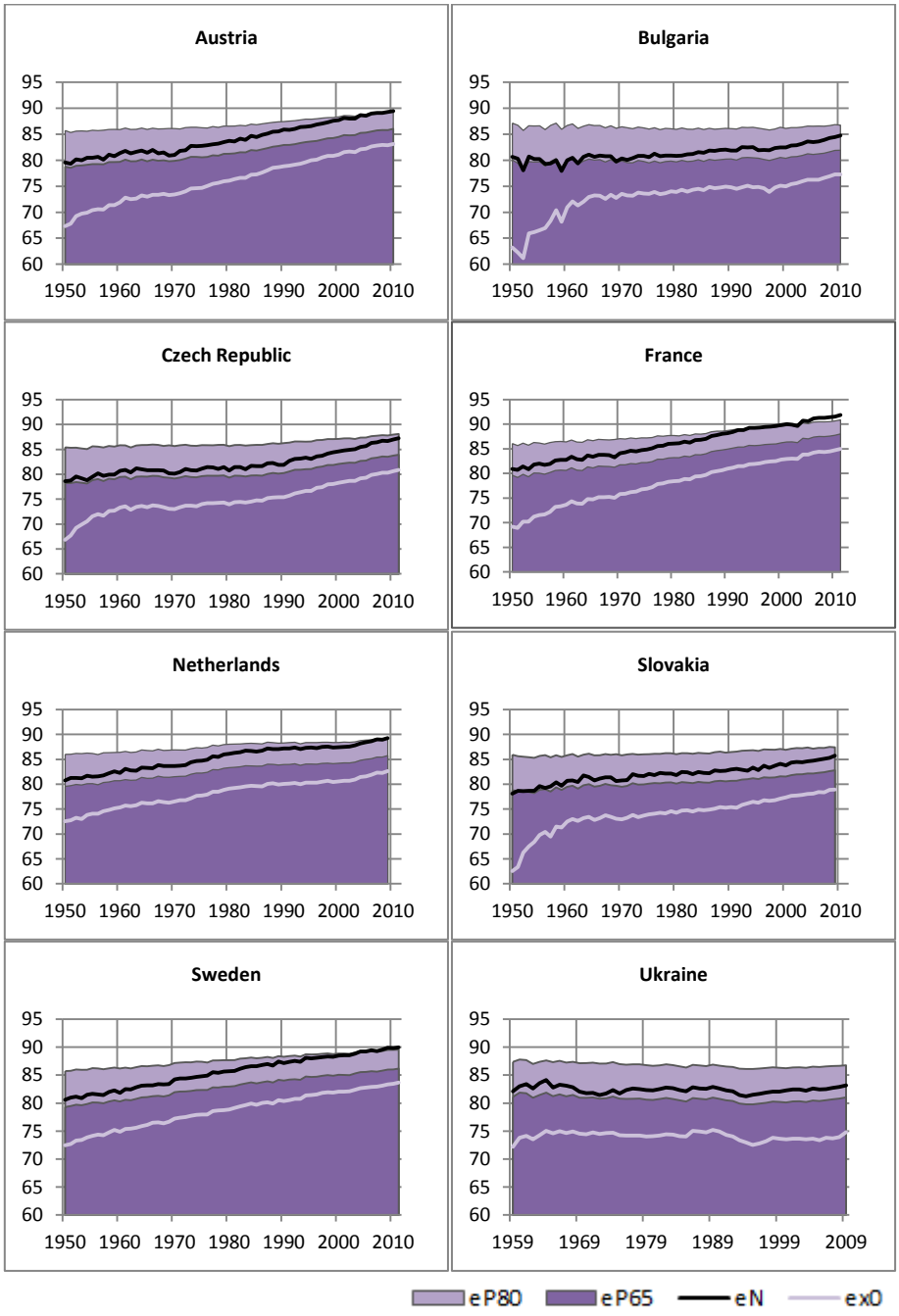


Fig. 5 Modal age at death, life expectancy at birth, probable age at death of 65-year-old and 80-year-old women in selected European countries in 1950-2012

Figure 4 and figure 5 compare the behaviour of modal age at death, life expectancy at birth and probable age at death for both men and women in monitored European countries in 1950-2012. The probable age of death we understand life expectancy for a *x*-year-old person plus age *x* years. Extending life expectancy and modal age at death for men and women are significant in Austria, Czech Republic, France, the Netherlands and Sweden. In Bulgaria, Slovakia and Ukraine the growth tendency of these indicators is not as high as in previous countries. From the figures 4 and 5 it is visible that modal age at death is copying the behaviour of probable age at death of 80-year-old men and women in the last decades and this trend is mainly in Western countries (Austria, France, the Netherlands, Sweden).

4 Conclusions

Life expectancy is influenced by infant mortality and by mortality at younger ages. Modal age at death is influenced by mortality in older ages. That is why it better reflects the typical life expectancy and longevity characteristics. Therefore, it is an appropriate complementary indicator for the evaluation of population aging. In terms of trends in life expectancy and modal age at death in the years 1950-2012 there is an increase of lifespan in most of the selected countries. Different trend is evident in the former socialist countries and in Western countries. In all countries male excess mortality is considerable. Expected future increase in life expectancy and modal age at death will considerably speed up the process of demographic aging. Population growth of Europe stagnates or is at very low levels. Acceleration of population aging means a burden for the productive part of the population and at this level we can not expect a significant improvement. According to projections, future working part of the population will be significantly formed by older people as well, due to improving mortality rates and longer lifespan. Evolution of mortality rates affects the process and intensity of population aging of developed countries. Mortality in the highest age groups in the next period may be affected by changes in the evolution of mortality. On the one hand, increase in the future longevity can be expected and modern technology will greatly help to reduce mortality at old ages (Illes et al., 2007). On the other hand, decrease of mortality can be slowed by epidemic of obesity and diabetes in developed countries (Olshansky et al., 2005). In the following period, these contradictory trends may influence the development of mortality and subsequently reflect on the process of demographic aging of populations of developed countries.

Acknowledgement

This article was supported by the Internal Grant Agency of University of Economics in Prague No. 68/2014 under the title "Economic and health connections of population ageing."

References

1. Canudas-Romo, V. Three Measures of Longevity: Time Trends and Record Values. *Demography: The Population Association of America*. 2010, No. 2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000019/
2. Demopaedia [online]. 2012 [2014-03-25]. http://www.demopaedia.org/
3. Fiala, T. Výpočty aktuárské demografie v tabulkovém procesoru. Praha: Oeconomica, 2005. 177 s. ISBN 80-245-0821-4.
4. Gavrilov, L.A. and Heuveline, P. (2003). Aging of Population. In: Demeny, P. and McNicoll, G. (eds.). *The Encyclopedia of Population*. New York: Macmillan Reference USA: 32-37.
5. Horiuchi S, Wilmoth JR, Pletcher S. "A Decomposition Method Based on a Model of Continuous Change" Demography. 2008;45:785–801.
6. Cheung – Jean-Marie Robine. Dissecting the compression of mortality in Switzerland. Demographic Research [online]. 2009, No. 21 [cit. 2014-03-30]. DOI: 10.4054/DemRes.2009.21.19. http://www.demographic-research.org/volumes/vol21/19/21-19.pdf
7. Illes, J. – A. de Grey – M. Rae. 2007. "Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime." Nature, 450 (7168), 351–352.
8. Kannisto V. "Mode and Dispersion of the Length of Life" Population: An English Selection. 2001;13:159–71.
9. Olshansky, S. J. – D. J. Passaro – R. C. Hershow – J. Layden – B. A. Carnes – J. Brody – L. Hayflick – R. N. Butler – D. B. Allison – D. S. Ludwig. 2005. "A potential decline in life expectancy in the United States in the 21st century." New England Journal of Medicine, 352 (11), 1138–1145.
10. Ouellette, N. – R. Bourbeau. Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research*. 2011, No. 25. DOI: 10.4054/DemRes.2011.25.19. http://www.demographic-research.org/volumes/vol25/19/25-19.pdf
11. Pavlík, Z. et al. *Demografie (nejen) pro demografy*. Praha: SLON, 2009, 241 s. ISBN 978-80-7419-012-4.
12. Vaupel, J. W. et al. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography [online]. 1979, No. 3, 439-454 [cit. 2014-03-30]. http://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/2d6493f9f0d93b50c125774b0045c00b/$FILE/Vaupel-Demography-16-1979-3.pdf
13. Wilmoth John R. "Demography of Longevity: Past, Present and Future Trends" Experimental Gerontology. 2000; 35:1111–29.
14. Wilmoth, John R. – S. Horiuchi. 1998. Do the oldest-old grow old more slowly? Paper presented at *Collogue médicine et recherche*, Paris
15. Wilmoth, John R. – S. Horiuchi. Rectangularization Revisited: Variability of Age at Death within Human Populations. Demography [online]. 1999, No. 36, 475-495 [cit. 2014-03-30]. http://www.jstor.org/stable/2648085

