Statistics and Operations Research Transactions, vol. 40, n. 2 (2016)

SORT 40 (2) July-December 2016, 227-242

Improving the resolution of the simple assembly line balancing problem type E

Albert Corominas, Alberto García-Villoria* and Rafael Pastor**

Abstract

The simple assembly line balancing problem type E (abbreviated as SALBP-E) occurs when the number of workstations and the cycle time are variables and the objective is to maximise the line efficiency. In contrast with other types of SALBPs, SALBP-E has received little attention in the literature. In order to solve optimally SALBP-E, we propose a mixed integer liner programming model and an iterative procedure. Since SALBP-E is NP-hard, we also propose heuristics derived from the aforementioned procedures for solving larger instances. An extensive experimentation is carried out and its results show the improvement of the SALBP-E resolution.

MSC: 90C27 (Combinatorial Optimisation).

Keywords: Assembly line balancing, SALBP, manufacturing optimisation.

1. Introduction

Assembly line balancing problems (ALBPs) consist in assigning optimally (according to a given objective function) the tasks of an assembly or production process to the ordered workstations of an assembly line (or several assembly lines) such that some specific conditions are satisfied. These NP-hard problems (Gutjahr and Nemhauser, 1964) have an important relevance in many production systems such as in automotive and electronic industries (Battaïa and Dolgui, 2013). Thus, ALBPs have been extensively studied in the literature and several surveys have been published. Some recent surveys are Erel and Sarin (1998), Rekiek et al. (2002), Becker and Scholl (2006), Scholl and Becker (2006a,b), Boysen et al. (2008), Battaïa and Dolgui (2013).

* *Corresponding author:* Alberto García-Villoria, Institute of Industrial and Control Engineering (IOC), Av. Diagonal 647 (Edif. ETSEIB), 11th floor, 08028 Barcelona, Spain; e-mail: alberto.garcia-villoria@upc.edu

** Institute of Industrial and Control Engineering (IOC), Universitat Politècnica de Catalunya (UPC). {albert.corominas/ alberto.garcia-villoria/rafael.pastor}@upc.ed

Received: January 2015

Accepted: July 2016

The most basic family of ALBPs is the simple assembly line balancing problem (SALBP). SALBP is defined with the following assumptions (Baybars, 1986): 1) a task cannot be split among workstations; 2) there are precedence relations between tasks; 3) all tasks must be processed; 4) the task process times are independent of the workstation, they are known with certainty, they are not sequence dependent and they are additive at any workstation; 5) all workstations have the same associated costs; 6) any task can be processed at any workstation; 7) the line is serial and without feeder or parallel subassembly lines; 8) the line is designed for a unique model of a single product.

According to Baybars' nomenclature, when the objective is to minimise the number m of workstations for a given upper bound on the cycle time ct, the SALBP is called SALBP-1; if the objective is to minimise ct for a given m, the problem is called SALBP-2. On the other hand, SALBP-E is a more general version of SALBP which consists in finding the combination of m and ct such that the line efficiency is maximised. The efficiency is measured as the sum of task process times, t_{sum}, divided by the product $m \cdot ct$. In practice SALBP-E has a lower bound on m (due to a desired degree of the division of the labour) and/or an upper bound on ct (due to a minimum desired production rate); otherwise, SALBP-E would be trivial since a line with $m = 1$ and $ct = t_{sum}$ has an efficiency equal to 1. When the aim is to find a feasible line balance for a given combination of m and ct, the problem is called SALBP-F.

SALBP-1 is the most studied problem in the ALBP literature and a lot of exact and heuristic procedures have been designed for its resolution (see Scholl and Becker, 2006b). SALBP-2 has been also studied although there exist fewer procedures and most of them are based on repeatedly solving SALBP-1 with different values (see Scholl and Becker, 2006b, Uğurdağ et al., 1997). SALBP-F can be solved with modified SALBP-1 or SALBP-2 procedures (Scholl and Becker, 2006b).

In the last years researchers have intensified their efforts studying ALBPs with additional characteristics. For instance, among others: general assignment constraints (e.g. Scholl et al., 2010), task times depending on the sequence (e.g. Capacho et al., 2009), setup times between tasks (e.g. Martino and Pastor, 2010), uncertainty on task times (e.g. Saif et al., 2014), task times dependent on the workers (e.g. Moreira et al., 2015), space constraints (e.g. Chica et al., 2016), constrained resources (e.g. Corominas et al., 2011), lengths of the workpieces larger than the accessibility windows of the workstations (e.g. Calleja et al., 2014), ergonomics considerations (e.g. Bautista et al., 2016), mixed-model lines (e.g. Battaïa et al., 2015), robotic lines (Levitin et al., 2006, Gao et al., 2009, Yoosefelahi et al., 2012), U-shaped lines (e.g. Ogan and Azizoglu, 2015), machining transfer lines (e.g. Battaïa and Dolgui, 2012) and parallel two-sided assembly lines (e.g. Tapkan et al., 2016).

In contrast, to the best of our knowledge, very few procedures have been discussed for SALBP-E. Plans and Corominas (1999) formulated a MILP model which solves optimally SALBP-E. The model is tested only on seven instances and, thus, the computational experiment is insufficient. Scholl and Becker (2006b) outlined the following exact approach. All combinations of m and ct values are (implicitly) considered and

SALBP-F is examined (that is, whether there is a feasible solution with m workstations and a cycle time equal to ct). The feasible combination with best efficiency would be the optimal SALBP-E solution. The drawback is that a lot of combinations may be considered and SALBP-F is itself NP-hard. Wei and Chao (2011) designed an exact procedure based on solving optimally as many SALBP-2 as admissible values for m; each SALBP-2 is solved by means of mixed integer linear programming (MILP). That work contains some errors which are corrected in García-Villoria and Pastor (2013). Additionally, the computational experiment in Wei and Chao (2011) is limited to small instances that can be solved optimally in a short computing time. An ALBP type-E considering stochastic task times has been dealt in Gurevsky et al. (2012) and Zacharia and Nearchou (2013), in which two heuristic procedures and a genetic algorithm are developed, respectively. A genetic algorithm is also used in Al-Hawari et al. (2015) to solve a multi-objective ALBP, which includes the maximization of the line efficiency. Esmaeilbeigi et al. (2015) proposed mathematical programming for SALBP-E and different variants of a model and redundant constraints are compared.

The aim of this study is to improve the resolution of SALBP-E. We propose a MILP model together with an enhanced procedure based on the iterative one designed by Wei and Chao (2011). A computational experiment shows that our proposal outperforms the previously published methods.

The remaining paper is organised as follows. First, Section 2 presents the terminology, bounds on the objective function and the cycle time, and the MILP model. The enhanced iterative procedure and its derived heuristics are presented in Section 3. In Section 4 the proposed procedures are tested on a well-known benchmark set of instances; the procedures are compared and the results show that the SALBP-E resolution is improved. We conclude with the final conclusions in Section 5.

2. Terminology, lower and upper bounds and MILP

The data that define a SALBP-E instance are the following:

- n Number of tasks
- t_i Process time for task i $(i = 1, \ldots, n)$
- IP Set of immediate precedence relations, such that $(h,i) \in IP$ means that task h must be performed before task i
- m_{min}, m_{max} The minimum and maximum number of workstations allowed, respectively

The SALBP-E objective is to maximise the efficiency E of the line (recall that $E = t_{sum}/(m \cdot ct)$, where $t_{sum} = \sum_{i=1\ldots n} t_i$). Note that this objective is equivalent to minimise the line capacity $Z = m \cdot ct$. We propose the following bounds on the cycle time and the line capacity, where $\lceil x \rceil$ ($\lfloor x \rfloor$) is the operator that returns the smallest (greatest) integer that is equal to or greater (smaller) than x:

LB^{ct} Lower bound of cycle time: $LB^{ct} = LB^{ct}_{m_{max}}$, where $LB^{ct}_m = \max(\max_{i=1\ldots n} t_i, \lceil t_{sum}/m \rceil)$
UB^{ct} Upper bound of cycle time: $UB^{ct} = UB^{ct}_{m_{min}}$, where $UB^{ct}_m = \max(\max_{i=1\ldots n} t_i, 2 \cdot \lfloor t_{sum}/m \rfloor)$
LB^{Z} Lower bound on the line capacity: $LB^{Z} = \min_{m=m_{min}\ldots m_{max}}(m \cdot LB^{ct}_m)$
UB^{Z} Upper bound on the line capacity: $UB^{Z} = \min_{m=m_{min}\ldots m_{max}}(m \cdot UB^{ct}_m)$

LB^{ct}_m and UB^{ct}_m are a lower and upper bound, respectively, on the cycle time for a given number of workstations m; that is, they are bounds of SALBP-2 (Scholl, 1999). Thus, $LB^{ct}_{m_{max}}$ and $UB^{ct}_{m_{min}}$ are lower and upper bounds on the cycle time of SALBP-E, respectively. With respect to the bounds on the line capacity, they are straightforwardly deduced from the bounds on the efficiency formulated in Scholl (1999).

Moreover, we define the following additional data that are derived from the above data and bounds:

P_i Set of all tasks which must precede task i $(i = 1\ldots n)$:
$P_i = \bigcup_{h=1\ldots n|(h,i)\in IP}(\{h\} \cup P_h)$
S_i Set of all tasks which must succeed task i $(i = 1\ldots n)$:
$S_i = \bigcup_{h=1\ldots n|(i,h)\in IP}(\{h\} \cup S_h)$
E_i Earliest workstation to which task i can be assigned $(i = 1\ldots n)$:
$E_i = \lceil (t_i + \sum_{h\in P_i} t_h)/UB^{ct} \rceil$
L_i Latest workstation to which task i can be assigned $(i = 1\ldots n)$:
$L_i = \max_{m=m_{min}\ldots m_{max}}(m + 1 - \lceil (t_i + \sum_{h\in S_i} t_h)/UB^{ct}_m \rceil)$
W_j Set of tasks that can be assigned to workstation j $(j = 1\ldots m_{max})$:
$W_j = \{i = 1\ldots n : E_i \leq j \leq L_i\}$

Formulations of SALBP-1 (also of SALBP-2) usually define the earliest and latest workstation in which each task can be assigned based on the precedence relations (e.g. Saltzman and Baybars, 1987). Analogously for SALBP-E, we define E_i and L_i as the earliest and latest workstation in which task i can be assigned, respectively. They are used to reduce significantly the size of the MILP model of the problem.

The mathematical model is formulated as follows:

Variables

z Line capacity: $LB^{Z} \leq z \leq UB^{Z}$
ct Cycle time: $LB^{ct} \leq ct \leq UB^{ct}$
$x_{ij} \in \{0,1\}$ 1 if task i is assigned to station j; 0 otherwise $(i = 1\ldots n; j = E_i\ldots L_i)$
$y_j \in \{0,1\}$ 1 if station j exists; 0 otherwise $(j = m_{min}+1\ldots m_{max})$

Model

$$[\text{MIN}]\ z \tag{1}$$

$$\sum_{j=E_i}^{L_i} x_{ij} = 1 \qquad i = 1 \ldots n \tag{2}$$

$$\sum_{i \in W_j} t_i \cdot x_{ij} \leq ct \qquad j = 1 \ldots m_{max} \tag{3}$$

$$\sum_{j=E_h}^{L_h} j \cdot x_{hj} \leq \sum_{j=E_i}^{L_i} j \cdot x_{ij} \qquad \forall (h,i) \in IP \tag{4}$$

$$\sum_{i \in W_j} x_{ij} \leq \|W_j\| \cdot y_j \qquad j = m_{min}+1 \ldots m_{max} \tag{5}$$

$$z \geq m_{min} \cdot ct \tag{6}$$

$$z \geq j \cdot ct - M_j \cdot (1 - y_j) \qquad j = m_{min}+1 \ldots m_{max}$$

$$\text{where } M_j = j \cdot UB^{ct} - \min_{m=m_{min}}^{j-1} (m \cdot LB_m^{ct}) \tag{7}$$

$$y_j \geq y_k \qquad j = m_{min}+1 \ldots m_{max}-1; k = j+1 \ldots m_{max} \tag{8}$$

Objective function (1) minimises the line capacity (recall that it is equivalent to maximise the line efficiency). Constraints (2) ensure that each task is assigned to one and only one workstation. Constraints (3) imply that the cycle time is not lower than the total task process time assigned to any workstation. Constraints (4) impose the precedence relations. Constraints (5) force a workstation to be open when some task is assigned to it. Constraints (6) and (7) link the line capacity with the number of open workstations and the cycle time. Finally, constraints (8) impose that the open workstations must be contiguous (as it is done in Pastor et al., 2011) and break symmetries.

The main differences between our enhanced MILP model and the model proposed in Plans and Corominas (1999) (let they be named Enh-MILP and CP-MILP, respectively) are: (i) the addition of a lower bound on the cycle time and a lower and upper bound on the line capacity; and (ii) the constraints that impose the contiguousness of the open workstations.

3. Iterative procedures

First we explain in Section 3.1 the procedure proposed in Wei and Chao (2011) and corrected in García-Villoria and Pastor (2013); let it be named Iterative Procedure (IP). Then we propose our enhanced iterative procedure in Section 3.2; let it be named Enhanced Iterative Procedure (EIP). When non-small instances are solved, the above procedures may need a huge computational time to solve optimally them, so a maximum

global time has to be set. Section 3.3 discuses several heuristics based on IP and EIP in which the maximum time is shared among their iterations in different ways.

3.1. Iterative Procedure (IP)

The Wei and Chao's iterative procedure (IP) consists in solving the corresponding SALBP-2 for each value of m between m_{min} and m_{max}. Figure 1 shows its pseudocode.

Let $CT(Sol)$ be the cycle time of solution Sol

$Z^* = \infty$

For $m = m_{min} \dots m_{max}$ **do**:

  $Sol =$ Solve SALBP-2 with m workstations

  If $m \cdot CT(Sol) \leq Z^*$ **then** $Sol^* = Sol$, $Z^* = m \cdot CT(Sol)$ **End if**

End for

Return Sol^*

Figure 1: *Pseudocode of* Iterative Procedure *for SALBP-E.*

The following MILP model is used to solve SALBP-2:

Additional data

m Number of workstations

E'_i Earliest workstation to which task i $(i = 1 \dots n)$ can be assigned:
$E'_i = \lceil (t_i + \sum_{h \in P_i} t_h) / UB^{ct}_m \rceil$

L'_i Latest workstation to which task i $(i = 1 \dots n)$ can be assigned:
$L'_i = m + 1 - \lceil (t_i + \sum_{h \in S_i} t_h) / UB^{ct}_m \rceil$

W'_j Set of tasks that can be assigned to workstation j $(j = 1 \dots m)$:
$W'_j = \{i = 1 \dots n : E'_i \leq j \leq L'_i\}$

Variables

ct Cycle time: $LB^{ct}_m \leq ct \leq UB^{ct}_m$

$x_{ij} \in \{0,1\}$ 1 if task i is assigned to station j; 0 otherwise $(i = 1 \dots n; j = E'_i \dots L'_i)$

Model

$$[\text{MIN}]\ ct \tag{9}$$

$$\sum_{j=E'_i}^{L'_i} x_{ij} = 1 \qquad i = 1 \dots n \tag{10}$$

$$\sum_{i \in W'_j} t_i \cdot x_{ij} \leq ct \qquad j = 1 \ldots m \tag{11}$$

$$\sum_{j=E'_h}^{L'_h} j \cdot x_{hj} \leq \sum_{j=E'_i}^{L'_i} j \cdot x_{ij} \qquad \forall (h,i) \in IP \tag{12}$$

3.2. Enhanced Iterative Procedure (EIP)

One drawback of IP is that each resolution of SALBP-2 does not use any information of the previous SALBP-2 solutions. To improve the performance of IP we propose an enhanced iterative procedure (EIP), which takes advantage of the best solution known up to the current iteration.

Let $ni = m_{max} - m_{min} + 1$ be the number of iterations of the IP and let Z_p^* be the best line capacity value found at the beginning of iteration p ($p = 1 \ldots ni$); we consider $Z_1^* = \infty$. The SALBP-2 to solve at iteration p has $m = p + m_{min} - 1$ workstations; let Sol_m be its optimum solution. The cycle time of Sol_m, $CT(Sol_m)$, must fulfil the following condition in order to have a better line capacity than the best SALBP-E solution known at the moment:

$$m \cdot CT(Sol_m) < Z_p^* \equiv CT(Sol_m) \leq \lfloor (Z_p^* - 1)/m \rfloor \tag{13}$$

Eq. 13 assumes, without loss of generality, that the process task times are integers, restricting cycle times and line capacities to integer values.