Comparing the Cumulative Rates of Cancer in two Regions

Christopher T. Lenard[1], Terence M. Mills[2], and Ruth F.G. Williams[3]

[1] Mathematics and Statistics, La Trobe University, Bendigo, Victoria, Australia
(E-mail: c.lenard@latrobe.edu.au)
[2] Loddon Mallee Integrated Cancer Service, Bendigo, Victoria, Australia
(E-mail: tmills@bendigohealth.org.au
[3] School of Economics, La Trobe University, Bendigo, Victoria, Australia
(E-mail: ruth.williams@latrobe.edu.au)

Abstract. As the incidence of cancer continues to rise, it is natural for a community to want to compare the incidence of cancer in the region with the incidence of cancer in another region, such as the rest of the nation. The cumulative incidence rate is a measure that was introduced in the cancer literature in 1976. This measure is easy to calculate and facilitates comparing the incidence of cancer in two regions. The aim of this paper is to promote this measure by means of a worked example based on illustrative data.
Keywords: Epidemiology, Health economics, Health services research, Health funding and financing

1 Introduction

"A cancer, or malignant growth, is now known to be a continuous, purposeless, unwanted, uncontrolled and damaging growth of cells." (Stephens and Fox[8].) There are many types of cancer - prostate cancer, lung cancer, bowel cancer, breast cancer to name a few. Although the term "cancer" refers to a broad range of diseases, we will use the term "cancer" to refer, collectively, to all malignant cancers. This is the practice of cancer agencies such as the Australian Institute of Health and Welfare[1].

The incidence of cancer is the number of new cases diagnosed in a particular region and a particular period of time, usually a year. Thus, incidence is a non-negative integer. The incidence rate is the incidence per 100,000 head of population.

It is natural to want to compare the incidence rates of cancer in two regions. For example, one may wish to compare the incidence rate of cancer in a small regional area with that in the rest of the nation. One may also wish to compare the incidence rate of cancer in a region at a particular time with that in the same

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

region at an earlier time. This could lead to a time series approach to tracking the incidence of cancer in a region over time.

Age is a risk factor associated with cancer (Mills[7]). Hence the incidence rate of cancer will tend to be higher in older populations, all other things being equal. The traditional approach to dealing with this difference in age profiles is to use age standardized incidence rates. This involves choosing some standard population on which to base the calculations (Estève *et al.*[4], p. 56).

The concept of cumulative rate of cancer was proposed as an alternative to the age standardized incidence rate by the distinguished epidemiologist N.E. Day[2] in 1976. The cumulative rate has the advantage that it avoids the arbitrariness of having to choose a pre-defined, standard population. The cumulative rate is also directly connected to the cumulative risk, or actuarial risk, of being diagnosed with cancer by a given age. For discussions of this connection, see works by Day[2], Estève *et al.*([4] p. 60), and Lenard *et al.* [5].

The primary aim of this paper is to demonstrate how one compares the cumulative rates of cancer in two regions. This is a fundamental problem for practitioners.

We will achieve this by means of a worked example that is based on data that are hypothetical, but realistic. The data are purely illustrative. This example can be used as a model by those who wish to make such comparisons in practice.

Despite the advantages of the cumulative rate, it is not often used in practice. The secondary aim of this paper is to promote further discussion of the cumulative rate.

2 Methods

Table 1 contains hypothetical data for two regions. We have chosen to use hypothetical, or illustrative, data because our aim is to demonstrate the method rather than compare the incidence of cancer in two particular regions. The data are presented in *5*-year age groups. For each age group, the number of persons in the population and the incidence of cancer are presented for each region. For example, in Region 1, there are 31,294 persons aged between 40 and 44, and 81 of these persons were diagnosed with cancer in the year in question.

Age group	Region 1		Region 2	
	Pop1 (n)	Inc1 (x)	Pop2 (m)	Inc2 (y)
0-4	29289	12	498381	118
5-9	29202	6	464389	55
10-14	31749	7	461565	66
15-19	32427	9	500763	91
20-24	27235	6	591423	175
25-29	24195	19	612961	313
30-34	23986	16	563386	492
35-39	27952	48	564801	793
40-44	31294	81	567399	1186
45-49	32571	132	540007	1969
50-54	33412	205	512283	2883
55-59	31051	274	455982	3675
60-64	30033	370	421669	5064
65-69	23739	423	325045	5382
70-74	18709	412	251007	4845

Table 1: Hypothetical population data and incidence data for two regions

Details of the methods, and the mathematical ideas that underpin them, have been discussed elsewhere; for example see Estève *et al*.[4] and Lenard *et al*.[6]. Here we present only the formulae that are necessary for the calculations.

Cumulative rate by age 75 for Region 1 := CumRate1 = $5\Sigma(x/n)$.

Approximate cumulative risk of being diagnosed with cancer by age 75 for Region 1 is $1-\exp(-5\Sigma(x/n))$.

Estimated standard deviation of cumulative rate by age 75 for Region 1 := $s_1 = 5\sqrt{(\Sigma(x/n^2))}$.

Alternatively, one could use the formulae in Dobson *et al.*[3] for estimating the standard deviation of the cumulative rate.

For Region 2, the formulae are similar: substitute y for x, and m for n.

The z-statistic for comparing the cumulative rates for the two regions is

$$z := (\text{CumRate1} - \text{CumRate2})/\sqrt{[(s_1)^2 + (s_2)^2]}.$$

Under the null hypothesis that the two populations have the same expected cumulative rates, this z statistic has, approximately, the standard Normal distribution.

Suppose that we expect, from experience, that the cumulative rate of cancer in Region 1 is larger than the cumulative rate in Region 2. Then we would conduct a one-sided statistical test and calculate the probability, p, that $Z > z$ where Z has the standard Normal distribution.

3 Results

The results of the analysis of the data are presented in Table 2. We have included the approximate cumulative risks only as a matter of interest; our main focus is on the cumulative rates.

	Region 1	**Region 2**
Cumulative rate by age 75	0.3913	0.3553
Approx. cumulative risk by age 75	0.3238	0.2990
Est. s.d. of cum. rate by age 75	0.0089	0.0022
z	3.9193	
$p=P(Z > z)$	4.4404E-05	

Table 2: Results of analysis of data in Table 1.

In this example, the p-value associated with the z-statistic is very small (p = 4.4404E-05). Hence, the data provide strong evidence, in which considerable confidence can be placed, that the cumulative rate of cancer up to age 75 is higher in Region 1 than in Region 2.

4 Conclusions

This paper has been written for practitioners who wish to use the cumulative rate to compare the incidence of cancer in two different regions. The cumulative rate might be used in making decisions about allocating resources for cancer care to different regions.

Although Estève *et al.* ([4], pp. 74-84) discuss methods for comparing the incidence of a disease in two populations, they do not discuss the use of the cumulative rate in this context. The present paper fills this gap in the literature.

The cumulative rate proposed by Day[2] avoids the arbitrariness of a pre-defined standard population on which to base the calculations. The only data required are the population data and incidence data stratified by *5*-year age groups. It is possible to deal with age groups of other widths; see Lenard *et al*. [6] as to how this might be done.

The method can be easily adapted to consider the cumulative rate of particular cancers. For example, in considering the cumulative rate of breast cancer among

female Australians, one would tabulate the population and incidence of female Australians.

Day[2] recommends that, for whole of life comparisons, 74 is an appropriate maximum age; there are many competing risks for people over this age. For childhood cancers, Day suggests that one might consider the maximum age as 14.

One could also use this method to compare the cumulative rates of cancer in one region in two different years.

It would be interesting to investigate multiple comparison procedures for comparing the cumulative rates of several regions. For example, if one were allocating resources for cancer care to four regions A, B, C, D it would be useful to be able to say that the cumulative rate in A was significantly higher than in B, C, and D but there is no significant difference between the rates in B, C, and D. Then one would have a sound basis for allocating equal resources to regions B, C, D and more to A.

Acknowledgement

We thank Professor Christos Skiadas and the organisers of SMTDA 2014 for the opportunity to present our ideas in this forum.

References

1. Australian Institute of Health and Welfare & Australasian Association of Cancer Registries. Cancer in Australia: an overview, 2012. Cancer series no. 74. Cat. no. CAN 70. Canberra, AIHW, 2012.
2. N.E. Day. A new measure of age standardized incidence, the cumulative rate. In: Payne R, Waterhouse J. eds. Cancer incidence in Five Continents. Vol. III. IARC Scientific Publications, No. 15. Lyon. International Agency for Research on Cancer, 443-452, 1976.
3. A.J. Dobson, K. Kuulasmaa, E. Eberle and J. Scherer. Confidence intervals for weighted sums of poisson parameters. Statistics in Medicine, 10, 3, 457–462, 1991
4. J. Estève, E. Benhamou, L. Raymond. Statistical Methods in Cancer Research, Vol. IV: Descriptive Epidemiology. Lyon, International Agency for Research on Cancer 1994.
5. C.T. Lenard, T.M. Mills and R.F.G. Williams. The risk of being diagnosed with cancer. Aust N Z J Public Health, 37, 5, 489, 2013.
6. C.T. Lenard, T.M. Mills and R.F.G. Williams. Cumulative incidence rates of cancer. The Mathematical Scientist (to appear).
7. T.M. Mills. Age and cancer. Significance Magazine. On-line: Available at URL: http://www.statslife.org.uk/health-medicine/892-age-and-cancer , 19 July, 2013.
8. F. Stephens and R. Fox. Cancer explained: The essential guide to diagnosis and management. Second Ed., Random House Australia, North Sydney, 2008.

Are credit ratings time-homogeneous and Markov?

Pedro Lencastre[1,2], Frank Raischel[3], Pedro G. Lind[4], Tim Rogers[5]

[1] ISCTE-IUL, Av. Forças Armadas, 1649-026 Lisboa, Portugal
(e-mail: pedro.lencastre.silva@gmail.com)
[2] Mathematical Department, FCUL, University of Lisbon, 1749-016 Lisbon, Portugal
[3] Instituto Dom Luiz, University of Lisbon, 1749-016 Lisbon, Portugal
(e-mail: raischel@cii.fc.ul.pt)
[4] ForWind and Institute of Physics, University of Oldenburg, Ammerländer Heerstrasse 136, DE-26111 Oldenburg, Germany
(e-mail: pedro.g.lind@forwind.de)
[5] Centre for Networks and Collective Behaviour, Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, Bath, UK

Abstract. We introduce a simple approach for testing the reliability of homogeneous generators and the Markov property of the stochastic processes underlying empirical time series of credit ratings. We analyze open access data provided by Moody's and show that the validity of these assumptions - existence of a homogeneous generator and Markovianity - is not always guaranteed. Our analysis is based on a comparison between empirical transition matrices aggregated over fixed time windows and candidate transition matrices generated from measurements taken over shorter periods. Ratings are widely used in credit risk, and are a key element in risk assessment; our results provide a tool for quantifying confidence in predictions extrapolated from rating time series.
Keywords: Generator matrices,Continuous Markov processes,Rating matrices,Credit Risk.

1 Motivation and Scope

After the Basel II accord in 2004 [1], ratings became an increasingly important instrument in Credit Risk, as they allow banks to base their capital requirements on internal as well as external rating systems. These ratings became instrumental in evaluating the risk of a bond or loan and in the calculation of the Value at Risk. As such, it is often desirable to quantify the uncertainty in these ratings, and predict the likelihood that an institution will be upgraded or downgraded in the near future. A common technique is to aggregate credit rating transition data over yearly or quarterly periods, and to model future transitions using these data. However, to be reliably the ratings' evolution

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

must obey particular features which we show below can be evaluated through analysis of the data published by rating agencies. Two sufficient properties for accepting the empirical data as a reliable indicator of future rating evolution are the existence of one generator and Markovianity.

The representation of the evolution of a time-continuous process by an aggregated transition matrix will not be adequate if the underlying process is not Markov. Moreover, if there is no generator associated to the transition matrix, the process underlying the ratings is not continuous. Different techniques to estimate a transition matrix from a finite sample of data should be employed depending on whether the process is time-homogeneous or not[2,3]. Theoretically, both the Markov and the time-homogeneous assumptions simplify considerably the models in question[4], but typically only the latter is at times dropped in order to build a more general theoretical framework.

In this paper we test how good both assumptions are in different periods of time for a homogeneous rating class in Moody's database. We compare transition matrices calculated under different assumptions and show that the quality of the time-homogeneous and Markov assumptions change considerably in time. Moreover, we argue that the wide fluctuations of the assumptions' quality may on the one hand provide evidence for detecting discontinuities in the rating process, e.g. when establishing new evaluation criteria for a bank rating, and, on the other hand, can be taken as a tool for ascertaining how complete and trustable such rating criteria are.

We start in Sec. 2 by describing the empirical data collected from Moody's and in Sec. 3 we describe how to test the validity of both the homogeneity and Markovianity assumptions. Section 4 concludes the paper and presents some discussion of our results in the light of finance rating procedures.

2 Data: Six Years of Rating Transitions in Europe

The data analyzed in this paper is publicly available data that Moody's needs to disclose and keep publicly available in compliance with Rule 17g-2(d)(3) of US. SEC regulations [5].

The rating time series of each bank has a sample frequency of one day, starting in January 1st 2007 and ending in January 1st of 2013. The data sample is the set of rating histories from the banks, in European countries, that had a rating at the final date. Each value indicates the rating class, according to the so-called *Banking Financial Strength*[6], at which the bank is evaluated at that particular day.

One first important feature of this rating database is its non-stationary character, as can be seen in Fig. 1. The number of banks N_R included in the data set increased almost monotonically during the total time-span analyzed by us (see Fig. 1a). On January 1st 2007 there are $N_R = 658$ rated companies in the data set, and this number increases until 2013 when one registers $N_R = 924$ rated banks. Therefore, we will consider our measures normalized to the number of banks in the database.

We count in Moody's database a total of $N_T = 932$ rating transitions, that distribute heterogeneously in time. Indeed, the number of transitions N_T per

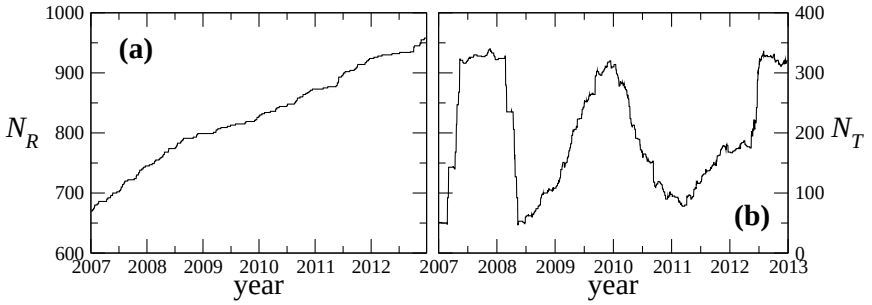


Fig. 1. (a) Number of bank entities in Moody's data sample as a function of time and **(b)** the number of transitions per bank, computed as moving averages during one-year periods.

bank also changes significantly, with three events of peaked activity, namely during the year of 2007, at the beginning of 2010 and in the last half year of 2012 (see Fig. 1b). This will be of importance when analyzing the evolution of the generator homogeneity and Markovianity of the corresponding transition matrices.

The rating category is a measure of the capacity of the institution to meet its financial obligations and avoids default or government bailout. We have $n_s = 15$ rating states, denoted by the letters A to E in alphabetic order and with the two possible extra suffixes, namely $+$ and $-$. State $A+$ represents the state corresponding to the best financial health and less credit risk, followed by A, $A-$, $B+$ and so on, until the bottom of the scale, $E-$, the state that represents the highest risk level. Figure 2 shows three plots (left) illustrating the histogram of rating states at three different time, namely the first day of 2007, 2009 and 2010.

Henceforth, we define $\tilde{R}_i(t)$ as the rating of the bank number i at the moment t, and we map the rating states to an increasing ordered number series: state $\tilde{R} = E+$ corresponding to label $R = 0$, and state $\tilde{R} = A+$ to label $R = 14$. With such a labelling it is possible to compute rating increments as

$$T_i(t, \tau) = R_i(t) - R_i(t - \tau). \tag{1}$$

When $T_i(t) > 0$ (resp. < 0) it means that bank i saw its rating increased (resp. decreased) during the last τ period of time. Unless stated otherwise we will use always $\tau = 365$ days. The plots in the right column of Fig. 2 show the histograms of the corresponding rating increments at the same three days.