EIP is an adaptation of IP in order to reduce the search space of each SALBP-2 thanks to the condition expressed in Eq. 13. Thus, each iteration of EIP may be more efficient. The EIP pseudocode is very similar to the one shown in Figure 1. The differences are at each iteration p the Z_p^* value is available and the domain of variable ct of the MILP model to solve SALBP-2 may be tighter. The ct domain is expressed in Eq. 14:

$$LB_m^{ct} \leq ct \leq \min(UB_m^{ct}, \lfloor (Z_p^* - 1)/m \rfloor) \tag{14}$$

Note that the SALBP-2 model used by EIP may be infeasible. Its infeasibility at iteration p means that the optimum solution of SALBP-2 with $m = p + m_{min} - 1$ workstations is not better than the best SALBP-E solution found up to iteration $p - 1$. Thus, EIP continues the search at the next iteration.

3.3. Heuristics derived from IP and EIP

Wei and Chao (2011) assumed that, at each iteration of IP, its corresponding SALBP-2 would be solved optimally. However, for non-small instances, the required time may be huge in practice. Heuristics can be derived from IP and EIP limiting the maximum total

computing time to D and returning the best solution found. In that case, one question that arises is how to distribute the available time among the ni iterations. We propose three ways similarly as it is done in the ALBP literature (see, for example,Pastor 2011):

T1 The maximum computing time for solving SALBP-2 at the first iteration is D, at the second iteration is the remaining time (if any), and so on. That is, the time limit at iteration p, TL_p, is the following:
$TL_p = D - \sum_{q=1\ldots p-1} \tau_q$ for $p = 1 \ldots ni$, where τ_q is the time spent for solving SALBP-2 at iteration q.

T2 The maximum computing time for solving SALBP-2 is half of the remaining time (except for the last SALBP-2, which is all remaining time). That is:
$TL_p = (1/2) \cdot (D - \sum_{q=1\ldots p-1} \tau_q)$ for $p = 1 \ldots ni-1$, and $TL_{ni} = D - \sum_{q=1\ldots ni-1} \tau_q$.

T3 The maximum computing time for solving SALBP-2 is the remaining time divided by the number of the remaining iterations. That is:
$TL_p = (D - \sum_{q=1\ldots p-1} \tau_q)/(ni - p + 1)$ for $p = 1 \ldots ni$.

Combining the two iterative procedures (IP and EIP) and the three ways of splitting the available computing time (T1, T2 and T3) results in a total of six heuristics: IP-T1, IP-T2, IP-T3, EIP-T1, EIP-T2 and EIP-T3. Note that two heuristics can also be derived from the introduced mathematical models (Section 2) limiting their maximum computing time to D.

Illustrative example of the heuristics mechanism

In order to clarify how the proposed heuristics work, we will show as example the iterations of EIP-T1 when it is applied to solve the instance named Lutz3 with $m_{min} = 12$ and $m_{max} = 15$ with a total computing time $D = 3600$ s. Thus, EIP-T1 will iterate 4 times ($ni = m_{max} - m_{min} + 1 = 4$).

EIP-T1

Iteration $p = 1$. $Z_1^* = \infty$. $TL_1 = D = 3600$. SALBP-2 is solved with $m = 12$ workstations and a solution is found; let it be called Sol^{12}, with $ct(Sol^{12}) = 138$. Let $Sol^* = Sol^{12}$ (whose line capacity is equal to $12 \cdot 138 = 1656$). The time spent in this iteration is $\tau_1 = 14$.

Iteration $p = 2$. $Z_2^* = 1656$. $TL_2 = D - \tau_1 = 3586$. SALBP-2 is solved with $m = 13$ workstations and the model is infeasible (therefore, there is no solution with 13 workstations and a line capacity smaller than Z_2^*, recall Eq. 14). The time spent in this iteration is $\tau_2 = 1$.

Iteration $p = 3$. $Z_3^* = 1656$. $TL_3 = D - \tau_1 - \tau_2 = 3585$. SALBP-2 is solved with $m = 14$ workstations and a solution is found; let it be called Sol^{14}, with $ct(Sol^{14}) = 118$.

Let $Sol^* = Sol^{14}$ (whose line capacity is equal to $14 \cdot 118 = 1652$). The time spent in this iteration is $\tau_3 = 5$.

Iteration $p = 4$. $Z_4^* = 1652$. $TL_4 = D - \tau_1 - \tau_2 - \tau_3 = 3580$. SALBP-2 is solved with $m = 15$ workstations and a solution is found; let it be called Sol^{15}, with $ct(Sol^{15}) = 110$. Let $Sol^* = Sol^{15}$ (whose line capacity is equal to $54 \cdot 110 = 1650$).

Return solution Sol^*

EIP-T2 and EIP-T3 would iterate in a similar way but the time limits at each iteration would be the following. For EIP-T2: $TL_1 = D/2 = 1800$, $TL_2 = (D - \tau_1)/2 = 1793$, $TL_3 = (D - \tau_1 - \tau_2)/2 = 1792$ and $TL_4 = D - \tau_1 - \tau_2 - \tau_3 = 3580$. And for EIP-T3: $TL_1 = D/4 = 900$, $TL_2 = (D - \tau_1)/3 = 1195$, $TL_3 = (D - \tau_1 - \tau_2)/2 = 1792$ and $TL_4 = D - \tau_1 - \tau_2 - \tau_3 = 3580$. Note that the results at each iteration would not be different since $\tau_p < TL_p$ for $p = 1 \ldots 4$.

4. Computational experiments

The MILP models are solved using the IBM ILOG CPLEX 12.2 Optimiser; the absolute MIP gap tolerance is set to 0.9999 (since process task times are integer values). The iterative procedures are implemented in Java SE 1.6.21. The experiments are run on a PC 3.16 GHz Pentium Intel Core 2 Duo E8500 with 3.46 GB of RAM. The maximum computing time D per instance and procedure is limited to 3,600 seconds. Note that in a real application, when the design of the assembly line is a strategic problem, the computational time could be much greater; however, one hour seems a reasonable compromise in order to use a variety of instances and, at the same time, make the computational experiment affordable.

Section 4.1 presents the test instances used in the experiments. Section 4.2 shows the obtained results and Section 4.3 compares the heuristic procedures between them. Section 4.4 studies the quality of the proposed heuristics according to the characteristics of the instances. Finally, Section 4.5 analyses how the distribution of the computing time among the iterations affects the performance of the heuristics.

4.1. Description of the test instances

Our experiments are performed on the 256 benchmark SALBP-E instances that are available in Scholl and Klein's assembly line balancing research website (www.assembly-line-balancing.de). Scholl (1993) generated these instances from twenty-four problems varying the m_{min} and m_{max} values. Table 1 shows the problem name, its number of tasks,

Table 1: *Test problems.*

Name	Number of tasks	Process time			Order strength	m_{min} range	m_{max} range
		Minimum	Maximum	Average			
Arcus1	83	233	3691	912.1	59.09	3 to 19	9 to 21
Arcus2	111	10	5689	1354.9	40.38	3 to 22	7 to 27
Barthold	148	3	383	38.1	25.8	3 to 14	5 to 15
Barthol2	148	1	83	28.6	25.8	3 to 51	10 to 52
Bowman	8	3	17	9.4	75	3 to 4	5 to 5
Buxey	29	1	25	11.2	50.74	3 to 10	7 to 13
Gunther	35	1	40	13.8	59.5	3 to 12	8 to 13
Hahn	53	40	1775	264.6	83.82	3 to 7	4 to 8
Heskiaoff	28	1	108	36.6	22.49	3 to 9	10 to 10
Jackson	11	1	7	4.2	58.18	3 to 5	7 to 7
Jaeschke	9	1	6	4.1	83.33	3 to 6	7 to 7
Killbridge	45	3	55	12.3	44.55	3 to 10	11 to 11
Lutz1	32	100	1400	441.9	83.47	3 to 10	11 to 11
Lutz2	89	1	10	5.4	77.55	3 to 46	13 to 49
Lutz3	89	1	74	18.5	77.55	3 to 19	9 to 23
Mansoor	11	2	45	16.8	60	3 to 4	5 to 5
Mertens	7	1	6	4.1	52.38	3 to 4	5 to 5
Mitchell	21	1	13	5	70.95	3 to 7	9 to 9
Roszieg	25	1	13	5	71.67	3 to 9	10 to 10
Sawyer	30	1	25	10.8	44.83	3 to 10	7 to 13
Scholl	297	5	1386	234.5	58.16	3 to 50	4 to 51
Tonge	70	1	156	50.1	59.42	3 to 21	12 to 23
Warnecke	58	7	53	26.7	59.1	3 to 28	13 to 30
Wee-Mag	75	2	27	20	22.67	3 to 36	9 to 38

their minimum, maximum and average process times, the order strength of the precedence graph and the ranges on the m_{min} and m_{max} values used to generate the instances.

4.2. Results of the procedures

When the objective function value of the obtained solution is equal to the lower bound, LB, value (or, strictly speaking, when the difference between them is less than one), then the solution optimality is demonstrated. In the case of the MILP procedures for SALBP-E, LB is equal to the lower bound returned by CPLEX. In the case of the iterative procedures, LB is calculated as follows. Let LBcplex_m^{ct} be the lower bound on the cycle time returned by CPLEX when it solves SALBP-2 with workstations; thus, LB is

equal to $\min_{m=m_{min}...m_{max}}(m \cdot \text{LBcplex}_m^{ct})$. Moreover, we consider that $\text{LBcplex}_m^{ct} = \text{LB}_m^{ct}$ when no lower bound is returned by CPLEX within the maximum time assigned and $\text{LBcplex}_m^{ct} = \infty$ when the SALBP-2 model is infeasible.

Table 2 summarises the type of solutions obtained with each procedure. The following information is given: the number of instances with a proved optimal solution (#Opt); the number of instances with a feasible solution whose optimality is not proven (#Fea); and the number of instances without finding a feasible solution (#Uns).

Table 2: *Type of the obtained solutions.*

	CP-MILP	Enh-MILP	IP-T1	IP-T2	IP-T3	EIP-T1	EIP-T2	EIP-T3
#Opt	72	90	121	124	120	144	143	141
#Fea	175	156	135	132	135	112	113	114
#Uns	9	10	0	0	1	0	0	1

In terms of proved optimal solutions, our proposed MILP model, Enh-MILP, outperforms the model proposed in Plans and Corominas (1999), CP-MILP. Nevertheless, both approaches are clearly worse than any iterative procedure. Moreover, Enh-MILP and CP-MILP cannot find a solution in 10 and 9 instances, respectively, whereas the iterative procedures always find a solution (IP-T1, IP-T2, EIP-T1 and EIP-T2) or only the same 1 instance remains unsolved (IP-T3 and EIP-T3).

4.3. Comparison between the heuristic procedures

We focus on the comparison of the heuristic pairs (IP-T1, EIP-T1), (IP-T2, EIP-T2) and (IP-T3, EIP-T3). Table 3 shows, for each pair, the number of instances in which both procedures guarantee the optimal solution (#Opt) and the average computational time, in seconds, for solving these instances ($\overline{\text{Time}}$). Table 3 also shows, for each pair, the number of instances when none of the procedures guarantees the optimal solution in the computing time allowed (#Fea) and the number of times that a procedure finds a better solution than the another procedure (#Best).

Table 3: *Results when both procedures guarantee the optimal solution or when neither procedure guarantees the optimal solution.*

Time distribution	Optimal solutions			Feasible solutions		
	#Opt	$\overline{\text{Time}}$		#Fea	#Best	
		IP	EIP		IP	EIP
T1	121	1057.1	201.39	112	0	24
T2	124	1189.39	180.38	113	0	42
T3	120	1045.42	191.94	114	0	32

Results in Tables 2 and 3 confirm the effectiveness of the proposed improvements regardless of how the global time is shared among the iterations of the procedures. EIP heuristics are able to find around 15% more proven optimal solutions than the IP heuristics and all instances solved optimally by one IP heuristic are also solved optimally by its analogous EIP heuristic. Regarding the computational time, the average times spent by EIP heuristics when an instance is solved optimally are at least five times less than the IP average times. Furthermore, all feasible solutions reached by an EIP heuristic has the same or better quality than the solutions reached by its analogous IP heuristic.

4.4. Quality of the proposed heuristics according to the instance characteristics

We now focus on the comparison of the three EIP heuristics, which are the best heuristics, taking into account the influence of the characteristics of the instances (as outlined in Table 1) on the quality of the results. Specifically, we create subset of instances according to the number of tasks and order strength. The number of tasks is respectively classified as low, medium and high according to the ranges (7, 35), (45, 111) and (148, 297). Likewise, the order strength is respectively considered low, medium and high according to the ranges (22.49, 25.80), (40.38, 60.00) and (70.95, 83.82).

Table 4: *Number of proven optimal solutions for each combination of number of tasks and order strength.*

Procedure	LL (6)	LM (30)	LH (20)	ML (22)	MM (67)	MH (40)	HL (39)	HM (32)
EIP-T1	6	30	20	14	25	31	15	3
EIP-T2	6	30	20	14	24	32	15	2
EIP-T3	6	30	20	14	22	31	15	3

Table 5: *Quality of the solutions for each combination of number of tasks and order strength.*

Procedure	LL (6)	LM (30)	LH (20)	ML (22)	MM (67)	MH (40)	HL (39)	HM (32)
=	6	30	20	20	32	30	19	2
EIP-T1	0 (0)	0 (0)	0 (0)	0 (2)	5 (9)	0 (1)	3 (4)	15 (2)
EIP-T2	0 (0)	0 (0)	0 (0)	0 (0)	12 (3)	1 (0)	4 (4)	7 (2)
EIP-T3	0 (0)	0 (0)	0 (0)	0 (0)	7 (15)	2 (7)	1 (8)	5 (23)

Tables 4 and 5 report the results obtained by each procedure for combinations of order strength and number of tasks. The column headers use L, M and H for low, medium and high, respectively. The first letter is for the number of tasks and the second letter for the order strength. This is followed by the number of instances in the subset, which

is shown between parentheses. Table 4 shows the number of instances with a proved optimal solution. Table 5 contains the number of times that all procedure obtains the same quality solution ("=") and, for each procedure, the number of times that it obtains a better solution (and a worst solution between parentheses) than the other two solutions obtained with the other procedures; we consider the non-existent solution of the only instance unsolved with EIP-T3 (see Table 2) worse than any feasible solution.

Results show that all EIP procedures are very similar with respect to the number of optimal solutions regardless of the characteristics of the instances. However, with respect to the quality of the solutions, we can see EIP-T3 that is, on average, worse than EIP-T1 and EIP-T2. EIP-T2 tends to be better than EIP-T1 in terms of quantity of best and worst solutions; the exception is in instances with high number of tasks and medium order strength, in which EIP-T1 performs, on average, better.

4.5. Analysis of the proposed heuristics

Lastly, we investigate the different performances of the EIP heuristics. To do so, we record for each instance the number of workstations of the best known solution, m^* (i.e. the best solution obtained with any of the eight procedures). If multiple best solutions have been obtained, then the one with the lowest number of workstations is considered. Table 6 reports the number of times (#Ins) that the difference between m^* and m_{min} is 0, 1, 2, etc.

Table 6: *Differences between the best number of workstations and m_{min}.*

$m^* - m_{min}$	$=0$	$=1$	$=2$	$=3$	$=4$	≥ 5
#Ins	168	44	26	11	3	4

We can observe that most of the number of workstations of the best found solutions are equal to or close to the minimum value allowed. Thus, it seems reasonable that EIP-T3, which tends to share equally the maximum global time among all SALBP-2 resolutions, performs worse than EIP-T1 and EIP-T2, which tend to give priority to SALBP-2 with fewer number of workstations. On the other hand, EIP-T1 might perform slightly worse than EIP-T2 because in some cases the time spent by EIP-T1 in the first SALBP-2 may be too much (potentially all time could be spent in it and no SALBP-2 is solved with other numbers of workstations).

5. Conclusions and future research

SALBP is the type of assembly line balancing problems most studied in the literature. However, most research efforts are reduced to SALBP-1 and SALBP-2. The resolution of SALBP-E, in which the number of workstations and cycle time are variables, has

not received enough attention in the literature with the notable exceptions of Plans and Corominas (1999) and Wei and Chao (2011).

The special interest in our work has been the exact resolution of SALBP-E. We propose an enhanced MILP model, together with an iterative procedure (based on solving SALBP-2 at each iteration) which improves the one proposed by Wei and Chao (2011). Since we expected that large instances cannot be solved optimally in a practical time, we proposed several heuristics based on limiting the maximum computing time and returning the best solution found. Specifically, we propose three ways of distributing the available time among the different SALBP-2 resolutions of the iterative procedures.

Through extensive experimentation, we have been able to determine the benefits of adding the proposed improvements to the existing iterative procedure. On the other hand, we have detected that a direct approach as the proposed MILP model performs worse than any iterative procedure. Nevertheless, other direct approaches should not be dismissed and a procedure based on, for instance, branch & bound will be studied. Regarding the heuristic resolution of this problem, another line of search that we will follow is the use of metaheuristics, which may obtain better results for large instances.

Acknowledgements

The authors wish to express their gratitude to the anonymous reviewers for their valuable comments, which have helped to improve the quality of this paper.

References

Al-Hawari, T., Ali, M., Al-Araidah, O. and Mumani, A. (2015). Development of a genetic algorithm for multi-objective assembly line balancing using multiple assignment approach. *The International Journal of Advanced Manufacturing Technology*, 77, 1419–32.

Battaïa, O. and Dolgui, A. (2012). Reduction approaches for a generalized assembly line balancing problem. *Computers and Operations Research*, 39, 2337–45.

Battaïa, O. and Dolgui, A. (2013). A taxonomy of line balancing problems and their solution approaches. *International Journal of Production Economics*, 142, 259–77.

Battaïa, O., Delorme, X., Dolgui., A., Hagemann, J., Horlemann, A., Kovalev, S. and Malyutin, S. (2015). Workforce minimization for a mixed-model assembly line in the automotive industry. *International Journal of Production Economics*, 170, 489–500.

Bautista, J., Batalla-García, C. and Alfaro-Pozo, R. (2016). Models for assembly line balancing by temporal, spatial and ergonomic risk attributes. *European Journal of Operational Research*, 251, 814–29.

Baybars, I. (1986). A survey of exact algorithms for the simple assembly line balancing problem. *Management Science*, 32, 909–32.

Becker, C. and Scholl, A. (2006). A survey on problems and methods in generalized assembly line balancing. *European Journal of Operational Research*, 168, 694–715.

Boysen, N. and Fliedner, M., Scholl, A. (2008). Assembly line balancing: Which model to use when?. *International Journal of Production Economics*, 111, 509–28.

Calleja, G., Corominas, A., García-Villoria, A. and Pastor, R. (2014). Combining matheuristics and MILP to solve the Accessibility Windows Assembly Line Balancing Problem Level 2 (AWALBP-L2). *Computers & Operations Research*, 48, 113–23.

Capacho, L., Pastor, R., Dolgui, A. and Guschinskaya, O. (2009). An evaluation of constructive heuristic methods for solving the alternative subgraphs assembly line balancing problem. *Journal of Heuristics*, 15, 109–32.

Chica, M., Bautista, J., Cordón, O. and Damas, S. (2016). A multiobjective model and evolutionary algorithms for robust time and space assembly line balancing under certain constraints. *Omega*, 58, 55–68.

Corominas, A., Ferrer, L. and Pastor, R. (2011). Assembly line balancing: general resource-constrained case. *International Journal of Production Research*, 49, 3527–42.

Erel, E. and Sarin, C.S. (1998). A survey of the assembly line balancing procedures. *Production Planning & Control*, 9, 414–34.

Esmaeilbeigi, R., Naderi, B. and Charkhgard, P. (2015). The type E simple assembly line balancing problem: A mixed integer linear programming formulation. *Computers & Operations Research*, 64, 168–177.

Gao, J., Sun, L., Wang, L. and Gen, M. (2009). An efficient approach for type II robotic assembly line balancing problems. *Computers & Industrial Engineering*, 56, 1065–80.

García-Villoria, A. and Pastor, R. (2013). Erratum to "A solution procedure for type E simple assembly line balancing problem". *Computers & Industrial Engineering*, 66, 201–22.

Gurevsky, E., Battaïa, O. and Dolgui, A. (2012). Balancing of simple assembly lines under variations of task processing times. *Annals of Operational Research*, 201, 265–86.

Gutjahr, A. L. and Nemhauser, G.L. (1964). An algorithm for the line balancing problem. *Management Science*, 11, 308–15.

Levitin, G., Rubinovitz, J. and Shnits, B. (2006). A genetic algorithm for robotic assembly line balancing. *European Journal of Operational Research*, 168, 811–25.

Martino, R. and Pastor, R. (2010). Heuristic procedures for solving the general assembly line balancing problem with setups. *International Journal of Production Research*, 48, 1787–804.

Moreira, M.C.O., Cordeau, J-F., Costa, A.M. and Laporte, G. (2015). Robust assembly line balancing with heterogeneous workers. *Computers & Industrial Engineering*, 88, 254–63.

Ogan, D. and Azizoglu, M. (2015). A branch and bound method for the line balancing problem in U-shaped assembly lines with equipment requirements. *Journal of Manufacturing Systems*, 36, 46–54.

Pastor, R. (2011). LB-ALBP: The lexicographical bottleneck assembly line balancing problem. *International Journal of Production Research*, 49, 2425–42.

Pastor, R., García-Villoria, A. and Corominas, A. (2011). Comparing ways of breaking symmetries in mathematical models for SALBP-1. *Assembly Automation*, 31, 385–7.

Plans, J. and Corominas, A. (1999). Modelling and solving the SALBP-E problem. *Proceedings of the 1999 IEEE International Symposium on Assembly and Task Planning*, Porto, Portugal, 356–60.

Rekiek, B., Dolgui, A., Deichambre, A. and Bratcu, A. (2002). State of art of optimization methods for assembly line design. *Annual Reviews in Control*, 26, 163–74.

Saif, U., Guan, Z., Liu, W., Zhang, C. and Wang, B. (2014). Pareto based artificial bee colony algorithm for multi objective single model assembly line balancing with uncertain task times. *Computers & Industrial Engineering*, 76, 1–15.

Saltzman, M.J. and Baybars, I. (1987). A two-process implicit enumeration algorithm for the simple assembly line balancing problem. *European Journal of Operational Research*, 32, 118–29.

Scholl, A. (1993). Data of assembly line balancing problem. *Schriften zur Quantitativen Betriebswirtschaftslehre*, Darmstadt University of Technology, Number 16/93.

Scholl, A. (1999). *Balancing and Sequencing of Assembly Lines*. Heidelberg, Physica-Verlag, 2nd rev.

Scholl, A. and Becker, C. (2006a). A survey on problems and methods in generalized assembly line balancing. *European Journal of Operational Research*, 168, 694–715.

Scholl, A. and Becker, C. (2006b). State-of-the-art exact and heuristic solution procedures for simple assembly line balancing. *European Journal of Operational Research*, 168, 666–93.

Scholl, A., Fliedner, M. and Boysen, N. (2010). Absalom: Balancing assembly lines with assignment restrictions. *European Journal of Operational Research*, 200, 688–701.

Tapkan, P., Özbakir, L. and Baykasoğlu, A. (2016). Bee algorithms for parallel two-sided assembly line balancing problem with walking times. *Applied Soft Computing*, 39, 275–91.

Uğurdağ, H.F., Rachamadugu, R. and Papachristou, C.A. (1997). Designing paced assembly lines with fixed number of stations. *European Journal of Operational Research*, 102, 488–501.

Wei, N.C. and Chao, I.M. (2011). A solution procedure for type E simple assembly line balancing problem. *Computers & Industrial Engineering*, 61, 824–30.

Yoosefelahi, A., Aminnayeri, M., Mosadegh, H. and Ardakani, D. (2012). Type II robotic assembly line balancing problem: An evolution strategies algorithm for a multi-objective model. *Journal of Manufacturing Systems*, 31, 139–51.

Zacharia, P.Th. and Nearchou, A.C. (2013). A meta-heuristic algorithm for the fuzzy assembly line balancing type-E problem. *Computers & Operations Research*, 12, 3033–44.

SORT 40 (2) July-December 2016, 243-266

Kernel-based estimation of $P(X > Y)$ in ranked set sampling

M. Mahdizadeh[1,*] and Ehsan Zamanzade[2]

Abstract

This article is directed at the problem of reliability estimation using ranked set sampling. A non-parametric estimator based on kernel density estimation is developed. The estimator is shown to be superior to its analog in simple random sampling. Monte Carlo simulations are employed to assess performance of the proposed estimator. Two real data sets are analysed for illustration.

MSC: 62G30, 62N05.

Keywords: Bandwidth selection, Judgment ranking, Stress-strength model.

1. Introduction

Ranked set sampling (RSS) is a cost-efficient alternative to simple random sampling (SRS) in situations where exact measurements of sample units are difficult or expensive to obtain but (judgment) ranking of them according to the variable of interest is relatively easy and cheap. A variety of methods can be used to implement the ranking, including visual inspection, expert opinion, or through the use of auxiliary variables, but it cannot entail actual measurements on the selected units. The RSS was first introduced by McIntyre (1952) in an agricultural experiment for estimating the mean pasture yield. Since then, it has been well adopted to environmental, ecological and health studies. The reader is referred to Chen (2007) for some novel applications in areas such as clinical trials and genetic quantitative trait loci mappings.

* *Corresponding author:* E-mail addresses: mahdizadeh.m@live.com (M. Mahdizadeh), e.zamanzade@sci.ui.ac.ir; ehsanzamanzadeh@yahoo.com (Ehsan Zamanzade)

[1] Department of Statistics, Hakim Sabzevari University, P.O. Box 397, Sabzevar, Iran

[2] Department of Statistics, University of Isfahan, Isfahan 81746-73441, Iran

Received: February 2016
Accepted: August 2016

The RSS procedure can be described as follows. First, m^2 units are collected as independent and identically distributed draws from the population. These units are randomly partitioned into m sets, each of size m. In the first set, the response judged to be smallest is taken for full quantification; in the second set, the response judged to be second smallest is taken; and so on, until in the last set, the response judged to be largest is taken. These measured values, along with the associated ranks form a ranked set sample of size m. The parameter m is called set size, which should be kept small to facilitate the judgment ranking process. Let $X_{[i]}\,(i = 1,\ldots,m)$ be the ith judgement order statistic from the ith set; then the resulting sample is denoted by $X_{[1]},\ldots,X_{[m]}$. Here, the square bracket is used to indicate that the judgement ranks may not be correct. If our ranking is accurate, then we replace the square brackets with the round ones, and $X_{(i)}$ becomes the ith true order statistic from the ith set. If a larger sample size is needed the above procedure may be repeated k times (cycles). So a ranked set sample, in its general setup, may be represented by $\{X_{[i]r} : i = 1,\ldots,m; r = 1,\ldots,k\}$, where $X_{[i]r}$ is the ith judgement order statistic in the rth cycle.

A ranked set sample contains more information than a simple random sample of comparable size because it contains not only information carried by quantified observations but also information provided by the ranking process. Thus, it is expected that statistical procedures based on RSS tend to be superior to their SRS analogues. For a good review of RSS and its applications, see Chen et el. (2004). The interested reader is also referred to Wolfe (2004, 2010) and the references therein. Mahdizadeh and Arghami (2013), and Tahmasebi and Jafari (2014) are examples of recently published papers on RSS methods.

The stress-strength model, in its simplest form, defines the reliability of a component as the probability that the strength of the unit (X) is greater than the stress (Y) imposed on it. The quantity $\theta = P(X > Y)$ is referred to as the reliability parameter. Although the use of stress-strength models was originally motivated by problems in physics and engineering, it is not limited to these contexts. It is worth mentioning that θ provides a general measure of the difference between two populations, and has found applications in different fields such as economics, quality control, psychology, medicine and clinical trials. For instance, if Y is the response of a control group, and X is that of a treatment group, then θ is a measure of the treatment effect. This situation is exemplified in Section 5.

There has been continuous interest in the problem of estimating θ when X and Y are independent variables, and belong to the same family of distributions. A comprehensive account of this topic appears in Kotz et al. (2003). The reliability estimation under RSS has also drawn some attention. Muttlak et al. (2010) derived estimators for θ using RSS in the case of the exponential distribution. Sengupta and Mukhuti (2008) studied unbiased estimation of θ using RSS in nonparametric setting based on the empirical distribution function. They showed that the proposed estimator is more efficient than its SRS counterpart, even in the presence of ranking errors. In this work, the kernel density estimator is used to suggest a new estimator.

Section 2 presents the estimator along with some notions and results which will be used in the sequel. Theoretical properties are studied in Section 3. Results from simulation experiments appear in Section 4. Two applications are provided in Section 5. A summary and concluding remarks are given in Section 6. Proofs are postponed to an appendix.

2. The proposed estimator

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be independent random samples from two continuous populations with density functions f and g, respectively. The corresponding distribution functions are denoted by F and G. The standard nonparametric estimator of θ is

$$\tilde{\theta}_{\mathrm{SRS}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i > Y_j). \tag{1}$$

Under the assumptions of independence, it is possible to write

$$\theta = P(X > Y) = E\left(I(X > Y)\right) = \int\int I(u > v) f(u) g(v)\,\mathrm{d}u\,\mathrm{d}v, \tag{2}$$

where $I(.)$ is the usual indicator function. An alternative estimator of θ can be made by replacing f and g in (2) with some estimates. To this end, the kernel density estimators may be utilized which are given by

$$\hat{f}(u) = \frac{1}{mh_1}\sum_{i=1}^{m} K\left(\frac{u - X_i}{h_1}\right)$$

and

$$\hat{g}(v) = \frac{1}{nh_2}\sum_{j=1}^{n} K\left(\frac{v - Y_j}{h_2}\right),$$

where the kernel K is a symmetric probability density, and the smoothing parameters h_1 and h_2 are known as the bandwidths.

Incorporating $\hat{f}$ and $\hat{g}$ in (2), we have

$$\begin{aligned}\hat{\theta}_{\mathrm{SRS}} &= \int\int I(u>v)\hat{f}(u)\hat{g}(v)\,\mathrm{d}u\,\mathrm{d}v \\ &= \int\int I(u>v)\left[\frac{1}{mh_1}\sum_{i=1}^{m}K\left(\frac{u-X_i}{h_1}\right)\right]\left[\frac{1}{nh_2}\sum_{j=1}^{n}K\left(\frac{v-Y_j}{h_2}\right)\right]\mathrm{d}u\,\mathrm{d}v \\ &= \int\int\int\int I(u>v)\left[\frac{1}{h_1}K\left(\frac{u-x}{h_1}\right)\right]\left[\frac{1}{h_2}K\left(\frac{v-y}{h_2}\right)\right]\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}\hat{F}(x)\,\mathrm{d}\hat{G}(y), \quad (3)\end{aligned}$$

where $\hat{F}$ and $\hat{G}$ are the empirical distribution functions. Using the change of variables $r=(u-x)/h_1$ and $s=(v-y)/h_2$ in (3), it follows that

$$\begin{aligned}\hat{\theta}_{\mathrm{SRS}} &= \int\int\left[\int\int I(h_1r+x>h_2s+y)K(r)K(s)\,\mathrm{d}r\,\mathrm{d}s\right]\mathrm{d}\hat{F}(x)\,\mathrm{d}\hat{G}(y) \\ &= \int\int H(x-y)\,\mathrm{d}\hat{F}(x)\,\mathrm{d}\hat{G}(y),\end{aligned}$$

where H is the distribution function of h_2S-h_1R and R and S are independent random variables with common density K. If K is the standard normal density, then H is the distribution function of a normal random variable with mean 0 and standard deviation $t=\sqrt{h_1^2+h_2^2}$. In this case, $\hat{\theta}_{\mathrm{SRS}}$ takes the form

$$\hat{\theta}_{\mathrm{SRS}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi\left(\frac{X_i-Y_j}{t}\right), \quad (4)$$

where $\Phi(.)$ is the standard normal distribution function. Baklizi and Eidous (2006) used the above estimator to construct confidence intervals for θ.

Proceeding in the same way, we arrive at the RSS analogue of (4) defined as

$$\hat{\theta}_{\mathrm{RSS}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right), \quad (5)$$

where $X_{[1]},\ldots,X_{[m]}$ and $Y_{[1]},\ldots,Y_{[n]}$ are ranked set samples drawn from f and g (with a single cycle), respectively. In the next section, properties of this estimator are studied. The results can be extended for other choices of the kernel function.

The success of RSS procedures hinges on how well the within-set rankings to select the units for measurement can be achieved. Although perfect rankings are the ideal case for any RSS-based method, it is unlikely to be feasible. Thus it is worth in practice to evaluate the effect of imperfect rankings on our procedures. The proper way to this would be using statistical models designed to capture possible errors in the ranking

process. A number of such imperfect ranking models can be found in the literature. We build on a model introduced by Bohn and Wolfe (1994). They consider the distributions of the judgment order statistics to be mixtures of distributions of the true order statistics. The model is now set forth for our two-sample problem.