We call henceforth $R(t)$ and $T(t)$ the aggregated processes of the ratings and rating increments respectively, over all N_R companies observed at time t. Figure 3 shows the evolution of the first four moments for both rating distributions (left) and transition distributions (right), with $\tau = 365$ days.

The average rating $\langle R \rangle$ (Fig. 3a) has decreased during most of the six year period records. We should note however that this is due to the new entries in the database whose initial rating is typically low, since $\langle T \rangle$ has positive periods during the first five years of the recorded set.

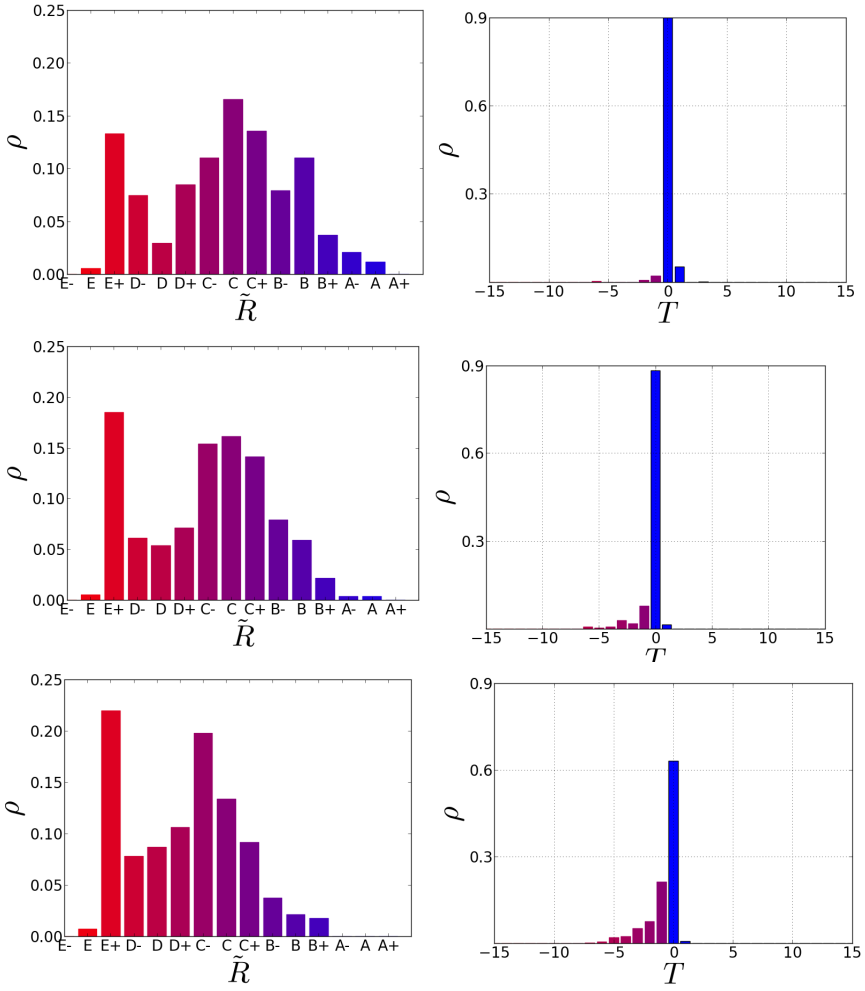


Fig. 2. Illustration of rating histograms for the rating state $\tilde{R}$ (left) and the corresponding rating variations $T = \Delta R$ (right), where R is an integer enconding the rating state, ranging from 0 ($E-$) to 14 ($A+$). Three different days are selected: first day of 2007 (first row), 2009(second row) and 2010 (third row); cf. Fig. 3.

As for the rating variance σ_R (Fig. 3e), after a slight increase, it also decreased since the middle of 2007, due to the concentration of rating states to the lower rating classes ($\langle T \rangle < 0$). The transitions however exhibit two periods of increased variance σ_T (Fig. 3f), which reflect probably the respective increase in the number of transitions (compare with Fig. 1b).

As the lowest states get more and more dominant, the rating skewness μ_R (Fig. 3c) increases steadily, until it changes sign around 2008, when transitions become negative on average. These two observations are consistent with each other: the negative skewness indicates the large majority of banks being below the average rating which corresponds to an average decrease of the rating $\langle T \rangle <$

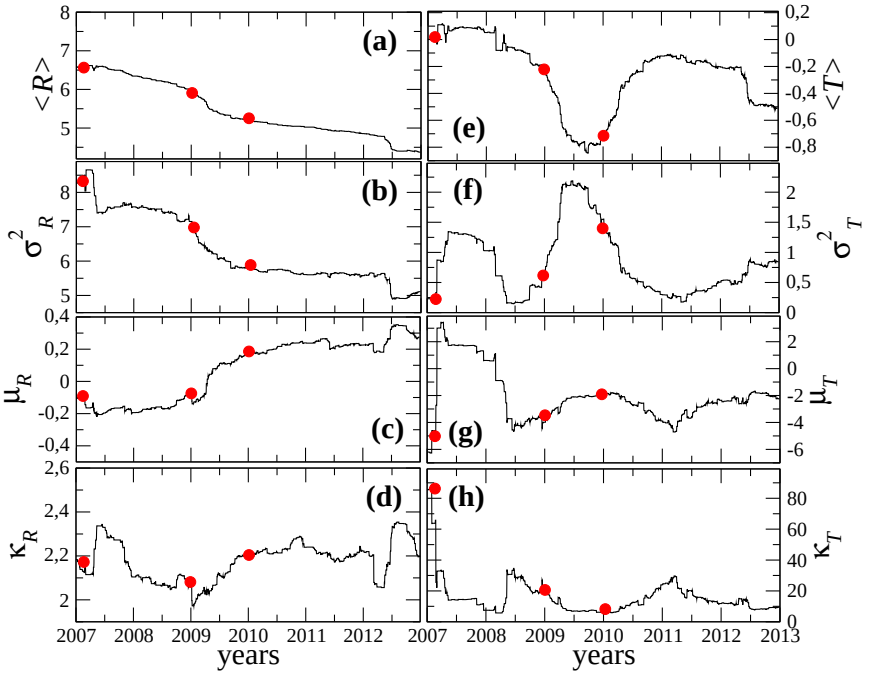


Fig. 3. Evolution of the first statistical moments of **(a-d)** the rating state R distribution and **(e-h)** its one-year-increment T distribution. From top to bottom: averages $\langle R\rangle$ and $\langle T\rangle$, variances σ_R^2 and σ_T^2, skewnesses μ_R and μ_T, and kurtosis κ_R and κ_T. Bullets indicate the days when the histograms in Fig. 2 were taken.

0. It also indicates that there are a few banks highly rated. This observation together with the observations regarding temporal homogeneity in the next section will justify some comments about the objectiveness of rating criteria.

The rating distribution is also typically platykurtic (see Fig. 3d), as its kurtosis is always below three (Gaussian kurtosis), indicating a more pronounced flatness around the average of rating distributions. Concerning the third and fourth moments of transition distributions, Figs. 3g and 3h respectively, we see large fluctuations during the periods with fewer transitions. One can clearly sees a very high kurtosis, and changes in the sign of the mean and skewness.

3 What is the Underlying Continuous Process?

In the following we assume that the set of rating transitions has a continuous processes underlying it, an assumption which has been the subject of previous investigations without a clear result, see e.g. Ref. [7]. Even in case that there is a continuous process, the corresponding generator may be constant (homogeneous generator) or vary in time (non-homogeneous).

The non-homogeneity is important in the finance context since it limits the range of models that can be used. In particular, it has been argued [2] that if we consider time-homogeneity a method for estimating a transition matrix better

than the one for the more general case. The main advantages of this method are to capture very small transition probabilities between two states, even when no transitions occurred between those two states, and to distinguish between transitions within the studied time-frame. The time-homogeneity condition is also important to check if the rating philosophies[8,9] allegedly used are being correctly followed or not, and they do not hold if criteria by which ratings are ascribed to banks are not constant in time, but vary according to artificial or externally imposed factors[10].

Furthermore, another important feature of continuous transition processes is their Markovianity. The Markov property is important if the current rating of a bank is to be considered a complete indicator of its future risk. In this section we will address both these conditions separately.