The density functions of the ith true and judgement order statistic of a random sample of size m from f are denoted by $f_{(i)}$ and $f_{[i]}$, respectively. Similar notations are used for a random sample of size n from G. We postulate an imperfect ranking model M_X under which $X_{[i]}$'s are assumed to be independently distributed as

$$P\left(X_{[i]} = X_{(r)}\right) = p_{ir}, \quad (r = 1, \ldots, m),$$

where p_{ir} is the probability that the rth order statistic is judged to have rank i, and thus $\sum_{r=1}^{m} p_{ir} = 1$. It is further assumed that $\sum_{i=1}^{m} p_{ir} = 1$. Obviously, this is true in the perfect ranking scenario, i.e. when $p_{ii} = 1$ and $p_{ir} = 0\,(r \neq i)$. Similarly, we postulate an imperfect ranking model M_Y under which $Y_{[j]}$'s are assumed to be independently distributed as

$$P\left(Y_{[j]} = Y_{(s)}\right) = q_{js}, \quad (s = 1, \ldots, n),$$

where q_{js} is the probability that the sth order statistic is judged to have rank j, and therefore $\sum_{s=1}^{n} q_{js} = 1$. Moreover, it is assumed that $\sum_{j=1}^{n} q_{js} = 1$. The model considering M_X and M_Y together is referred to as M. Also, misplacement probability matrices are denoted by $\mathbf{P} = [p_{ir}]_{m\times m}$ and $\mathbf{Q} = [q_{js}]_{n\times n}$.

According to a basic identity in RSS, which simply follows from the binomial expansion, we have

$$\frac{1}{m}\sum_{i=1}^{m} f_{(i)}(x) = f(x), \quad \frac{1}{n}\sum_{j=1}^{n} g_{(j)}(y) = g(y). \tag{6}$$

For details, see Chen et al. (2004, Chapter 2). It is easy to verify that these equations also hold under the model M, i.e.

$$\frac{1}{m}\sum_{i=1}^{m} f_{[i]}(x) = f(x), \quad \frac{1}{n}\sum_{j=1}^{n} g_{[j]}(y) = g(y). \tag{7}$$

These identities are repeatedly used in the sequel.

3. Main results

Theoretical properties of the suggested estimator are studied in this section. It can be seen that (4) and (5) are both biased and their expectations are $E\left\{\Phi\left(\frac{X-Y}{t}\right)\right\}$. The next proposition presents variance expression for the two estimators.

Proposition 1 *The variances of $\hat{\theta}_{SRS}$ and $\hat{\theta}_{RSS}$ are given by*

$$m^2n^2\,\mathrm{Var}(\hat{\theta}_{SRS}) = mn(n-1)EE^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\} + nm(m-1)EE^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|Y\right\} + mnE\left\{\Phi^2\left(\frac{X-Y}{t}\right)\right\} + \left(mn-m^2n-n^2m\right)E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\right\} \tag{8}$$

and

$$\begin{aligned} m^2n^2\,\mathrm{Var}(\hat{\theta}_{RSS}) &= mE\left(n^2E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\} - \sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right) \\ &\quad + E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2 - \sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right) \\ &\quad + mnE\left\{\Phi^2\left(\frac{X-Y}{t}\right)\right\} - m^2n^2E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\right\}. \end{aligned} \tag{9}$$

The variances of $\hat{\theta}_{\mathrm{SRS}}$ and $\hat{\theta}_{\mathrm{RSS}}$ are compared in the next proposition.

Proposition 2 *Under model M,* $\mathrm{Var}(\hat{\theta}_{RSS}) \leq \mathrm{Var}(\hat{\theta}_{SRS})$*, and the equality holds if $f_{[i]} = f\,(i = 1,\ldots,m)$ and $g_{[j]} = g\,(j = 1,\ldots,n)$. The latter happens when $p_{ir} = 1/m\,(i,r = 1,\ldots,m)$ and $q_{js} = 1/n\,(j,s = 1,\ldots,n)$.*

The RSS-based procedures tend to outperform their SRS analogues as long as the judgment ranking is not random. In the case of estimating θ, this was formally shown (under model M) in the previous proposition. The maximum efficiency is expected to happen in the perfect ranking setup. We now give a result confirming this property. It should be mentioned that the approach adopted in proof is distinctly different from that of similar result in Sengupta and Mukhuti (2008).

Proposition 3 *Under model M, the variance of $\hat{\theta}_{RSS}$ is minimized in the absence of ranking errors.*

It is worth noting that the case of perfect rankings is not the only one where the minimum variance is achieved. It would also be attained in the case where model M holds, and $\mathbf{P}$ and $\mathbf{Q}$ are permutation matrices. In addition, there are cases in the RSS literature where an appropriately chosen imperfect rankings scheme can lead to more efficient estimation than is possible with perfect rankings. To put it another way, Proposition 3 may not hold under other imperfect ranking models.

We close this section by some remarks on a general form of our proposed estimator. The estimator (5) is defined for the case where RSS is done with a single cycle. The ranked set sample size, however, is increased not by increasing the set size, but by increasing the number of cycles. It is therefore important to study the multi-cycle case as well. In this setup, the estimator is given by

$$\hat{\theta}_{\mathrm{RSS}} = \frac{1}{mkn\ell} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{r=1}^{k} \sum_{s=1}^{\ell} \Phi\left(\frac{X_{[i]r} - Y_{[j]s}}{t} \right), \tag{10}$$

where $\{X_{[i]r} : i = 1, \ldots, m; r = 1, \ldots, k\}$ and $\{Y_{[j]s} : j = 1, \ldots, n; s = 1, \ldots, \ell\}$ are ranked set samples of size mk and $n\ell$ drawn from f and g, respectively. The above estimator can be represented as

$$\hat{\theta}_{\mathrm{RSS}} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} h\left(X_{[i]}, Y_{[j]}\right),$$

where

$$h\left(X_{[i]}, Y_{[j]}\right) = \frac{1}{k\ell} \sum_{r=1}^{k} \sum_{s=1}^{\ell} \Phi\left(\frac{X_{[i]r} - Y_{[j]s}}{t} \right).$$

Now, one can proceed with proving analogues of Propositions 1-3. The steps are similar to current proofs in which

$$h\left(X_{[i]}, Y_{[j]}\right) = \Phi\left(\frac{X_{[i]} - Y_{[j]}}{t} \right).$$

As a reviewer pointed out, the estimator (10) was also studied by Yin et al. (2016). The authors, however, build on the theory of U-statistics in computing the variance expression. Moreover, they only show that this estimator is asymptotically more efficient than its counterpart in SRS. And last but not least, no theoretical result in the imperfect ranking setup is provided in the aforesaid paper.

For simplicity, we consider the estimator (5) in the next section. But the data analysis in Section 5 is based on the estimator (10).

4. Simulation study

This section contains results of simulation studies performed to evaluate behaviours of $\hat{\theta}_{\text{SRS}}$ and $\hat{\theta}_{\text{RSS}}$. It is assumed that both populations follow normal, exponential or uniform distribution. Suppose X and $Y-\mu$ are standard normal random variables. Then, it is simply shown that

$$\theta = \Phi\left(\frac{-\mu}{\sqrt{2}}\right),$$

where $\Phi(.)$ is the distribution function of X. Similarly, for standard exponential random variables X and Y/α, we have

$$\theta = \frac{1}{1+\alpha}.$$

Finally, let X and Y/β be uniformly distributed on the unit interval. Then, it follows that

$$\theta = \begin{cases} 1-\beta/2 & 0<\beta<1 \\ 1/(2\beta) & \beta \geq 1 \end{cases}.$$

Under each parent distribution, five values were assigned to the associated parameter so as to produce $\theta = 0.1, 0.3, 0.5, 0.7, 0.9$. The appropriate parameter values are given in Table 1. Also, set sizes $(m,n) = (3,3),(3,7),(5,5),(10,10)$ were selected.

Table 1: *Parameter values corresponding to different reliability parameters.*

	θ				
Parameter	0.1	0.3	0.5	0.7	0.9
μ	1.812388	0.7416143	0	−0.7416143	−1.812388
α	9	7/3	1	3/7	1/9
β	5	5/3	1	3/5	1/5

We first consider the perfect ranking situation. For each combination of distributions and sample sizes, 10,000 pairs of samples were generated in SRS and RSS settings. The two estimators were computed from each pair of samples in the corresponding designs, and their mean squared errors (MSEs) were determined. The relative efficiency (RE) is defined as the ratio of $\widehat{MSE}(\hat{\theta}_{\text{SRS}})$ to $\widehat{MSE}(\hat{\theta}_{\text{RSS}})$. The RE values larger than one indicate that $\hat{\theta}_{\text{RSS}}$ is more efficient than $\hat{\theta}_{\text{SRS}}$. Tables 2 and 3 display the results (to save space, tables for the uniform distribution are provided as supplementary material), where RE1-RE4 are based on the following four methods for bandwidth selection, re-

Table 2: *Estimated REs under normal distribution (RE1, RE2, RE3 and RE4 are based on bandwidth selection using AMISE, UCV, BCV and PI methods, respectively).*

(m,n)	θ	RE1	RE2	RE3	RE4
(3,3)	0.1	1.00 (0.38)	1.02 (0.40)	1.00 (0.38)	1.56 (1.17)
	0.3	1.48 (1.74)	1.53 (1.76)	1.48 (1.74)	1.95 (1.34)
	0.5	2.38 (4.35)	2.36 (4.02)	2.38 (4.35)	2.04 (1.36)
	0.7	1.44 (1.71)	1.49 (1.73)	1.44 (1.71)	1.91 (1.32)
	0.9	0.99 (0.38)	1.00 (0.41)	0.99 (0.38)	1.51 (1.18)
(3,7)	0.1	1.00 (0.25)	1.04 (0.32)	1.00 (0.25)	1.68 (1.00)
	0.3	1.45 (1.23)	1.64 (1.30)	1.45 (1.23)	2.23 (1.29)
	0.5	2.71 (4.03)	2.72 (3.07)	2.71 (4.03)	2.42 (1.38)
	0.7	1.43 (1.25)	1.61 (1.32)	1.43 (1.25)	2.18 (1.31)
	0.9	0.99 (0.25)	1.01 (0.32)	0.99 (0.25)	1.59 (1.00)
(5,5)	0.1	1.01 (0.20)	1.09 (0.30)	1.01 (0.20)	1.81 (0.91)
	0.3	1.46 (0.96)	1.83 (1.13)	1.46 (0.95)	2.73 (1.30)
	0.5	3.51 (4.27)	3.52 (2.93)	3.51 (4.28)	3.14 (1.43)
	0.7	1.48 (0.95)	1.87 (1.14)	1.48 (0.95)	2.77 (1.31)
	0.9	1.02 (0.20)	1.11 (0.30)	1.02 (0.20)	1.84 (0.92)
(10,10)	0.1	1.02 (0.08)	1.19 (0.19)	1.02 (0.08)	2.14 (0.60)
	0.3	1.43 (0.37)	2.30 (0.64)	1.43 (0.37)	4.13 (1.06)
	0.5	5.78 (3.42)	6.12 (2.18)	5.78 (3.42)	5.81 (1.47)
	0.7	1.44 (0.38)	2.28 (0.65)	1.44 (0.38)	4.09 (1.07)
	0.9	1.02 (0.09)	1.18 (0.19)	1.02 (0.09)	2.11 (0.61)

spectively. Minimizing asymptotic mean integrated squared error (AMISE) of the kernel density estimator is a basic scheme. Rudemo (1982) and Bowman (1984) proposed unbiased (least-squares) cross-validation (UCV) method. Biased cross-validation (BCV) was studied by Scott and George (1987). A plug-in (PI) method was suggested by Sheather and Jones (1991). All these techniques are developed for SRS, and more details on them can be found in Sheather (2004). The methods can be implemented in R statistical software using the `kedd` and `KernSmooth` packages. In the RSS setup, we treat data as if collected by SRS to choose bandwidth.

Sengupta and Mukhuti (2008) introduced the RSS competitor of (1) defined as

$$\tilde{\theta}_{\mathrm{RSS}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(X_{[i]} > Y_{[j]}). \tag{11}$$

The entries of Tables 2–4 appearing in parentheses show efficiency of $\hat{\theta}_{\mathrm{RSS}}$ relative to $\tilde{\theta}_{\mathrm{RSS}}$.

Table 3: *Estimated REs under exponential distribution (RE1, RE2, RE3 and RE4 are based on bandwidth selection using AMISE, UCV, BCV and PI methods, respectively).*

(m,n)	θ	RE1	RE2	RE3	RE4
(3,3)	0.1	0.96 (0.33)	0.97 (0.35)	0.96 (0.33)	1.37 (1.19)
	0.3	1.51 (1.93)	1.56 (1.96)	1.51 (1.93)	1.99 (1.51)
	0.5	2.23 (4.11)	2.21 (3.79)	2.23 (4.11)	2.04 (1.38)
	0.7	1.50 (1.90)	1.55 (1.94)	1.50 (1.90)	2.00 (1.50)
	0.9	0.97 (0.33)	0.97 (0.35)	0.97 (0.33)	1.35 (1.20)
(3,7)	0.1	0.93 (0.13)	1.04 (0.33)	0.93 (0.13)	1.27 (0.71)
	0.3	1.39 (1.33)	1.88 (1.58)	1.39 (1.33)	2.44 (1.52)
	0.5	2.45 (3.76)	2.43 (2.55)	2.45 (3.76)	2.41 (1.37)
	0.7	1.45 (1.34)	1.64 (1.47)	1.45 (1.34)	2.21 (1.43)
	0.9	0.93 (0.26)	0.93 (0.28)	0.93 (0.26)	1.26 (0.98)
(5,5)	0.1	0.92 (0.14)	1.02 (0.28)	0.92 (0.14)	1.32 (0.74)
	0.3	1.44 (1.07)	2.13 (1.48)	1.44 (1.07)	2.83 (1.50)
	0.5	2.94 (3.78)	3.00 (2.38)	2.94 (3.78)	2.96 (1.40)
	0.7	1.43 (1.08)	2.07 (1.47)	1.43 (1.08)	2.78 (1.50)
	0.9	0.93 (0.15)	1.01 (0.28)	0.93 (0.15)	1.31 (0.75)
(10,10)	0.1	0.93 (0.05)	1.05 (0.19)	0.93 (0.05)	1.22 (0.32)
	0.3	1.30 (0.38)	3.42 (1.14)	1.30 (0.38)	4.33 (1.32)
	0.5	4.77 (3.14)	5.33 (1.67)	4.77 (3.14)	5.50 (1.41)
	0.7	1.27 (0.37)	3.34 (1.12)	1.27 (0.37)	4.30 (1.30)
	0.9	0.93 (0.05)	1.04 (0.19)	0.93 (0.05)	1.23 (0.33)

It is observed that $\hat{\theta}_{\mathrm{SRS}}$ is outperformed by $\hat{\theta}_{\mathrm{RSS}}$, using one of the bandwidths, at least. Moreover, the results from AMISE and BCV methods are in close agreement. For each pair of sample sizes, the RE values are generally larger when the reliability parameter is 0.5. Also, the PI method works better than the others for $\theta \neq 0.5$. It is to be noted that in the case of $\theta = 0.1, 0.9$, only RE4 values exceed unity markedly. Given a total sample size, the efficiency gain is generally larger for equal sample sizes setup. Compare similar REs for $(m,n) = (3,7), (5,5)$ under different parent distributions.

As to the comparison of the suggested estimator with its rival based on empirical distribution function, the following conclusions can be made (based on entries given in parentheses in Tables 2 and 3). Again, the RE values are generally larger for the reliability parameter 0.5, given a pair of sample sizes. There are cases that $\hat{\theta}_{\mathrm{RSS}}$ is less efficient than $\tilde{\theta}_{\mathrm{RSS}}$. For example, see the results when $(m,n) = (5,5), (10,10)$ and $\theta = 0.1, 0.9$. Sometimes the REs from AMISE, UCV and BCV methods fall much below unity. In such instances, the PI method is still the best one.

Table 4: *Estimated REs for $(m,n) = (3,3)$ (upper panel) and $(m,n) = (5,5)$ (lower panel) with imperfect ranking (RE1, RE2, RE3 and RE4 are based on bandwidth selection using AMISE, UCV, BCV and PI methods, respectively).*

Dist.	θ	RE1	RE2	RE3	RE4
Normal	0.1	1.03 (0.42)	1.05 (0.45)	1.03 (0.42)	1.47 (1.18)
	0.3	1.41 (1.89)	1.44 (1.90)	1.41 (1.89)	1.65 (1.31)
	0.5	1.94 (4.10)	1.92 (3.78)	1.94 (4.10)	1.70 (1.31)
	0.7	1.34 (1.82)	1.37 (1.82)	1.34 (1.82)	1.63 (1.29)
	0.9	0.99 (0.43)	1.00 (0.46)	0.99 (0.43)	1.37 (1.18)
Exponential	0.1	0.98 (0.38)	0.98 (0.40)	0.98 (0.38)	1.30 (1.26)
	0.3	1.40 (2.03)	1.43 (2.05)	1.40 (2.03)	1.71 (1.46)
	0.5	1.82 (4.01)	1.81 (3.70)	1.82 (4.01)	1.69 (1.36)
	0.7	1.36 (2.01)	1.39 (2.03)	1.36 (2.01)	1.66 (1.45)
	0.9	0.96 (0.37)	0.96 (0.39)	0.96 (0.37)	1.21 (1.22)
Normal	0.1	1.02 (0.22)	1.10 (0.33)	1.02 (0.22)	1.65 (0.94)
	0.3	1.42 (1.08)	1.70 (1.24)	1.42 (1.08)	2.31 (1.30)
	0.5	2.81 (4.07)	2.80 (2.79)	2.81 (4.07)	2.58 (1.40)
	0.7	1.39 (1.07)	1.70 (1.23)	1.39 (1.07)	2.33 (1.30)
	0.9	1.00 (0.23)	1.07 (0.34)	1.00 (0.23)	1.60 (0.95)
Exponential	0.1	0.94 (0.16)	1.00 (0.30)	0.94 (0.16)	1.26 (0.79)
	0.3	1.39 (1.19)	1.91 (1.55)	1.39 (1.19)	2.41 (1.49)
	0.5	2.59 (3.82)	2.63 (2.38)	2.59 (3.82)	2.58 (1.39)
	0.7	1.42 (1.16)	1.99 (1.52)	1.42 (1.16)	2.57 (1.49)
	0.9	0.94 (0.16)	1.02 (0.30)	0.94 (0.16)	1.29 (0.78)

As mentioned before, although perfect rankings are ideal case for any RSS-based method, it is unlikely to be feasible. Let $\mathbf{P}$ and $\mathbf{Q}$ be misplacement probability matrices defined in Section 2. The perfect ranking setup corresponds to the case that $\mathbf{P}$ and $\mathbf{Q}$ are the identity matrices. We conducted a partial simulation study to assess performance of the suggested estimator in the presence of ranking errors. To do so, the REs were estimated when $(m,n) = (3,3),(5,5)$ and the matrices $\mathbf{P}$ and $\mathbf{Q}$ are selected to be

$$\mathbf{P} = \mathbf{Q} = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.1 & 0.9 \end{bmatrix}$$

and

$$\mathbf{P} = \mathbf{Q} = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0.8 & 0.1 \\ 0 & 0 & 0 & 0.1 & 0.9 \end{bmatrix},$$

respectively. The results are given in Table 4. The entries outside parentheses are generally smaller than similar entries in Tables 2 and 3. We note, however, that these REs still exceed the unity, and this is consistent with our theoretical results. There is not a uniform trend for entries inside parentheses as compared with analogous ones under perfect ranking assumption. It is to be mentioned that these REs are associated with $\hat{\theta}_{\mathrm{RSS}}$ and $\tilde{\theta}_{\mathrm{RSS}}$ that both are affected by ranking errors.

All simulation studies in this work are programmed using R statistical software, and the corresponding code is available from the first author.

5. Application

The RSS is applicable in the following situations: (i) the ranking of a set of sampling units can be done easily by judgment relating to their latent values of the variable of interest through visual inspection, expert opinion, etc. (ii) there are certain easily accessible concomitant variables. We now illustrate the proposed procedure using some real data from two different fields.

5.1. Agriculture

Murray et al. (2000) conducted an experiment in which apple trees are sprayed with chemical containing fluorescent tracer, Tinopal CBS-X, at 2% concentration level in water. Two nine-tree plots were chosen for spraying. One plot was sprayed at high volume, using coarse nozzles on the sprayer to give a large average droplet size. The other plot was sprayed at low volume, using fine nozzles to give a small average droplet size. Fifty sets of five leaves were identified from the central five trees of each plot, and used to draw a ranked set sample with set size 5 and cycle size 10, from each plot. The variable of interest is the percentage of area covered by the spray on the surface of the leaves. The formal measurement entails chemical analysis of the solution collected from the surface of the leaves, and thereby is a time-consuming and expensive process. The judgment ranking within each set is based on the visual appearance of the spray deposits on the leaf surfaces when viewed under ultraviolet light. Clearly, the latter method is cheap, and fairly accurate if implemented by an expert observer.

The data are given in Table 5, where measurements obtained from the plot sprayed at high (low) volume constitute the control (treatment) group. The interest centres on knowing whether the sprayer settings affect the percentage area coverage. If X (Y) denotes the response variable from treatment (control) group, then $\hat{\theta}_{\mathrm{RSS}}$ and $\tilde{\theta}_{\mathrm{RSS}}$ can serve as measures of the treatment effect.

Let $\breve{\theta}$ be either $\hat{\theta}_{\mathrm{RSS}}$ or $\tilde{\theta}_{\mathrm{RSS}}$. Then the bootstrap method, introduced by Efron (1979), can be used to estimate the variance of $\breve{\theta}$, and to construct confidence interval. Modarres et al. (2006) suggested three bootstrap algorithms in RSS design. Bootstrap ranked set sampling (BRSS) and bootstrap RSS by rows (BRSSR) are the most efficient methods

Table 5: *Ranked set sample data for the percentage area covered on the surface of the leaves of apple trees.*

Group	Cycle	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Control	1	0.003	0.028	0.244	0.057	0.143
	2	0.039	0.119	0.126	0.105	0.565
	3	0.034	0.118	0.130	0.218	0.296
	4	0.051	0.104	0.193	0.210	0.150
	5	0.032	0.141	0.130	0.250	0.229
	6	0.069	0.070	0.260	0.225	0.285
	7	0.100	0.091	0.244	0.130	0.347
	8	0.012	0.096	0.069	0.373	0.133
	9	0.046	0.117	0.126	0.223	0.273
	10	0.028	0.083	0.108	0.212	0.261
Treatment	1	0.036	0.137	0.183	0.270	0.487
	2	0.250	0.181	0.290	0.328	0.715
	3	0.089	0.032	0.269	0.419	0.315
	4	0.180	0.111	0.130	0.194	0.742
	5	0.100	0.009	0.184	0.277	0.122
	6	0.042	0.089	0.199	0.269	0.395
	7	0.044	0.083	0.227	0.177	0.742
	8	0.044	0.171	0.067	0.192	0.336
	9	0.009	0.017	0.217	0.438	0.544
	10	0.071	0.132	0.310	0.343	0.379

which are used here. Suppose B pairs of bootstrap samples are drawn from the two ranked set samples by either of the algorithms. If $\breve{\theta}_b$ is value of the estimator based on data in the b th ($b = 1,\ldots,B$) replication, then the bootstrap variance estimator is given by

$$\widehat{\mathrm{Var}}_{\mathrm{boot}}(\breve{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\breve{\theta}_b - \bar{\theta}\right)^2, \tag{12}$$

where $\bar{\theta} = \sum_{b=1}^{B}\breve{\theta}_b/B$. An approximate $(1-\alpha)$ normal interval for θ is then constructed as

$$\left(\breve{\theta} - z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_{\mathrm{boot}}(\breve{\theta})}, \breve{\theta} + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_{\mathrm{boot}}(\breve{\theta})}\right), \tag{13}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. We may alternatively use $(1-\alpha)$ bootstrap percentile interval defined as

$$\left(\breve{\theta}_{\alpha/2}, \breve{\theta}_{1-\alpha/2}\right), \tag{14}$$

where $\breve{\theta}_\beta$ is the β quantile of $\breve{\theta}_1,\ldots,\breve{\theta}_B$.

Table 6: *Estimates of θ along with their estimated variances, and the corresponding 0.95 confidence intervals.*

Estimator	Value	Estimated variance	Normal interval	Bootstrap interval
$\hat{\theta}_{\mathrm{RSS}}$ (AMISE)	0.5903	0.000456	(0.548, 0.632)	(0.550, 0.633)
		0.000468	*(0.548, 0.633)*	*(0.549, 0.634)*
$\hat{\theta}_{\mathrm{RSS}}$ (UCV)	0.6161	0.001021	(0.553, 0.679)	(0.557, 0.680)
		0.001052	*(0.553, 0.680)*	*(0.556, 0.684)*
$\hat{\theta}_{\mathrm{RSS}}$ (BCV)	0.6118	0.000748	(0.558, 0.665)	(0.558, 0.664)
		0.000789	*(0.557, 0.667)*	*(0.557, 0.667)*
$\hat{\theta}_{\mathrm{RSS}}$ (PI)	0.6168	0.000927	(0.557, 0.676)	(0.559, 0.678)
		0.000964	*(0.556, 0.678)*	*(0.558, 0.680)*
$\tilde{\theta}_{\mathrm{RSS}}$	0.6184	0.001163	(0.552, 0.685)	(0.553, 0.685)
		0.001224	*(0.550, 0.687)*	*(0.552, 0.688)*

Table 6 displays the estimates along with their estimated variances computed using (12). Two 0.95 intervals (13) and (14) are also reported. The number of bootstrap replications is chosen to be 5000, and entries associated with BRSSR method are in italic. Clearly, the kernel-based estimators have smaller estimated variances as expected. It is concluded that the treatment effect is significant at 0.05 level as none of the intervals contain 0.5.

5.2. Medicine

The RSS can be used in studying certain medical measures, which usually involves expensive laboratory tests. Samawi et al. (2009) employed this design in comparing bilirubin level between male and female jaundice babies. To this end, blood sample must be taken from the sampled babies and tested in a laboratory. But, on the other hand, the ranking of the bilirubin levels of a small number of babies can be done by observing whether their face, chest, lower parts of the body and the terminal parts of the whole body are yellowish. The yellowish color goes from face to the terminal parts of the whole body, the level of bilirubin in blood goes higher.

Table 7 shows the results of 15 measurements for male/female babies collected by RSS with set size 3 and cycle size 5. Assume that X and Y represent the response variable for male and female babies, respectively. Then $\hat{\theta}_{\mathrm{RSS}}$ and $\tilde{\theta}_{\mathrm{RSS}}$ can be used to decide whether male babies are more likely to experience jaundice. Table 8 displays the estimates along with their estimated variances. The corresponding 0.95 confidence intervals are also provided. Again, the kernel-based estimators have smaller estimated variances. All the intervals contain 0.5, and the null hypothesis that male and female babies are equally likely to experience jaundice is not rejected, at 0.05 level.

Table 7: *Ranked set sample data of bilirubin level in jaundice babies.*

Group	Cycle	Rank 1	Rank 2	Rank 3
Male	1	7.50	10.50	7.30
	2	7.50	15.00	8.60
	3	8.90	14.60	13.53
	4	7.00	11.90	15.70
	5	10.24	13.18	18.47
Female	1	1.20	8.94	15.00
	2	7.50	12.82	10.80
	3	8.00	8.82	10.70
	4	8.90	8.94	14.59
	5	8.53	8.20	18.29

Table 8: *Estimates of θ along with their estimated variances, and the corresponding 0.95 confidence intervals.*

Estimator	Value	Estimated variance	Normal interval	Bootstrap interval
$\hat{\theta}_{\text{RSS}}$ (AMISE)	0.5549	0.002183	(0.463, 0.646)	(0.464, 0.649)
		0.001813	*(0.471, 0.638)*	*(0.474, 0.638)*
$\hat{\theta}_{\text{RSS}}$ (UCV)	0.5753	0.005465	(0.430, 0.720)	(0.409, 0.704)
		0.004715	*(0.441, 0.710)*	*(0.421, 0.687)*
$\hat{\theta}_{\text{RSS}}$ (BCV)	0.5576	0.002398	(0.462, 0.654)	(0.464, 0.657)
		0.002016	*(0.470, 0.646)*	*(0.473, 0.647)*
$\hat{\theta}_{\text{RSS}}$ (PI)	0.5774	0.005067	(0.438, 0.717)	(0.434, 0.717)
		0.004172	*(0.451, 0.704)*	*(0.450, 0.700)*
$\tilde{\theta}_{\text{RSS}}$	0.5467	0.006564	(0.388, 0.705)	(0.382, 0.707)
		0.005709	*(0.399, 0.695)*	*(0.400, 0.689)*

6. Conclusion

The RSS design employs ranking of the characteristic of interest via auxiliary information to improve estimation of population attributes. The rankings can be performed through subjective judgment, concomitant variable, or a combination of them. These preparatory rankings are made before any actual measurements on the variable of interest, and are utilized to select more informative units to include in our sample for measurement.

In this article, a nonparametric reliability estimator based on kernel density estimation is suggested. Some theoretical results are presented under an imperfect ranking model. The perfect ranking setup is treated separately. Monte Carlo simulations are

used to compare the estimator with its SRS competitor, and the RSS analogue based on empirical distribution function. The results confirm preference of the new estimator in many situations. In a subsequent work, we plan to study interval estimation of the reliability parameter under the RSS scheme.

Acknowledgments

We are most grateful to the reviewers and the Editor-in-Chief for constructive comments on an earlier version of this paper.

References

Baklizi, A. and Eidous, O. (2006). Nonparametric estimation of $P(X < Y)$ using kernel methods. *Metron*, LXIV, 47–60.

Bohn, L.L. and Wolfe, D.A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association*, 89, 168–176.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of kernel density estimates. *Biometrika*, 71, 353–360.

Chen, Z. (2007). Ranked set sampling: its essence and some new applications. *Environmental and Ecological Statistics*, 14, 355–363.

Chen, Z., Bai, Z. and Sinha, B.K. (2004). *Ranked set sampling: Theory and Applications*. Springer, New York.

Datta, B.N. (2010). *Numerical Linear Algebra and Applications*, Second Edition. SIAM.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Kotz, S., Lumelskii, Y. and Pensky, M. (2003). The stress-strength model and its generalizations. *Theory and applications. World Scientific*, Singapore.

Mahdizadeh, M. and Arghami, N.R. (2013). Improved entropy based test of uniformity using ranked set sample. *SORT*, 37, 3–18.

McIntyre, G.A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385–390.

Modarres, R., Hui, T.P. and Zheng, G. (2006). Resampling methods for ranked set samples. *Computational Statistics and Data Analysis*, 51, 1039–1050.

Murray, R.A., Ridout, M.S. and Cross, J.V. (2000). The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology*, 57, 141–146.

Muttlak, H.A., Abu-Dayyeh, W.A., Al-Saleh, M.F. and Al-Sawi, E. (2010). Estimating $P(Y < X)$ using ranked set sampling in case of the exponential distribution. *Communications in Statistics: Theory and Methods*, 39, 1855–1868.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9, 65–78.

Samawi, H.M., Al-Saleh, M.F. and Al-Saidy, O. (2009). The matched pair sign test using bivariate ranked set sampling for different ranking based schemes. *Statistical Methodology*, 6, 397–407.

Scott, D.W. and George, R.T. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82, 1131–1146.

Sengupta, S. and Mukhuti, S. (2008). Unbiased estimation of $P(X > Y)$ using ranked set sample data. *Statistics*, 42, 223–230.

Sheather, S.J. (2004). Density estimation. *Statistical Science*, 19, 588–597.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B 53, 683–690.

Tahmasebi, S. and Jafari, A.A. (2014). Estimators for the parameter mean of Morgenstern type bivariate generalized exponential distribution using ranked set sampling. *SORT*, 38, 161–180.

Wolfe, D.A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, 19, 636–643.

Wolfe, D.A. (2010). Ranked set sampling. Wiley interdisciplinary reviews. *Computational Statistics*, 2, 460–466.

Yin, J., Hao, Y., Samawi, H. and Rochani, H. (2016). Rank-based kernel estimation of the area under the ROC curve. *Statistical Methodology*, 32, 91–106.