3.1 Testing Time-Homogeneity

Mathematically, if a time-continuous Markov process is time-homogeneous then there is a constant matrix $\mathbf{Q}$, called a generator, solution of

$$\frac{d\mathbf{M}(t)}{dt} = \mathbf{Q}\mathbf{M}(t), \tag{2}$$

where $\mathbf{M}$ is the transition matrix, with entries M_{ij} given the probability for observing a transition from state i to state j $(i, j = 1, \ldots, n_s)$. In other words, a time-continuous process is time-homogeneous if, being Markov, its transition matrix can be expressed as $\mathbf{M}(t) = e^{\mathbf{Q}t}$, and therefore it has a well-defined logarithm. We take the analogue from ordinary differential equations and loosely call $\mathbf{Q}$ the logarithm of $\mathbf{M}$.

The mathematical conditions for the existence of a homogeneous generator give a bivalent result[7,11] that does not take into consideration neither noise generated from finite samples nor how distant an empirical process is from being time-continuous. Therefore, we neglect several mathematical results that determine if a generator exists or not, and assume that the process is Markov and time-continuous. Being Markov and time-continuous means that there is a generator satisfying Eq. (2) and that it either is constant or varies in time.

Next, we estimate the closest constant generator $\mathbf{Q}$ directly from the empirical data, compute the associated matrix $\mathbf{M} = e^{\mathbf{Q}t}$, and compare it with the empirical transition matrix $\mathbf{M}^{(e)}$. For estimating the generator matrix $\mathbf{Q}$ we follow the approach described in Ref. [3], calculating its off-diagonal elements as

$$Q_{ij} = \frac{N_T^{(ij)}}{\int_{t_0}^{t_f} N_R^{(i)}(t)dt}, \tag{3}$$

where $N_T^{(ij)}$ represents the number of transitions from i to j between the times t_0 and t_f, and $N_R^{(i)}(t)$ stands for the number of banks in state i at time t. The diagonal elements Q_{ii} follow from the condition $\sum_j Q_{ij} = 0$.

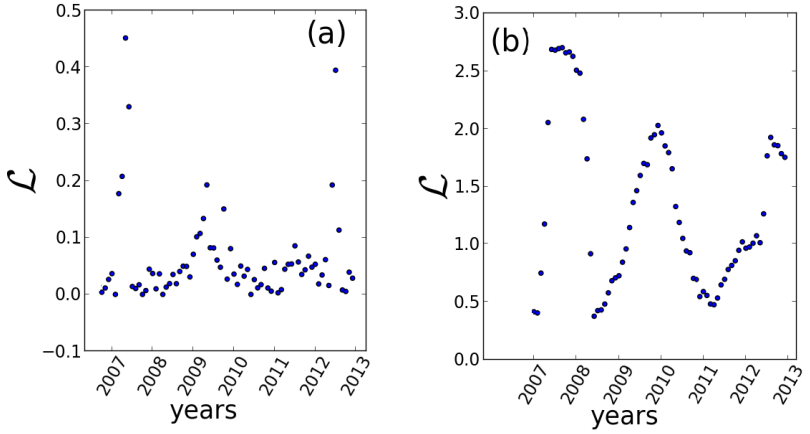


Fig. 4. Testing for temporal homogeneity: difference between the log-likelihood $\mathcal{L}$ of the transition matrix $\mathbf{M}^{(e)}$ and the transition matrix $\mathbf{M}$ calculated assuming time-homogeneity. Both matrices are calculated over a time interval **(a)** one month and **(b)** one year. The log-likelihood was calculated using Eq. (4) at the first day of each month from January 2007 to December 2012.

To compute the distance between a time-homogeneous process and the empirical process we compare $\mathbf{M}$ with $\mathbf{M}^{(e)}$, and plot the statistic:

$$\mathcal{L} = \frac{\sum_{i,j} N_T^{(ij)} \left(\log M_{ij} - \log M_{ij}^{(e)} \right)}{\sum_{i,j} N_T^{(ij)}} \,. \tag{4}$$

This is a log-likelihood ratio; loosely speaking it quantifies the error introduced by making the assumption of time homogeneity. The results are shown in Fig. 4: in panel (a) we aggregate the data in periods of one month while in panel (b) the aggregation period is one year.

It can be seen that there are three periods when the time-homogeneity condition becomes an insufficient approximation to the dynamics of the process marked by significant increases in $\mathcal{L}$. The first period starts in the early 2007, the second period around the middle of 2009, and the third period in the last half of 2012. The profile of the time-inhomogeneity is different for each time-period. It shows a sharp peak in 2007, concentrated in just a few months, and wider in the other periods.

These three periods can be better analysed taking also observations from Fig. 3. In 2007 there was an unusually high number of rating transitions, even considering that only about 700 companies were rated at the time. In Fig. 3 it can be seen that in this period the variance σ_R of the ratings decreased, the skewness μ_R had slight negative burst, and there was an increase in the kurtosis κ_R. As for the statistics of transitions, one can see in this period the average $\langle T \rangle$ becoming positive, the skewness (μ_T) changing signal and becoming positive and the kurtosis (κ_T) decreasing. The variance of T increases, but again that can be explained by the high number of rating transitions in that period.

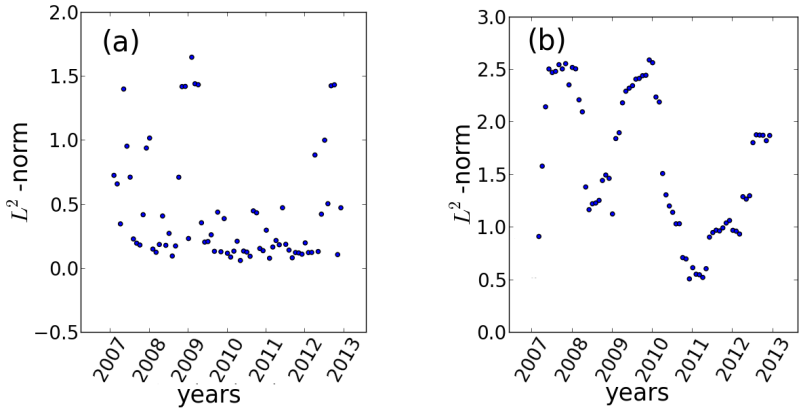


Fig. 5. Testing Markovianity: difference between the empirical transition matrix $\mathbf{M}^{(e)}_{0\tau}$ calculated over a time-interval $[0, \tau]$ and the product of the half-period matrices, $\mathbf{M}^{(e)}_{0\frac{\tau}{2}}$ and $\mathbf{M}^{(e)}_{\frac{\tau}{2}\tau}$, using the L^2-norm defined in Eqs. (7) and (8). Both matrices are calculated over a time interval of **(a)** one month and **(b)** one year. The difference was calculated at the first day of each month between January 2007 and December 2012.

In late 2009 and early 2010 we have a very different profile. In this period the downgrades are the rule, as one can see by the negative values of $\langle T \rangle$. The relatively low values of κ_T and the absolute value of μ_T tells us that this was a general trend, and not a very drastic movement by just a few banks.

In 2012 the scenario is similar to 2010. Again there are more downgrades, and this a general trend. The companies are now much more clustered, i.e. with short dispersion in their ratings, as one can see by the low values in σ_R.

3.2 Testing the Markov Hypothesis

Mathematically, a Markov process x_t obeys the following condition:

$$\Pr(x_{t_1}|x_{t_2}, x_{t_3}, \dots) = \Pr(x_{t_1}|x_{t_2}) \tag{5}$$

with $t_1 > t_2 > t_3 > \dots$. The conditional probability in the right hand-side of Eq. (5), $\Pr(x_{t_1}|x_{t_2})$, is exactly specified by the transition matrix $\mathbf{M}$.

The rating process must be assumed to be Markov, otherwise a rating would not represent a uniform risk class, as its elements could be distinguished according to their previous series of rating states.

From the definition of a Markov process in Eq. (5) it is straightforward to show that a Markov process also obeys

$$\mathbf{M}_{t_0 t_f} = \prod_{n=1}^{N} \mathbf{M}_{t_{n-1} t_n}, \tag{6}$$

where N is the number of subintervals in $[t_0, t_f]$ and labels $t_i t_j$ denote the time interval $[t_i, t_j]$ considered when determining $\mathbf{M}_{t_i t_j}$. Here we fix $N = 2$ and

consider two equally spaced intervals with $\tau \equiv t_f - t_0 = 1$ month and $\tau = 1$ year. Equation (6) is known as the Chapman-Kolmogorov equation[12] and it does not hold in general either when the process is non-Markov or when we have an insufficiently short sample of data.