Appendix

Proof of Proposition 1. It is easy to show that

$$m^2n^2E(\hat{\theta}^2_{\mathrm{SRS}}) = E(A_1 + A_2 + A_3 + A_4), \tag{15}$$

where

$$\begin{aligned} E(A_1) &= E\left\{ \sum_{i\neq i'=1}^{m} \sum_{j\neq j'=1}^{n} \Phi\left(\frac{X_i - Y_j}{t}\right) \Phi\left(\frac{X_{i'} - Y_{j'}}{t}\right) \right\} \\ &= m(m-1)n(n-1)E^2\left\{ \Phi\left(\frac{X-Y}{t}\right) \right\}, \end{aligned} \tag{16}$$

$$\begin{aligned} E(A_2) &= E\left\{ \sum_{i=1}^{m} \sum_{j\neq j'=1}^{n} \Phi\left(\frac{X_i - Y_j}{t}\right) \Phi\left(\frac{X_i - Y_{j'}}{t}\right) \right\} \\ &= mn(n-1)EE^2\left\{ \Phi\left(\frac{X-Y}{t}\right) \Big| X \right\}, \end{aligned} \tag{17}$$

$$\begin{aligned} E(A_3) &= E\left\{ \sum_{j=1}^{n} \sum_{i\neq i'=1}^{m} \Phi\left(\frac{X_i - Y_j}{t}\right) \Phi\left(\frac{X_{i'} - Y_j}{t}\right) \right\} \\ &= nm(m-1)EE^2\left\{ \Phi\left(\frac{X-Y}{t}\right) \Big| Y \right\} \end{aligned} \tag{18}$$

and

$$E(A_4) = E\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi^2\left(\frac{X_i - Y_j}{t}\right)\right\} = mnE\left\{\Phi^2\left(\frac{X-Y}{t}\right)\right\}. \tag{19}$$

From (15)-(19) and the expectation of $\hat{\theta}_{\mathrm{SRS}}$, the proof of the first part is complete. Similarly,

$$m^2n^2E(\hat{\theta}^2_{\mathrm{RSS}}) = E(B_1 + B_2 + B_3), \tag{20}$$

where

$$\begin{aligned}
E(B_1) &= E\left\{\sum_{i\neq i'=1}^{m}\sum_{j\neq j'=1}^{n}\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Phi\left(\frac{X_{[i']}-Y_{[j']}}{t}\right)\right.\\
&\quad\left.+\sum_{j=1}^{n}\sum_{i\neq i'=1}^{m}\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Phi\left(\frac{X_{[i']}-Y_{[j]}}{t}\right)\right\}\\
&= E\left(\sum_{i\neq i'=1}^{m}\sum_{j\neq j'=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}E\left\{\Phi\left(\frac{X_{[i']}-Y_{[j']}}{t}\right)\Big|Y_{[j']}\right\}\right.\\
&\quad\left.+\sum_{j=1}^{n}\sum_{i\neq i'=1}^{m}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}E\left\{\Phi\left(\frac{X_{[i']}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right)\\
&= E\left(\left[\sum_{i=1}^{m}\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2 - \sum_{i=1}^{m}\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right.\\
&\quad\left.-\sum_{i=1}^{m}\sum_{j\neq j'=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j']}}{t}\right)\Big|Y_{[j']}\right\}\right)\\
&= E\left(\left[\sum_{i=1}^{m}\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2 - \sum_{i=1}^{m}\left[\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right.\right.\\
&\quad\left.\left.+\sum_{j\neq j'=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j']}}{t}\right)\Big|Y_{[j']}\right\}\right]\right)\\
&= E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2 - \sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right),
\end{aligned} \tag{21}$$

$$
\begin{aligned}
E(B_2) &= E\left\{\sum_{i=1}^{m}\sum_{j\neq j'=1}^{n}\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Phi\left(\frac{X_{[i]}-Y_{[j']}}{t}\right)\right\}\\
&= mE\left\{\sum_{j\neq j'=1}^{n}\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Phi\left(\frac{X-Y_{[j']}}{t}\right)\right\}\\
&= mE\left(\sum_{j\neq j'=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}E\left\{\Phi\left(\frac{X-Y_{[j']}}{t}\right)\Big|X\right\}\right)\\
&= mE\left(\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right]^2-\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right)\\
&= mE\left(n^2E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\}-\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right), \qquad (22)
\end{aligned}
$$

and

$$
E(B_3) = E\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi^2\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\right\} = mnE\left\{\Phi^2\left(\frac{X-Y}{t}\right)\right\}. \qquad (23)
$$

Now the second part follows from (20)-(23) and the expectation of $\hat{\theta}_{\mathrm{RSS}}$. ■

Proof of Proposition 2. Using equations (8) and (9), it can be shown

$$
m^2n^2\left[\mathrm{Var}(\hat{\theta}_{\mathrm{SRS}})-\mathrm{Var}(\hat{\theta}_{\mathrm{RSS}})\right] = \Delta_1+\Delta_2+\Delta_3, \qquad (24)
$$

where

$$
\begin{aligned}
\Delta_1 &= E\left(\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2-m\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right)\\
&= E\left(\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}-\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right), \qquad (25)
\end{aligned}
$$

$$\begin{aligned}\Delta_2 &= mn(n-1)EE^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\}\\ &\quad - mE\left(n^2E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\} - \sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right)\\ &= mE\left(\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\} - nE^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\}\right)\\ &= mE\left(\sum_{j=1}^{n}\left[E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\} - E\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|X\right\}\right]^2\right) \qquad (26)\end{aligned}$$

and

$$\begin{aligned}\Delta_3 &= m(m-1)n(n-1)E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\right\} + nm(m-1)EE^2\left\{\Phi\left(\frac{X-Y}{t}\right)\Big|Y\right\}\\ &\quad - m(m-1)E\left(\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right)\\ &= m(m-1)\left[n(n-1)E^2\left\{\Phi\left(\frac{X-Y}{t}\right)\right\}\right.\\ &\quad \left. + E\left(\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\} - \left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right)\right]\\ &= m(m-1)\left[\left(1-\frac{1}{n}\right)\left(\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\right\}\right)^2\right.\\ &\quad \left. - \sum_{j\neq j'=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\right\}E\left\{\Phi\left(\frac{X-Y_{[j']}}{t}\right)\right\}\right]\\ &= m(m-1)\left[\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\right\} - \frac{1}{n}\left(\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\right\}\right)^2\right]\\ &= m(m-1)\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right) - \Phi\left(\frac{X-Y}{t}\right)\right\}. \qquad (27)\end{aligned}$$

Clearly, $\Delta_i \geq 0\,(i = 1,2,3)$, as was to be shown. Proof of the next part is straightforward, and is omitted. ■

The next lemma paves the way for Proposition 3.

Lemma 1 *If* $\ell = \sum_{j=1}^{n} E\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}$ *and* $L = \sum_{j=1}^{n} E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}$, *then* $\mathrm{Var}(\ell) \leq \mathrm{Var}(L)$.

Proof of Lemma 1. Using conditional variance formula, we have

$$\begin{aligned}\mathrm{Var}(L) &= \sum_{j=1}^{n} \mathrm{Var}\left(E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right)\\ &\geq \sum_{j=1}^{n}\sum_{k=1}^{n} q_{jk}\mathrm{Var}\left(E\left\{\Phi\left(\frac{X-Y_{(k)}}{t}\right)\Big|Y_{(k)}\right\}\right)\\ &= \sum_{k=1}^{n} \mathrm{Var}\left(E\left\{\Phi\left(\frac{X-Y_{(k)}}{t}\right)\Big|Y_{(k)}\right\}\right) = \mathrm{Var}(\ell),\end{aligned}$$

as was asserted. ∎

Proof of Proposition 3. First, some necessary notions and results from matrix algebra are provided.

The L_1, L_∞ and L_2 norms for an $r \times c$ matrix $\mathbf{A} = [a_{ij}]$ are defined as

$$\|\mathbf{A}\|_1 = \max_{j=1,\ldots,c} \sum_{i=1}^{r} a_{ij},$$

$$\|\mathbf{A}\|_\infty = \max_{i=1,\ldots,r} \sum_{j=1}^{c} a_{ij}$$

and

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}'\mathbf{A})},$$

where $\lambda_{\max}(\mathbf{A}'\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}'\mathbf{A}$ matrix. If the product of matrices $\mathbf{A}$ and $\mathbf{B}$ is defined, then

$$\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2 \tag{28}$$

and

$$\|\mathbf{A}\|_2^2 \leq \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty. \tag{29}$$

See Datta (2010) for more details.

In view of (9), it suffices to show that

$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}\right]^2\right)\leq$$
$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right) \quad (30)$$

and

$$E\left(\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}\right)\leq E\left(\sum_{j=1}^{n}E^2\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|X\right\}\right). \quad (31)$$

We begin with proving the first inequality. Assume that $Z_{(i)}=\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}$ and $Z_{[i]}=\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}$. Then one can write

$$Z_{[i]}=\sum_{j=1}^{n}\sum_{k=1}^{m}p_{ik}E\left\{\Phi\left(\frac{X_{(k)}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}=\sum_{k=1}^{m}p_{ik}Z_{(k)}. \quad (32)$$

Let Ω_Y be the sample space on which Y is defined. If $\mathbf{P}=[p_{ir}]_{m\times m}$ and $\mathbf{Z}'=\left(Z_{(1)}(\vartheta),\ldots,Z_{(m)}(\vartheta)\right)$ given a fixed $\vartheta\in\Omega_Y$, then using (28), (29) and (32) it follows that

$$\begin{aligned}\sum_{i=1}^{m}Z_{[i]}^2(\vartheta)&=\sum_{i=1}^{m}\left(\sum_{k=1}^{m}p_{ik}Z_{(k)}(\vartheta)\right)^2=\|\mathbf{PZ}\|_2^2\leq\|\mathbf{P}\|_2^2\,\|\mathbf{Z}\|_2^2\\&\leq\|\mathbf{P}\|_1\,\|\mathbf{P}\|_\infty\sum_{i=1}^{m}Z_{(i)}^2(\vartheta)\\&=\sum_{i=1}^{m}Z_{(i)}^2(\vartheta).\end{aligned}$$

The last equality holds because $\sum_{i=1}^{m}p_{ik}=\sum_{k=1}^{m}p_{ik}=1$. Hence,

$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right)\leq$$
$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{[i]}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right).$$

Now, (30) is deduced if

$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}\right]^2\right)\leq$$
$$E\left(m^2\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2-\sum_{i=1}^{m}\left[\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}\right]^2\right) \quad (33)$$

For $i=1,\ldots,m$, suppose $\ell_{(i)}=\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{(j)}}{t}\right)\Big|Y_{(j)}\right\}$ and ℓ be as in Lemma 1. We note that $\ell_{(1)}<\cdots<\ell_{(m)}$ are order statistics from a sample of size m. Therefore,

$$\sum_{i=1}^{m}E(\ell_{(i)}^2)=\sum_{i=1}^{m}\int t^2 f_{\ell_{(i)}}(t)\,\mathrm{d}t=m\int t^2 f_{\ell}(t)\,\mathrm{d}t=mE(\ell^2), \quad (34)$$

where $f_{\ell_{(i)}}$ and f_{ℓ} denote the density function of $\ell_{(i)}$ and ℓ, respectively. Similarly, one can define $L_{(i)}=\sum_{j=1}^{n}E\left\{\Phi\left(\frac{X_{(i)}-Y_{[j]}}{t}\right)\Big|Y_{[j]}\right\}$, and conclude that

$$\sum_{i=1}^{m}E(L_{(i)}^2)=mE(L^2), \quad (35)$$

where L is as in Lemma 1. From (34) and (35), (33) reduces to $E(\ell^2)\leq E(L^2)$. This is equivalent to $\mathrm{Var}(\ell)\leq\mathrm{Var}(L)$ which holds thanks to Lemma 1.

Assume that $W_{(j)}=E\left\{\Phi\left(\frac{X-Y_{(j)}}{t}\right)\Big|X\right\}$ and $W_{[j]}=E\left\{\Phi\left(\frac{X-Y_{[j]}}{t}\right)\Big|X\right\}$. Then, it can be shown that $W_{[j]}=\sum_{k=1}^{n}q_{jk}W_{(j)}$. Let Ω_X be the sample space on which X is defined. If $\mathbf{Q}=[q_{js}]_{n\times n}$ and $\mathbf{W}^{\mathrm{T}}=\left(W_{(1)}(\eta),\ldots,W_{(n)}(\eta)\right)$ for each fixed $\eta\in\Omega_X$, then applying (28) and (29) using $\mathbf{Q}$ and $\mathbf{W}$ yields

$$\begin{aligned}\sum_{j=1}^{n}W_{[j]}^2(\eta)&=\sum_{j=1}^{n}\left(\sum_{k=1}^{m}q_{jk}W_{(k)}(\eta)\right)^2=\|\mathbf{QW}\|_2^2\leq\|\mathbf{Q}\|_2^2\,\|\mathbf{W}\|_2^2\\&\leq\|\mathbf{Q}\|_1\,\|\mathbf{Q}\|_\infty\sum_{i=1}^{m}W_{(i)}^2(\eta)\\&=\sum_{i=1}^{m}W_{(i)}^2(\eta).\end{aligned}$$

This completes the proof of (31). ∎

SORT 40 (2) July-December 2016, 267-302

A construction of continuous-time ARMA models by iterations of Ornstein-Uhlenbeck processes

Argimiro Arratia[1], Alejandra Cabaña[2] and Enrique M. Cabaña[3]

Abstract

We present a construction of a family of continuous-time ARMA processes based on p iterations of the linear operator that maps a Lévy process onto an Ornstein-Uhlenbeck process. The construction resembles the procedure to build an AR(p) from an AR(1). We show that this family is in fact a subfamily of the well-known CARMA(p,q) processes, with several interesting advantages, including a smaller number of parameters. The resulting processes are linear combinations of Ornstein-Uhlenbeck processes all driven by the same Lévy process. This provides a straightforward computation of covariances, a state-space model representation and methods for estimating parameters. Furthermore, the discrete and equally spaced sampling of the process turns to be an ARMA($p,p-1$) process. We propose methods for estimating the parameters of the iterated Ornstein-Uhlenbeck process when the noise is either driven by a Wiener or a more general Lévy process, and show simulations and applications to real data.

MSC: 60G10, 62M10, 62M99 60M99.

Keywords: Ornstein-Uhlenbeck process, Lévy process, Continuous ARMA, stationary process.

1. Introduction

The link between discrete time autoregressive moving average (ARMA) processes and stationary processes with continuous-time has been of interest for many years, see for instance, Doob (1944), Durbin (1961), Bergstrom (1984, 1996) and more recently Brockwell (2009), Thornton and Chambers (2013). Continuous time ARMA processes are better suited than their discrete counterparts for modelling irregularly spaced data, and when the white noise is driven by a non-Gaussian process it becomes a more realistic model in finance and other fields of application.

[1] Universitat Politècnica de Catalunya, Barcelona, Spain. argimiro@cs.upc.edu
Supported by Spain's MINECO project APCOM (TIN2014-57226-P) and Generalitat de Catalunya 2014SGR 890 (MACDA).

[2] Universitat Autònoma de Barcelona, Spain. acabana@mat.uab.cat
Supported by Spain's MINECO project MTM2015-69493-R.

[3] Universidad de la República, Montevideo, Uruguay. ecabana@ccee.edu.uy
Received: November 2015
Accepted: May 2016

A popular continuous-time representation of ARMA(p,q) process (known as CARMA(p,q)) can be obtained via a state-space representation of the formal equation

$$a(D)Y(t) = \sigma b(D)D\Lambda(t),$$

where $\sigma > 0$ is a scale parameter, D denotes differentiation with respect to t, Λ is a second-order Lévy process, $a(z) = z^p + a_1 z^{p-1} + \ldots + a_p$ is a polynomial of order p and $b(z) = b_0 + b_1 z + \ldots + b_q z^q$ a polynomial of order $q \leq p-1$ with coefficient $b_q \neq 0$ (see, e.g., Brockwell, 2004, 2009, Thornton and Chambers, 2013). The parameters of this model are estimated by adjusting first an ARMA(p,q), $q < p$ to regularly spaced data. Then obtain the parameters of the continuous version whose values at the observation times have the same distribution of the fitted ARMA. Hence, $p+q+1$ parameters have to be estimated.

We propose in this work a parsimonious model for continuous autoregression, with fewer parameters (as we shall see exactly p plus the variance). Our construction departs from the observation that a Ornstein-Uhlenbeck (OU) process can be thought of as continuous-time interpolation of an autoregressive process of order one (i.e. an AR(1)). This is shown in Section 2, where we also review some well known facts on Lévy processes, ARMA models and their representations. The model is obtained by a procedure that resembles the one that allows to build an AR(p) from an AR(1). Departing from this analogy, we define and analyse the result of iterating the application of the operator that maps a Wiener process onto an OU process. This operator is defined in Section 3 and denoted $\mathcal{OU}$, with subscripts denoting the parameters involved.

The p iterations of $\mathcal{OU}$, for each positive integer p, give rise to a new family of processes, the Ornstein-Uhlenbeck processes of order p, denoted OU(p). They can be used as models for either stationary continuous-time processes or the series obtained by observing these continuous processes at equally spaced instants. We show that an OU(p) process can be expressed as a linear combination of ordinary OU processes, or generalized OU processes, also defined in Section 3. This result resembles the aggregations of Gaussian (and non-Gaussian) processes studied with the idea of deconstructing a complicated economic model into simpler constituents. In the extensive literature on aggregations (or superpositions) of stochastic processes the aggregated processes are driven by independent Lévy processes (see, e.g., Granger and Morris, 1976, Granger, 1980, Barndorff-Nielsen, 2001, Eliazar and Klafter, 2009, among many others). A distinctive point of our construction is that the stochastic processes obtained by convolution of the OU operator result in a linear combination comprised of processes driven by the same Lévy process.

Another consequence of writing the OU(p) process as the aggregation of simpler ones is the derivation of a closed formula for its covariance. This has important practical implications since it allows to easily estimate the parameters of a OU(p) process by matching correlations (a procedure resembling the method of moments, to be described in Section 6.2), and by maximum likelihood.

In Section 4 we show how to write the discrete version of a OU(p) as a state-space model, and from this representation we show in Section 5 that for $p > 1$, a OU(p) behaves like an aggregation of AR processes (in the manner considered in Granger and Morris (1976)), that turns out to be an ARMA(p,q), with $q \leq p-1$. Consequently the OU(p) processes are a subfamily of the CARMA(p,q) processes. Notwithstanding this structural similarity, the family of discretized OU(p) processes is more parsimonious than the family of ARMA$(p,p-1)$ processes, and we shall see empirically that it is able to fit well the autocovariances for large lags. Hence, OU processes of higher order appear as a new continuous model, competitive in a discrete time setting with higher order autoregressive processes (AR or ARMA). The estimation of the parameters of OU(p) processes is attempted in Section 6. Simulations and applications to real data are provided in Section 6.5. Our concluding remarks are in Section 7.

2. Preliminaries

Let us recall that a Lévy process $\Lambda(t)$ is a càdlàg function, with independent and stationary increments, that vanishes in $t = 0$. As a consequence, $\Lambda(t)$ is, for each t, a random variable with an infinitely divisible law (Sato, 1999). A Wiener process W is a centred Gaussian process, with independent increments and variance $\mathrm{E}(W(t) - W(s))^2 = \sigma^2|t-s|$. Wiener processes are the only Lévy processes with almost surely continuous paths. For parameter $\lambda > 0$ the classical Ornstein-Uhlenbeck process is defined as $\int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-s)} dW(s)$ (Uhlenbeck and Ornstein, 1930).

Wiener process can be replaced by a second order Lévy process Λ to define a Lévy driven Ornstein-Uhlenbeck process as

$$x(t)(= x_{\lambda,\Lambda}(t)) := \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-s)} d\Lambda(s) \tag{1}$$

The previous equation can be formally written in differential form

$$dx(t) = -\lambda x(t)dt + d\Lambda(t) \tag{2}$$

We may think of x as the result of accumulating a random noise $d\Lambda$, with reversion to the mean (that we assume to be 0) of exponential decay with rate λ.

When the Ornstein-Uhlenbeck process x is sampled at equally spaced times $\{h\tau : h = 0,1,2,\ldots,n\}$, $\tau > 0$, the series $X_h = x(h\tau)$ obeys an autoregressive model of order 1 (i.e. an AR(1)), because $X_{h+1} = e^{-\lambda\tau} X_h + Z_{h+1}$, where $Z_{h+1} = \int_{h\tau}^{(h+1)\tau} e^{-\lambda((h+1)\tau - s)} d\Lambda(s)$, is the stochastic innovation.

Hence, we can consider the OU process as continuous-time interpolation of an AR(1) process. Notice that both models are stationary. This link between AR(1) and OU(1) suggests the definition of iterated OU processes introduced in Section 3.

An ARMA(p,q) or autoregressive moving average process of order (p,q) has the following form

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta_0 \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

where $\phi_1, \ldots, \phi_p$ are the autoregressive parameters, $\theta_0, \ldots, \theta_q$ are the moving average parameters, and the white-noise process ϵ_t has variance one. Denote by B the backshift operator that carries x_t into x_{t-1}. By considering the polynomials in the backshift operator,

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \text{ and } \theta(B) = \theta_0 + \theta_1 B + \cdots + \theta_q B^q$$

the ARMA(p,q) model can be written as

$$\phi(B) x_t = \theta(B) \epsilon_t \tag{3}$$

This compact expression comes in handy for analysing structural properties of time series. It also links to the representation of ARMA processes as a *state-space model*, useful for simplifying maximum likelihood estimation and forecasting. A state-space model has the general form

$$\boldsymbol{Y}_t = \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\eta}_t \tag{4}$$

$$x_t = \boldsymbol{K}^\top \boldsymbol{Y}_t + \boldsymbol{N}_t \tag{5}$$

where (4) is the state equation and (5) is the observation equation, with $\boldsymbol{Y}_t$ the m-dimensional state vector, $\boldsymbol{A}$ and $\boldsymbol{K}$ are $m \times m$ and $m \times k$ coefficient matrices, $\boldsymbol{K}^\top$ denotes the transpose of $\boldsymbol{K}$, $\boldsymbol{\eta}$ and $\boldsymbol{N}$ are m and k dimensional white noises. $\boldsymbol{N}$ would be present only if the process x_t is observed subject to additional noise (see Box, Jenkins, and Reinsel, 1994 for further details). We present in Section 4 a state-space model representation of our generalized OU process.

3. Ornstein-Uhlenbeck processes of order p

The AR(1) process $X_t = \phi X_{t-1} + \epsilon_t$, where ϵ_t, $t \in$Z, is a white noise, can be written as $(1 - \phi B)X_t = \epsilon_t$ using the back-shift operator B. Equivalently, X_t can be written as $X_t = \mathcal{MA}_{1/\rho}\epsilon_t$, where $\mathcal{MA}_{1/\rho}$ is the moving average that maps ϵ_t onto $\mathcal{MA}_{1/\rho}\epsilon_t, = \sum_{j=0}^{\infty} \frac{1}{\rho^j}\epsilon_{t-j}$, and ρ $(= 1/\phi)$ is the root of the characteristic polynomial $1 - \phi z$.

Moreover, the AR(p) process $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t$ (or $\phi(B)X_t = \epsilon_t$), where $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j = \prod_{j=1}^{p}(1 - z/\rho_j)$ has roots $\rho_j = \mathrm{e}^{\lambda_j}$, $j = 1, \ldots, p$, can be obtained by applying the composition of the moving averages $\mathcal{MA}_{1/\rho_j}$ to the noise, that is:

$$X_t = \prod_{j=1}^{p} \mathcal{MA}_{1/\rho_j} \epsilon_t$$

Now consider the operator $\mathcal{MA}_{\mathrm{e}^{-\lambda}}$ that maps ϵ_t onto

$$\mathcal{MA}_{\mathrm{e}^{-\lambda}} \epsilon_t = \sum_{l \le t, \text{integer}} \mathrm{e}^{-\lambda(t-l)} \epsilon_l$$

A continuous version of this operator is $\mathcal{OU}_\lambda$ that maps $y(t), t \in \mathbb{R}$ onto

$$\mathcal{OU}_\lambda y(t) = \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-s)} dy(s), \tag{6}$$

whenever the integral can be defined. The definition of $\mathcal{OU}_\lambda$ is extended to include complex processes, by replacing λ by $\kappa = \lambda + i\mu$, $\lambda > 0$, $\mu \in \mathbb{R}$ in (6). The set of complex numbers with positive real part is denoted by $\mathbf{C}^+$, and the conjugate of κ is denoted by $\bar{\kappa}$.

For $p \ge 1$ and parameters $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$, the previous argument suggests to define the following process obtained as repeated compositions of operators $\mathcal{OU}_{\kappa_j}$, $j = 1, \ldots, p$:

$$\mathcal{OU}_{\boldsymbol{\kappa}} y(t) := \mathcal{OU}_{\kappa_1} \mathcal{OU}_{\kappa_2} \cdots \mathcal{OU}_{\kappa_p} y(t) = \prod_{j=1}^{p} \mathcal{OU}_{\kappa_j} y(t) \tag{7}$$

This is called *Ornstein-Uhlenbeck process of order p with parameters* $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p) \in (\boldsymbol{C}^+)^p$. The composition $\prod_{j=1}^{p} \mathcal{OU}_{\kappa_j}$ is unambiguously defined because the application of $\mathcal{OU}_{\kappa_j}$ operators is commutative as shown in Theorem 1(i) below.

The particular case of interest where the underlying noise is a second order Lévy process Λ, namely,

$$\mathcal{OU}_{\boldsymbol{\kappa}} \Lambda(t) := \mathcal{OU}_{\kappa_1} \mathcal{OU}_{\kappa_2} \cdots \mathcal{OU}_{\kappa_p} \Lambda(t) = \prod_{j=1}^{p} \mathcal{OU}_{\kappa_j} \Lambda(t) \tag{8}$$

is called the *Lévy-driven Ornstein-Uhlenbeck process of order p with parameters* $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p) \in (\boldsymbol{C}^+)^p$.

For technical reasons, it is convenient to introduce the *Ornstein-Uhlenbeck operator* $\mathcal{OU}_\kappa^{(h)}$ *of degree h with parameter* κ that maps y onto

$$\mathcal{OU}_\kappa^{(h)} y(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} dy(s) \tag{9}$$

and Λ onto

$$\xi_\kappa^{(h)}(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} d\Lambda(s) \tag{10}$$

We call the process (10) *generalized Ornstein-Uhlenbeck process* of order 1 and degree h. For the remainder of the paper we restrict the underlying noise to a second order Lévy Λ, but note that the general properties of the $\mathcal{OU}_\kappa$ operator that we are going to show hold for any random function $y(t)$ for which the integral (6) is defined.

3.1. Properties of the operator $\mathcal{OU}_\kappa$

The following statements summarize some properties of products (compositions) of the operators defined by (7) and (9), and correspondingly, of the stationary centred processes $\xi_\kappa^{(h)}$, $h \geq 0$. In particular, the Ornstein-Uhlenbeck processes of order 1 and degree 0, $\xi_\kappa^{(0)} = \xi_\kappa$ are the ordinary Ornstein-Uhlenbeck processes (1).

Theorem 1

(i) When $\kappa_1 \neq \kappa_2$*, the product* $\mathcal{OU}_{\kappa_2}\mathcal{OU}_{\kappa_1}$ *can be computed as*

$$\frac{\kappa_1}{\kappa_1 - \kappa_2}\mathcal{OU}_{\kappa_1} + \frac{\kappa_2}{\kappa_2 - \kappa_1}\mathcal{OU}_{\kappa_2}$$

and is therefore commutative.

(ii) The composition $\prod_{j=1}^{p} \mathcal{OU}_{\kappa_j}$ *constructed with pairwise different* $\kappa_1, \ldots, \kappa_p$ *is equal to the linear combination*

$$\prod_{j=1}^{p} \mathcal{OU}_{\kappa_j} = \sum_{j=1}^{p} K_j(\kappa_1, \ldots, \kappa_p)\mathcal{OU}_{\kappa_j}, \tag{11}$$

with coefficients

$$K_j(\kappa_1, \ldots, \kappa_p) = \frac{1}{\prod_{\kappa_l \neq \kappa_j}(1 - \kappa_l/\kappa_j)}. \tag{12}$$

(iii) *For* $i = 1,2,\ldots$, $\mathcal{OU}_\kappa \mathcal{OU}_\kappa^{(i)} = \mathcal{OU}_\kappa^{(i)} - \kappa \mathcal{OU}_\kappa^{(i+1)}$.

(iv) *For any positive integer* p *the* p*-th power of the Ornstein-Uhlenbeck operator has the expansion*

$$\mathcal{OU}_\kappa^p = \sum_{j=0}^{p-1} \binom{p-1}{j} \mathcal{OU}_\kappa^{(j)}. \tag{13}$$

(v) *Let* $\kappa_1, \ldots, \kappa_q$ *be pairwise different complex numbers with positive real parts, and* $p_1, \ldots, p_q$ *positive integers, and let us denote by* $\boldsymbol{\kappa}$ *a complex vector in* $(\boldsymbol{C}^+)^p$ *with components* κ_h *repeated* p_h *times,* $p_h \geq 1$, $h = 1, \ldots, q$, $\sum_{h=1}^q p_h = p$. *Then, with* $K_h(\boldsymbol{\kappa})$ *defined by (12),*

$$\prod_{h=1}^q \mathcal{OU}_{\kappa_h}^{p_h} = \sum_{h=1}^q \frac{1}{\prod_{l \neq h}(1 - \kappa_l/\kappa_h)^{p_l}} \mathcal{OU}_{\kappa_h}^{p_h} = \sum_{h=1}^q K_h(\boldsymbol{\kappa}) \mathcal{OU}_{\kappa_h}^{p_h}.$$

An immediate consequence is that the operator $\mathcal{OU}_{\boldsymbol{\kappa}}$ with p-vector parameter $\boldsymbol{\kappa}$ can be written as a linear combination of p operators $\mathcal{OU}_\kappa$ or $\mathcal{OU}_\kappa^{(h)}$ for suitable scalar values κ and non-negative integer h. Therefore, the process $\mathcal{OU}_{\boldsymbol{\kappa}}\Lambda$ can be written as a linear combination of OU processes driven by the same Lévy process, as stated in the following Corollary.

Corollary 1

(i) *The process* $\mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda) = \prod_{h=1}^q \mathcal{OU}_{\kappa_h}^{p_h}(\Lambda)$ *can be expressed as the linear combination*

$$\mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda) = \sum_{h=1}^q K_h(\boldsymbol{\kappa}) \sum_{j=0}^{p_h-1} \binom{p_h-1}{j} \xi_{\kappa_h}^{(j)} \tag{14}$$

of the p *processes* $\{\xi_{\kappa_h}^{(j)} : h = 1, \ldots, q, j = 0 \ldots, p_h - 1\}$ *(see (10)).*

(ii) *Consequently,*

$$\mathcal{OU}_{\boldsymbol{\kappa}}\Lambda(t) = \sum_{h=1}^q K_h(\boldsymbol{\kappa}) \sum_{j=0}^{p_h-1} \binom{p_h-1}{j} \int_{-\infty}^t \mathrm{e}^{-\kappa_h(t-s)} \frac{(-\kappa_h(t-s))^j}{j!} d\Lambda(s)$$

Corollary 2 *For real* λ, μ, *with* $\lambda > 0$, *the product* $\mathcal{OU}_{\lambda+i\mu}\mathcal{OU}_{\lambda-i\mu}$ *is real, that is, applied to a real process produces a real image.*

The proofs of Theorem 1 and corollaries are in Appendix A.

3.2. Computing the covariances

The representation

$$x := \mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda) = \sum_{h=1}^{q} K_h(\kappa) \sum_{j=0}^{p_h-1} \binom{p_h-1}{j} \mathcal{OU}_{\kappa_h}^{(j)}(\Lambda)$$

of x as a linear combination of the processes $\xi_{\kappa_h}^{(i)} = \mathcal{OU}_{\kappa_h}^{(i)}(\Lambda)$ allows a direct computation of the covariances $\gamma(t) = \mathrm{E}x(t)\bar{x}(0)$ through a closed formula, in terms of the covariances $\gamma_{\kappa_1,\kappa_2}^{(i_1,i_2)}(t) = \mathrm{E}\xi_{\kappa_1}^{(i_1)}(t)\bar{\xi}_{\kappa_2}^{(i_2)}(0)$:

$$\gamma(t) = \sum_{h'=1}^{q} \sum_{i'=0}^{p_{h'}-1} \sum_{h''=1}^{q} \sum_{i''=0}^{p_{h''}-1} K_{h'}(\boldsymbol{\kappa}) \bar{K}_{h''}(\boldsymbol{\kappa}) \binom{p_{h'}-1}{i'} \binom{p_{h''}-1}{i''} \gamma_{\kappa_{h'},\kappa_{h''}}^{(i',i'')}(t) \tag{15}$$

with $v^2 = \mathrm{Var}\Lambda(1)$,

$$\begin{aligned}
\gamma_{\kappa_1,\kappa_2}^{(i_1,i_2)}(t) &= v^2(-\kappa_1)^{i_1}(-\bar{\kappa}_2)^{i_2} \int_{-\infty}^{0} \mathrm{e}^{-\kappa_1(t-s)} \frac{(t-s)^{i_1}}{i_1!} \mathrm{e}^{-\bar{\kappa}_2(-s)} \frac{(-s)^{i_2}}{i_2!} ds \\
&= v^2(-\kappa_1)^{i_1}(-\bar{\kappa}_2)^{i_2} \mathrm{e}^{-\kappa_1 t} \sum_{j=0}^{i_1} \binom{i_1}{j} \frac{t^j}{i_1! i_2!} \int_{-\infty}^{0} \mathrm{e}^{(\kappa_1+\bar{\kappa}_2)s} (-s)^{i_1+i_2-j} ds \\
&= \frac{v^2(-\kappa_1)^{i_1}(-\bar{\kappa}_2)^{i_2} \mathrm{e}^{-\kappa_1 t}}{i_2!} \sum_{j=0}^{i_1} \frac{t^j (i_1+i_2-j)!}{j!(i_1-j)!(\kappa_1+\bar{\kappa}_2)^{(i_1+i_2-j+1)}}
\end{aligned} \tag{16}$$

A real expression for the covariance when the imaginary parameters appear as conjugate pairs can be obtained but it is much more involved than this one.