We will use the Chapman-Kolmogorov equation as a test indicating whether the rating database of Moody's is Markov. To that end, we consider empirical matrices $\mathbf{M}^{(e)}_{0\tau}$ computed for one month and one year intervals, and compare it with the associated product of the two corresponding half-periods, $\overline{\mathbf{M}^{(e)}}_{0\tau} = \mathbf{M}^{(e)}_{0\frac{\tau}{2}}\mathbf{M}^{(e)}_{\frac{\tau}{2}\tau}$. For the comparison we now use the L_2-norm instead of the the $\mathcal{L}$ log-likelihood, since the latter creates singularities when dealing with zero entries in the matrices, and which occur now more frequently. The L_2-norm of the transition matrix is the maximum singular value of $\mathbf{A}$,

$$\|\mathbf{A}\| = \sigma_{\max}(\mathbf{A}), \tag{7}$$

and we compute it for as the difference

$$\mathbf{A} = \mathbf{M}^{(e)}_{0\tau} - \overline{\mathbf{M}^{(e)}}_{0\tau}\,, \tag{8}$$

where $\|\cdot\|$ represents the usual Euclidian norm.

Results are shown in Fig. 5. Clearly, there are two periods when the Markov assumption seems less valid. The first period is in early 2007, and the second in the middle of 2009, followed by another, less significant increase at the end of 2012. As said before, this coincides with an abrupt change in the statistics of T and R.

4 Discussion and conclusions

We have addressed time series of credit ratings publicly available at Moody's online site and studied simple ways to compute the validity of the time-homogeneous and Markovianity assumptions. We have shown how the accuracy of these assumptions varies with time. Naturally, when the Markov assumption fails, so does the time-homogeneous assumption, in particular during 2007 and in the latest half of 2009 and beginning of 2010. In these periods the statistics of the process changed considerably. In the end of the year of 2012 the accuracy of the time-homogeneous assumption is low but the Markov approximation is within the usual fluctuation range. In this period there is a less abrupt change in the statistics of the process.

One must stress that when the Markov assumption does not hold, the ratings are not a complete measure of the risk of a given entity, since further information besides the actual rating needs to be specified. Moreover, our results present evidence that perhaps in 2007 new rating criteria were introduced, imposing a discontinuity in the series of ratings, or that new rating transition were correlated with previous ones, which could support the claim that rating agencies were an active part in the crisis that followed.

Our approach can be improved by introducing for instance a more sophisticated procedure for extracting the histograms for the ratings and their increments, namely using the kernel based density, which is known to converge

faster to the real distribution than the usual binning procedure. From this first approach to investigate Moody's rating database one can now attack the embedding problem for the series of transition matrices, where different generators estimates can be compared. These and other issues will be addressed elsewhere.

Acknowledgments

The authors thank Fundação para a Ciência e a Tecnologia for financial support under PEst-OE/FIS/UI0618/2011, PEst-OE/MAT/UI0152/2011, FCOMP-01-0124-FEDER-016080, SFRH/BPD/65427/2009 (FR). This work is part of a bilateral cooperation DRI/DAAD/1208/2013 supported by FCT and Deutscher Akademischer Auslandsdienst (DAAD). PL thanks Global Association of Risks Professionals (GARP) for the "Spring 2014 GARP Research Fellowship".

References

1. Basel Committee on Banking Supervision. Basel II: International convergence of capital measurement and capital standards: a revised framework. *(available at http://bis.org/publ/bcbs107.htm)*, 2004.
2. T. Charitos, P.R. de Waal, and L. C. van der Gaag. Computing short-interval transition matrices of a discrete-time markov chain from partially observed data. *Statistics in medicine*, 27(6):905–921, 2008.
3. D. Lando and T. M. Skødeberg. Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, 26(2):423–444, 2002.
4. R. Weißbach, P. Tschiersch, and C. Lawrenz. Testing time-homogeneity of rating transitions after origination of debt. *Empirical Economics*, 36(3):575–596, 2009.
5. https://www.moodys.com/pages/reg001004.aspx.
6. Moody's Investors Service. Moody's rating symbols & definitions. *Report*, 79004(08):1–52, 2004.
7. R.B. Israel, J.S. Rosenthal, and J.Z. Wei. Finding generators for markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2):245–265, 2001.
8. J. Kiff, M. Kisser, and L. Schumacher. An inspection of the through-the-cycle rating methodology. *IMF Working Paper*, 2013.
9. Z. Varsanyi. Rating philosophies: some clarifications. *Report*, 2007.
10. J. Mathis, J. McAndrews, and J.-C. Rochet. Rating the raters: are reputation concerns powerful enough to discipline rating agencies? *Journal of Monetary Economics*, 56(5):657–674, 2009.
11. EB Davies. Embeddable markov matrices. *Electronic Journal of Probability*, 15:1474–1486, 2010.
12. H. Risken. *The Fokker-Planck Equation*. Springer, Berlin, 2nd edition, 1989.

Asian Options, Jump-Diffusion Processes on a Lattice, and Vandermonde Matrices

Karl Lundengård[1], Carolyne Ogutu[2], Sergei Silvestrov[1], and Patrick Weke[2]

[1] Division of Applied Mathematics, School of Education, Culture and Communication, Mälardalen University, Box 883, SE-721 23 Västerås, Sweden (e-mail: karl.lundengard@mdh.se, sergei.silvestrov@mdh.se)
[2] School of Mathematics, University of Nairobi, Box 30197 - 00100, Nairobi, Kenya (e-mail: cogutu@uonbi.ac.ke, pweke@uonbi.ac.ke)

Abstract. Asian options are options whose value depends on the average asset price during its lifetime. They are useful because they are less subject to price manipulations. We consider Asian option pricing on a lattice where the underlying asset follows MertonBates jump-diffusion model. We describe the construction of the lattice using the moment matching technique which results in an equation system described by a Vandermonde matrix. Using some properties of Vandermonde matrices we calculate the jump probabilities of the resulting system. Some conditions on the possible jump sizes in the lattice are also given.
Keywords: Jump-diffusion process, lattices, Vandermonde matrix, Asian options, option pricing.

1 Pricing of Asian options and jump-diffusion option pricing

Asian options are path dependent options whose payoffs depend on the average price during a specific period of time before maturity. The averages are considered to be either geometric or arithmetic averages. Assuming the geometric average results in a closed-form formula for the European option price within the classical Black–Scholes model. This is because the geometric average of log-normally distributed random variables also has a lognormal distribution and this simplifies the mathematics involved in the pricing problem. In contrast, the arithmetic average of lognormal random variables is not log-normally distributed, thus there exists no closed-form formula for European Asian options based on the arithmetic average of the underlying asset prices. There are, however, approximation methods that have been developed to aid in pricing Asian options, here we will consider lattice methods. Lattice methods are based on a discrete approximation of the process such that the time span is divided into n time steps and specifies asset price at each time step. At each time step the process can jump to L different asset prices, henceforth referred to as nodes.

3rd SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal
C. H. Skiadas (Ed)

We will model the options underlying asset using a Merton-Bates jump-diffusion process. Popular methods for pricing options of processes include binomial methods ($L = 2$, see Cox *et al.*[2], Amin[1], Hilliard and Schwartz[4] among others) and trinomial methods ($L = 3$, see Dai *et al.*[3] among others). For any of these methods it is required that the first L moments match the asset return and that the probabilities of moving to any given node is between zero and one.

Here we will consider multinomial methods with higher L. For further details on the construction and use of this type of method see Lundengård *et al.*[6].

2 Moment-matching multinomial lattice methods

We want to match the moments of a random variable X with a discrete random variable Z. Let Z denote a discrete random variable as given below (Primbs *et al.*[8]):

$$Z = m_1 + (2i - L - 1)\,\alpha \quad \text{with probability } p_i,\ i = 1, 2, ..., L,$$

where α is the jump size (distance between two outcomes), m_1 is the mean of X and L is the number of lattice nodes. Here α must be real and positive.

The requirement that the k:th moment matches the asset return on an L-node lattice can be written:

$$\sum_{i=1}^{L} p_i (2i - L - 1)^k \alpha^k = \mu_k$$

where μ_k is the k:th moment. We will also use the notation $\mu_0 = 1$ which means that matching to μ_0 is equivalent to the sum of all probabilities being equal to one.

Matching the first L moments can be written $A\boldsymbol{p} = \boldsymbol{\mu}$ where $\boldsymbol{p}$ is a column vector containing the jump probabilities, $\boldsymbol{\mu}$ is a column vector containing the moments and A is the general lattice matrix for the jump diffusion process that takes the following form:

$$A = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ (1-L)\alpha & \cdots & (2n-L-1)\alpha & \cdots & (L-1)\alpha \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ ((1-L)\alpha)^k & \cdots & ((2n-L-1)\alpha)^k & \cdots & ((L-1)\alpha)^k \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ ((1-L)\alpha)^L & \cdots & ((2n-L-1)\alpha)^L & \cdots & ((L-1)\alpha)^L \end{bmatrix}.$$

Note that the general lattice matrix A is a *Vandermonde matrix*.

Definition 1. A *Vandermonde matrix* is a square matrix of the form

$$V_L = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_L \\ x_1^2 & x_2^2 & \dots & x_L^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{L-1} & x_2^{L-1} & \dots & x_L^{L-1} \end{bmatrix}, \tag{1}$$

where all x_i are distinct numbers.

Note that the requirement that all x_i are distinct is usually not included in the definition. Here it has been added since two x_i being equal would indicate two overlapping nodes which can be combined to a single node. If all x_i are distinct the matrix is also guaranteed to be invertible.

Choosing elements

$$x_i = (2i - L - 1)\alpha \ , \ 1 \leq i \leq L \tag{2}$$

will give the general lattice matrix with the final row missing.

The inverse of the Vandermonde matrix is known, see Macon and Spitzbart[5], and can be used to calculate the transition probabilities.

Theorem 1. *The elements of the inverse of an L-dimensional Vandermonde matrix V_L can be calculated by*

$$\left(V_L^{-1}\right)_{ij} = \frac{(-1)^{j-1}\sigma_{L-j,i}}{\prod_{\substack{k=1 \\ k\neq i}}^{L}(x_k - x_i)},$$

where $\sigma_{j,i}$ is the jth elementary symmetric polynomial with variable x_i set to zero:

$$\sigma_{j,i} = \sum_{1\leq m_1<m_2<\dots \ <m_j\leq L} \prod_{n=1}^{j} x_{m_n}(1-\delta_{m_n,i}) \ , \ \delta_{a,b} = \begin{cases} 1 \ , \ a = b, \\ 0 \ , \ a \neq b. \end{cases}$$

For x_i of the form (2) the expression for the elements of the inverse matrix can be simplified.

Theorem 2. *For a Vandermonde matrix, V_L, with elements defined by* (2)*, the elements of the inverse are given by*

$$\left(V_L^{-1}\right)_{ij} = \frac{(-1)^{j-i}}{2^{L-2}\alpha^{j-1}} \cdot \frac{\tilde{\sigma}_{L-j,i}}{(i-1)!(L-i)!} \tag{3}$$

where

$$\tilde{\sigma}_{j,i} = \sum_{1\leq m_1<\dots \ <m_j\leq L} \prod_{n=1}^{j} (2n - L - 1)(1-\delta_{m_n,i}) \ \ , \ \delta_{a,b} = \begin{cases} 1 \ , \ a = b, \\ 0 \ , \ a \neq b. \end{cases} \tag{4}$$

For more details, see Lundengård *et al.*[6].

Matching the lattice to the first $L-1$ moments gives the equation

$$\boldsymbol{p} = V_L^{-1}\boldsymbol{\mu}, \tag{5}$$

where $\boldsymbol{p}$ and $\boldsymbol{\mu}$ are vectors containing the probabilities and moments respectively. Using formulas (1) and (5) gives

$$p_i = \sum_{j=1}^{L} \left(V^{-1}\right)_{ij} \mu_{j-1} = \sum_{j=1}^{L} \frac{(-1)^{j-i}}{2^{L-1}\alpha^{j-1}} \cdot \frac{\tilde{\sigma}_{L-j,i}}{(i-1)!(L-i)!}\mu_{j-1}\,. \tag{6}$$

The lattice models also require that the L:th moment is matched:

$$\sum_{i=1}^{L} p_i x_i^L = \mu_L.$$

Using equation (6) this requirement can be rewritten as a polynomial equation $P_L(\alpha) = 0$ where

$$P_L(\alpha) = -\mu_L + \sum_{j=1}^{L} \alpha^{L-j+1}\mu_{j-1}\left(\sum_{i=1}^{L} \frac{(-1)^{j-i}(2i-L-1)^L\tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!}\right) \tag{7}$$

and α is a real, positive number. For further details, see Lundengård *et al.*[6]. Next we will show that expression (7) can be simplified further.

Lemma 1. *Let A be a set of n distinct real values evenly distributed around zero. Denote the set of combinations of k elements from A with A_k. Let $\pi : A_k \mapsto \mathbb{R}$ be the product of all elements in a given combination.*
If k is odd

$$\sum_{s\in A_k} \pi_k(s) = 0 \tag{8}$$

and if k is even

$$\sum_{s\in A_k} \pi_k(s) = \sum_{s\in \tilde{A}_{\frac{k}{2}}} \tilde{\pi}_{\frac{k}{2}}(s) \tag{9}$$

where $\tilde{A} = \{a \in A | a > 0\}$, $\tilde{A}_k$ is the set of combinations of k elements from $\tilde{A}$ and $\tilde{\pi} : \tilde{A}_k \mapsto \mathbb{R}$ is the product of the square of the elements in a combination multiplied by $(-1)^k$.

Proof. When k is odd it is possible for all $s \in A_k$ to rewrite the product $\pi_k(s) = a_l\pi_{k-1}(r)$ such that a_l is not a factor in $\pi_{k-1}(r)$ for some $a_l \in A$, $r \in A_{k-1}$. If $a_l = 0$ it is obvious that $\pi_k(s) = 0$ and for any other $a_l \in A$ there is another combination $t \neq s$ such that $\pi_k(t) = -a_l\pi_{k-1}(r) = -\pi_k(s)$ and thus

$$\sum_{s\in A_k} \pi_k(s) = 0 + \sum_{r\in A_{k-1}} (a_r - a_r)\pi_k(s) = 0.$$

When k is even we can use an argument analogous to the odd case and conclude that for each combination $s \in A_k$ that can be rewritten such that $\pi_k(s) = a_l \pi_{k-1}(r)$ where a_l is not a factor in $\pi_{k-1}(r)$ for some $a_l \in A$, $r \in A_{k-1}$ there is also a combination that generate an annihilating term in the sum over all the products. Thus the only remaining terms in the sum over the products will contain both a_l and $-a_l$ as factors and thus any term can be written on the form

$$\pi_k(s) = \prod_{a \in s} a \cdot (-a) = (-1)^{\frac{k}{2}} \prod_{a \in s} a^2.$$

Lemma 2. *The integer values given by* (4) *can be simplified in the following way:*

$$\begin{aligned} j = 2k+1 \quad &: \quad \tilde{\sigma}_{j,i} = x_{L-i+1} \sum_{\substack{s \in \tilde{A}_k \\ x_i \notin s}} \tilde{\pi}_k(s), \\ j = 2k \quad &: \quad \tilde{\sigma}_{j,i} = \sum_{\substack{s \in \tilde{A}_k \\ x_i \notin s}} \tilde{\pi}_k(s). \end{aligned}$$

From this it is also clear that $\tilde{\sigma}_{j,i} = (-1)^j \tilde{\sigma}_{j,L-i+1}$.

Proof. With the notation used in Lemma 1 the expression in (4) can be rewritten as a sum of products of combinations of the elements in $\boldsymbol{x}$,

$$\begin{aligned} \tilde{\sigma}_{j,i} &= \sum_{1 \le m_1 < \ldots < m_j \le L} \prod_{n=1}^{j} (2n - L - 1)(1 - \delta_{m_n,i}) = \sum_{\substack{s \in A_j \\ i \notin s}} \pi_j(s) \\ &= \sum_{\substack{s \in A_{j-1} \\ x_i \notin s \\ -x_i \notin s}} -x_i \pi_{j-1}(s) + \sum_{\substack{s \in A_j \\ x_i \notin s \\ -x_i \notin s}} \pi_j(s), \end{aligned}$$

where A is the set formed by the values of the elements in $\boldsymbol{x}$. Now Lemma 2 follows by directly applying Lemma 1.

Lemma 3. *Let*

$$c(j) = \sum_{i=1}^{L} \frac{(-1)^{L-j}(2i - L - 1)^L \tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!}. \tag{10}$$

Then $c(j) = 0$ *if* $L - j$ *is even and if* $L - j$ *is odd*

$$c(j) = \sum_{i=1}^{\lfloor \frac{L}{2} \rfloor} \frac{(-1)^{j-i}(2i - L - 1)^L \tilde{\sigma}_{L-j,i}}{2^{L-2}(i-1)!(L-i)!}.$$

Proof. Split the sum into two parts:

$$c(j) = \sum_{i=1}^{\lfloor \frac{L}{2} \rfloor} \frac{(-1)^{j-i}(2i-L-1)^L \tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!} + \sum_{i=\lfloor \frac{L}{2} \rfloor + 1}^{L} \frac{(-1)^{j-i}(2i-L-1)^L \tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!}.$$

Changing index in the second sum according to $k = L - i + 1$ gives:

$$c(j) = \sum_{i=1}^{\lfloor \frac{L}{2} \rfloor} \frac{(-1)^{j-i}(2i-L-1)^L \tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!} + \sum_{k=1}^{\lfloor \frac{L}{2} \rfloor + a} \frac{(-1)^{j-k-L+1}(2k-L+1)^L \tilde{\sigma}_{L-j,L-k+1}}{2^{L-1}(n-k)!(k-1)!}.$$

where $a = 1$ when L is odd and $a = 0$ when L is even.

Lemma 2 gives $\tilde{\sigma}_{L-j,i} = (-1)^{L-j}\tilde{\sigma}_{L-j,L-i+1}$ and thus recombining the two sums gives:

$$c(j) = \left(1 - (-1)^{L-j}\right) \sum_{i=1}^{\lfloor \frac{L}{2} \rfloor} \frac{(2i-L-1)^L \tilde{\sigma}_{L-j,i}}{2^{L-1}(i-1)!(L-i)!}.$$

Since the factor in front of the sum is zero when j is odd and two otherwise this concludes the proof.

Lemma 4. *The polynomial given by* (7) *can be written on the form*

$$P(\alpha) = \begin{cases} \mu_L - \sum_{j=1}^{\lfloor \frac{L}{2} \rfloor} c(2j-1)\ \mu_{L-2j}\ \alpha^{L-2j} & \text{if } L \text{ even.} \\ \mu_L - \sum_{j=1}^{\lfloor \frac{L}{2} \rfloor} c(2j)\ \mu_{L-2j+1}\ \alpha^{L-2j+1} & \text{if } L \text{ odd.} \end{cases} \tag{11}$$

where $c(j)$ *is defined by* (10).

Proof. This lemma follows from substituting the $c(j)$ in Lemma 3 in the polynomial defined by (7).

3 On the existence of suitable jump sizes

There are conditions that must be satisfied for the moment matching lattice methods to work. The distance between the lattice nodes, α, must be a positive real root to the polynomial given by (11). For this α the probabilities to move to node i, given by (6), must be between zero and one.

To examine whether there are any positive real roots for (11) Sturm's theorem will be used.

Theorem 3 (Sturm's theorem).
Let $p_0(x)$ be a polynomial and $p_1(x) = p_0'(x)$. Using polynomial division we can find $p_2(x)$, ..., $p_n(x)$ such that

$$p_{k-2}(x) = q_{k-1}(x)p_{k-1}(x) - p_k(x),$$
$$p_{n-1}(x) = q_n(x)p_n(x).$$

The sequence $S_L(x) = \{p_0(x),\ p_1(x), \ldots,\ p_n(x)\}$ is called the canonical Sturm chain. The number of real roots of $p_0(x)$, m, confined between a and b, $a < b$, $p(a) \neq 0$, $p(b) \neq 0$, is given by $m = v_L(a) - v_L(b)$ where $v_L(x)$ is the number of sign variations in $S(x)$ ignoring zeros.

There are many sources for proofs of Sturm's theorem, e.g. Prasolov[7].

For the quadrinomial ($L = 4$) and pentanomial ($L = 5$) lattices the following canonical Sturm chains correspond to the polynomial given by (11):

$$S_4(\alpha) = \Big\{ -9\alpha^4 + 10\mu_2\alpha^2 - \mu_4,\ \ -36\alpha^3 + 20\mu_2\alpha,$$
$$\left(\frac{36}{5}\frac{\mu_4}{\mu_2} - 20\mu_2\right)\alpha,\ \ -\mu_4\Big\},$$
$$S_5(\alpha) = \Big\{ -64\mu_1\alpha^4 + 20\mu_3\alpha^2 - \mu_5,\ \ -256\mu_1\alpha^3 + 40\mu_3\alpha,$$
$$-10\mu_3\alpha^2 + \mu_5,\ \ \left(\frac{256}{10}\frac{\mu_1}{\mu_3}\mu_5 - 40\mu_3\right)\alpha,\ \mu_5\Big\}.$$

Finding all the positive real roots can now be done by noting that there for any polynomial must be some value, $r > 0$, that is large enough that the highest order term in the polynomial dominates and the signs in the Sturm chain will be determined by the corresponding coefficients. Thus the number of real positive roots can be found by calculating $q = v_L(0) - v_L(r)$.

Since all even-numbered moments are positive it is easy the see that $v_4(0) = 0$ and $v_4(r) = 0$ unless $36\mu_4 > 100\mu_2^2$ which will give $v(r) = 2$. In this case we have an underlying asset described by a Merton-Bates jump-diffusion process and if we denote the Lévy measure for the process with

$$l(\mathrm{d}x) = \frac{\lambda}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(\mathrm{d}x - \eta)^2}{2\delta^2}\right)$$

the second and fourth moments will be

$$\mu_2 = \frac{\lambda\delta^2(1+\eta^2)}{\sigma^2 + \lambda\delta^2 + \lambda\eta^2},$$
$$\mu_4 = \frac{\lambda\delta^4(3+6\eta^2+\eta^4)}{(\sigma^2 + \lambda\delta^2 + \lambda\eta^2)^2},$$

for derivations of these expressions see Lundengård *et al.*[6].

Using the explicit formulas for the moments it can be shown that the condition $36\mu_4 > 100\mu_2^2$ is equivalent to $\lambda < \frac{36}{100}\frac{3+6\eta^2+\eta^4}{(1+\eta^2)^2}$.

The pentanomial case is less straight forward, $v_5(0) = 2$ and $v_5(r)$ can vary between zero and two depending on the odd-numbered moments μ_1, μ_3 and μ_5 in a complicated way which we choose not to describe in detail here.

Note that the existence of real positive roots does not guarantee that the probabilities given by (11) will be between zero and one.

Acknowledgements This work was partially supported by The Royal Physiographic Society in Lund, The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Swedish Research Council, The Royal Swedish Academy of Sciences, The Crafoord Foundation, as well as by The International Science Program, SIDA foundation and by Mälardalen University. Carolyne Ogutu is grateful to the research environment in Mathematics and Applied Mathematics at the Division of Applied Mathematics of the School of Education, Culture and Communication (UKK) at Mälardalen University for their hospitality and creating excellent conditions for research, research education and cooperation. She is also grateful for the support of the International Science Program, Uppsala University, Sweden, through collaboration with The Eastern African Universities Mathematics Programme.

References

1. Amin, K. L. "Jump diffusion option valuation in discrete time", *The Journal of Finance* **48**(5), 1833–1863 (1993)
2. Cox, J. C., Ross, S. A., and Rubinstein, M., "Option pricing. A simplified approach", *Journal of Financial Economics* **7**(3), 229–263 (1979)
3. Dai, T.-S., Wang, C.-J., Lyuu, Y.-D., and Liu, Y.-C., "An efficient and accurate lattice for pricing derivatives under jump-diffusion process", *Applied Mathematics and Computation* **217**, 3174–3189 (2010)
4. Hilliard, J. E., and Schwartz, A., "Pricing European and American derivatives under a jump-diffusion process: A bivariate tree approach", *Journal of Financial and Quantitative Analysis* **40**(3), 671–691 (2005)
5. Macon, N. and Spitzbart, A., "Inverses of Vandermonde matrices", *The American Mathematical Monthly* **65**(2), 95–100 (1958)
6. Lundengård, K., Ogutu, C., Silvestrov S., and Weke, P., "Asian Options, Jump-Diffusion Processes on a Lattice, and Vandermonde Matrices", in *Modern Problems in Insurance Mathematics*, Silvestrov, Dmitrii, Martin-Löf, Anders, Eds., Springer-Verlag, Berlin, 337–364 (2014)
7. Prasolov V. V., *Polynomials*, *Algorithms and Computation in Mathematics* **11**, Springer-Verlag, Berlin (2004)
8. Primbs, J. A., Rathinam, M., and Yamada, Y., "Option pricing with a pentanomial lattice model that incorporates skewness and kurtosis", *Applied Mathematical Finance* **14**(1), 1–17, February (2007)