4. The OU(p) process as a state-space model

Theorem 1 and its corollaries lead to express the OU(p) process by means of linear state-space models. The state-space modelling provides a unified methodology for the analysis of time series (see Durbin and Koopman, 2001).

In the simplest case, where the elements of $\boldsymbol{\kappa}$ are all different, the process $x(t) = \mathcal{OU}_{\boldsymbol{\kappa}}\Lambda(t)$ is a linear combination of the state vector $\boldsymbol{\xi}_{\boldsymbol{\kappa}}(t) = (\xi_{\kappa_1}(t), \xi_{\kappa_2}(t), \ldots, \xi_{\kappa_p}(t))^{\intercal}$, where $\xi_{\kappa_j} = \mathcal{OU}_{\kappa_j}(\Lambda)$.

More precisely, the vectorial process

$$\boldsymbol{\xi}_{\boldsymbol{\kappa}}(t) = (\xi_{\kappa_1}(t), \xi_{\kappa_2}(t), \ldots, \xi_{\kappa_p}(t))^{\intercal},\ \xi_{\kappa_j} = \mathcal{OU}_{\kappa_j}(\Lambda)$$

and $x(t) = \mathcal{OU}_{\boldsymbol{\kappa}}\Lambda(t)$ satisfy the linear equations

$$\boldsymbol{\xi}_{\boldsymbol{\kappa}}(t) = \operatorname{diag}(\mathrm{e}^{-\kappa_1\tau}, \mathrm{e}^{-\kappa_2\tau}, \ldots, \mathrm{e}^{-\kappa_p\tau})\boldsymbol{\xi}_{\boldsymbol{\kappa}}(t-\tau) + \boldsymbol{\eta}_{\boldsymbol{\kappa},\boldsymbol{\tau}}(t) \tag{17}$$

$$\text{and} \qquad x(t) = \boldsymbol{K}^{\intercal}(\boldsymbol{\kappa})\boldsymbol{\xi}(t), \tag{18}$$

$$\boldsymbol{\eta}_{\boldsymbol{\kappa},\boldsymbol{\tau}}(t) = (\eta_{\kappa_1,\tau}(t), \eta_{\kappa_2,\tau}(t), \ldots, \eta_{\kappa_p,\tau}(t))^{\intercal},\ \eta_{\kappa_j,\tau}(t) = \int_{t-\tau}^{t} \mathrm{e}^{-\kappa_j(t-s)} d\Lambda(s),$$

$$\operatorname{Var}(\boldsymbol{\eta}_{\boldsymbol{\kappa},\boldsymbol{\tau}}(t)) = v^2((v_{j,l})),\ v_{j,l} = \int_{t-\tau}^{t} \mathrm{e}^{-(\kappa_j+\bar{\kappa}_l)(t-s)} ds = \frac{1-\mathrm{e}^{-(\kappa_j+\bar{\kappa}_l)\tau}}{\kappa_j+\bar{\kappa}_l} \tag{19}$$

and the coefficients from (12), $\boldsymbol{K}^{\intercal}(\boldsymbol{\kappa}) = (K_1(\boldsymbol{\kappa}), K_2(\boldsymbol{\kappa}), \ldots, K_p(\boldsymbol{\kappa}))$.

The initial value $\boldsymbol{\xi}(0)$ is estimated by means of its conditional expectation $\hat{\boldsymbol{\xi}}(0) = \mathrm{E}(\xi(0)|x(0)) = \frac{\boldsymbol{K}^{\intercal}(\boldsymbol{\kappa})Vx(0)}{\boldsymbol{K}^{\intercal}(\boldsymbol{\kappa})V\overline{\boldsymbol{K}}}$, with $V = \operatorname{Var}(\boldsymbol{\xi}(0)) = \left(\left(\frac{1}{\kappa_j+\bar{\kappa}_l}\right)\right)$.

An application of Kalman filter to this state-space model leads to compute the likelihood of $\boldsymbol{x} = (x(0), x(\tau), \ldots, x(n\tau))$. Some Kalman filter programs included in software packages require the processes in the state-space to be real. That condition is not fulfilled by the model described by equations (17) and (18). An equivalent description by means of real processes can be obtained by ordering the parameters $\boldsymbol{\kappa}$ with the imaginary components paired with their conjugates in such a way that $\kappa_{2h} = \bar{\kappa}_{2h-1}$, $h = 1, 2, \ldots, c$ and the imaginary component $\Im(\kappa_j) = 0$ if and only if $2c < j \leq p$.

Then the matrix $\boldsymbol{M} = ((M_{j,k}))$ with all elements equal to zero except $M_{2h-1,2h-1} = M_{2h-1,2h} = 1$, $-M_{2h,2h-1} = M_{2h,2h} = i$, $h = 1, 2, \ldots, c$ and $M_{j,j} = 1$, $2c < j \leq p$, induces the linear transformation $\boldsymbol{\xi} \mapsto \boldsymbol{M\xi}$ that leads to the new state-space description

$$\boldsymbol{M\xi}(t) = \boldsymbol{M}\operatorname{diag}(\mathrm{e}^{-\kappa_1\tau}, \mathrm{e}^{-\kappa_2\tau}, \ldots, \mathrm{e}^{-\kappa_p\tau})\boldsymbol{M}^{-1}\boldsymbol{M\xi}(t-\tau) + \boldsymbol{M\eta}(t), \tag{20}$$

$$x(t) = \boldsymbol{K}^{\intercal}\boldsymbol{M}^{-1}M\boldsymbol{\xi}(t), \tag{21}$$

where the processes $\boldsymbol{M\xi}$ are real.

Observe that there is no loss of generality in choosing the spacing τ between observations as unity for the derivation of the state-space equations. Hence, we set $\tau = 1$ in the sequel and, in addition, τ will be omitted from the notation.

When $\kappa_1, \ldots, \kappa_q$ are all different, $p_1, \ldots, p_q$ are positive integers, $\sum_{h=1}^{q} p_h = p$ and $\boldsymbol{\kappa}$ is a p-vector with p_h repeated components equal to κ_h, the OU(p) process $x(t) = \mathcal{OU}_{\boldsymbol{\kappa}}\Lambda(t)$ is a linear function of the state-space vector

$$\left(\xi_{\kappa_1}^{(0)}, \xi_{\kappa_1}^{(1)}, \ldots, \xi_{\kappa_1}^{(p_1-1)}, \ldots, \xi_{\kappa_q}^{(0)}, \xi_{\kappa_q}^{(1)}, \ldots, \xi_{\kappa_q}^{(p_q-1)}\right)$$

where the components are given by (10), and the transition equation is no longer expressed by a diagonal matrix. In this case the state-space model has the following form

$$\begin{aligned} \boldsymbol{\xi}(t) &= \boldsymbol{A}\boldsymbol{\xi}(t-1) + \boldsymbol{\eta}(t) \\ x(t) &= \boldsymbol{K}^{\intercal}\boldsymbol{\xi}(t) \end{aligned} \tag{22}$$

We leave the technical details of this derivation to Appendix B. The terms $\boldsymbol{\xi}(t)$, $\boldsymbol{A}$, $\boldsymbol{\eta}(t)$ and $\boldsymbol{K}$ are precisely defined in (36). The real version of (22), when the process $\boldsymbol{\xi}$ has imaginary components is obtained by multiplying both equations by a block-diagonal matrix $\boldsymbol{C}$ (which is defined precisely in the Appendix), giving us the real state-space model

$$\boldsymbol{C}\boldsymbol{\xi}(t) = (\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^{-1})(\boldsymbol{C}\boldsymbol{\xi}(t-1)) + \boldsymbol{C}\boldsymbol{\eta}(t), \tag{23}$$

$$x(t) = (\boldsymbol{K}^{\intercal}C^{-1})(C\boldsymbol{\xi}(t)). \tag{24}$$

5. The OU(p) as an ARMA(p, $p-1$)

The studies of properties of linear transformations and aggregations of similar processes have produced a great amount of work stemming from the seminal paper by Granger and Morris (1976) on the invariance of MA and ARMA processes under these operations. These results and extensions to vector autoregressive moving average (VARMA) processes are compiled in the textbook by Lütkepohl (2005).

The description of the OU(p) process $x = \mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda)$ with parameters $\boldsymbol{\kappa}$ as a linear state-space model, given in the previous section, will allow us to show that the series $x(0)$, $x(1)$, …, $x(n)$ satisfies an ARMA(p,q) model with q smaller than p. We refer the reader to (Lütkepohl, 2005, Ch. 11) for a presentation on VARMA processes and, in particular, to the following result on the invariance property of VARMA processes under linear transformations, which we quote with a minor change of notation:

Theorem 2 (Lütkepohl, 2005, Cor. 11.1.2) *Let $y(t)$ be a d-dimensional, stable, invertible VARMA($\tilde{p}$,$\tilde{q}$) process and let $\boldsymbol{F}$ be an $(m \times d)$ matrix of rank m. Then the process $z_t = \boldsymbol{F}y_t$ has a VARMA($\check{p},\check{q}$) representation with $\check{p} \leq (d-m+1)\tilde{p}$ and $\check{q} \leq (d-m)\tilde{p}+\tilde{q}$.* ■

Equation (23) shows that $C\boldsymbol{\xi}(t)$ is a p-dimensional autoregressive vector (a p-dimensional VARMA(1,0) process) and Equation (24) expresses $x(t)$ as a linear transformation of $C\boldsymbol{\xi}(t)$ by the $(1 \times p)$ matrix $\boldsymbol{F} = \boldsymbol{K}^\intercal \boldsymbol{C}^{-1}$. Using Theorem 2 (with $d = p$, $\tilde{p} = 1$, $\tilde{q} = 0$, $m = 1$) we conclude that $(x(t) : t = 0, 1, \ldots, n)$ is an ARMA($\check{p}$,$\check{q}$) process with $\check{p} \leq p$ and $\check{q} \leq p-1$:

$$x(h) = \sum_{j=1}^{p} \phi_j x(h-j) + \sum_{l=0}^{p-1} \theta_l \epsilon_{h-l} \tag{25}$$

where ϵ is a Gaussian white noise with variance 1 and the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\intercal$, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})^\intercal$ of the ARMA process are functions of the parameters $\boldsymbol{\kappa}$ of the OU process. When the noise is any other second order Lévy process the corresponding OU(p) process has the same covariances as the process (25).

By using the backshift operator B, and the polynomials $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$, $\theta(z) = \sum_{l=0}^{p-1} \theta_l z^l$, (25) is written as

$$\phi(B)x = \theta(B)\epsilon. \tag{26}$$

5.1. Identifying the ARMA(p, $p-1$) from a given OU(p) process

We proceed now to identify the coefficients $\boldsymbol{\phi} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \mathbb{R}^{p-1}$ of the ARMA$(p, p-1)$ model that has the same autocovariances as $x = \mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda)$.

Case 1. Consider first that all components of $\boldsymbol{\kappa}$ are pairwise different, and hence $x(t)$ $= \sum_{j=1}^{p} K_j \xi_{\kappa_j}(t)$ is a linear combination of the OU(1) processes

$$\xi_{\kappa_j}(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa_j(t-s)} d\Lambda(s) = \mathrm{e}^{-\kappa_j} \xi_{\kappa_j}(t-1) + \int_{t-1}^{t} \mathrm{e}^{-\kappa_j(t-s)} d\Lambda(s)$$

with innovations $\boldsymbol{\eta}_{\boldsymbol{\kappa}}$ with components $\eta_{\kappa_j}(t) = \int_{t-1}^{t} \mathrm{e}^{-\kappa_j(t-s)} d\Lambda(s)$.

For each j, the series $\xi_{\kappa_j} = (\xi_{\kappa_j}(h))_{h \in \boldsymbol{Z}}$ satisfies the AR(1) model

$$(1 - \mathrm{e}^{-\kappa_j} B)\xi_{\kappa_j} = \eta_{\kappa_j}$$

(see (17)), and from (18) the series $x = (x(h))_{h\in\mathbf{Z}}$ follows the ARMA model

$$\prod_{j=1}^{p}(1-\mathrm{e}^{-\kappa_j}B)x = \sum_{j=1}^{p} K_j(\boldsymbol{\kappa})\prod_{l\neq j}(1-\mathrm{e}^{-\kappa_l}B)\eta_{\kappa_j}.$$

The sum of moving averages in the right-hand term is distributed as the moving average

$$\zeta = \sum_{h=0}^{p-1}\theta_h B^h\epsilon$$

where ϵ is a white noise with variance one and the coefficients θ_h are suitably chosen. It is readily verified that the autocovariances $c_l = \mathrm{E}\zeta(h)\bar{\zeta}(h-l)$ of this MA are the coefficients in the sum of powers of z

$$\left(\sum_{h=0}^{p-1}\theta_h z^h\right)\left(\sum_{k=0}^{p-1}\bar{\theta}_k z^{-h}\right) = \sum_{l=-p+1}^{p-1} c_l z^l. \tag{27}$$

A similar formula that takes into account the correlations (19) between the noises η_{κ_k} indicates that the same autocovariances are given by the identity

$$J(z) := \sum_{j=1}^{p}\sum_{l=1}^{p} K_j\bar{K}_l G_j(z)\bar{G}_l(1/z)v_{j,l} = \sum_{l=-p+1}^{p-1} c_l z^l \tag{28}$$

where $G_j(z) = \prod_{l\neq j}(1-\mathrm{e}^{-\kappa_l}z) = \sum_{l=0}^{p-1} g_{j,l}z^l$.

The coefficients $g_{j,l}$ and the function J are completely determined from the parameters of the OU process. In order to express the parameters of the ARMA$(p,p-1)$ process in terms of $\boldsymbol{\kappa}$ and $v^2 = \mathrm{Var}\Lambda(1)$ it remains to obtain the coefficients θ_h in the factorization (27). The roots ρ_j $(j = 1,2,\ldots,p-1)$ of

$$\theta(z) = \sum_{h=0}^{p-1}\theta_h z^h = \theta_0\prod_{j=1}^{p-1}(1-z/\rho_j) \tag{29}$$

are obtained by choosing the roots of the polynomial $z^{p-1}\theta(z)\bar{\theta}(1/z) = z^{p-1}J(z)$ with modules greater than one (the remaining roots are their inverses). Then all θ_h are written in terms of the ρ_h and the size factor θ_0 by applying (29). The value of θ_0 follows by using an additional equation, namely, the equality of the terms of degree zero in $J(z)$ and $\theta(z)\bar{\theta}(1/z)$, thus obtaining

$$\sum_{l=0}^{p-1}|\theta_l|^2=\sum_{j=1}^{p}\sum_{l=1}^{p}K_j\bar{K}_l v_{j,l}\sum_{h=0}^{p-1}g_{j,h}\bar{g}_{l,h}.$$

The general result, for arbitrary $\boldsymbol{\kappa}$ is much more involved and its derivation is deferred to Appendix C.

6. Estimation of the parameters of OU(p)

6.1. Reparameterization by means of real parameters

Our purpose is to insert the expression (15) for the covariance $\gamma(t)$ of the process $x(t)=\mathcal{OU}_{\boldsymbol{\kappa}}\Lambda(t)$ in a numeric optimization procedure in order to compute the maximum likelihood estimates of the parameters. Although $\gamma(t)$ depends continuously on $\boldsymbol{\kappa}$, the same does not happen with each term in the expression (15), because of the lack of boundedness of the coefficients of the linear combination when two different values of the components of $\boldsymbol{\kappa}$ approach each other. Since we wish to consider real processes x and the process itself and its covariance $\gamma(t)$ depend only of the unordered set of the components of $\boldsymbol{\kappa}$, we shall reparameterize the process. For the sake of simplicity, but without losing generality, consider the case where the components in $\boldsymbol{\kappa}$ are pairwise different. Let $K_{j,i}=\frac{1}{(-\kappa_j)^i\prod_{l\neq j}(1-\kappa_l/\kappa_j)}$ (in particular, $K_{j,0}$ is the same as K_j). Then the processes $x_i(t)=\sum_{j=1}^{p}K_{j,i}\xi_j(t)$ and the coefficients $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$ of the polynomial

$$g(z)=\prod_{j=1}^{p}(1+\kappa_j z)=1-\sum_{j=1}^{p}\beta_j z^j. \tag{30}$$

satisfy

$$\sum_{i=1}^{p}\beta_i x_i(t)=x(t).$$

The resulting process is real, because of Corollary 2. This works likewise for the general case of $\boldsymbol{\kappa}$ with some repetitions. Therefore the new parameter $\boldsymbol{\beta}$ shall be adopted.

6.2. Matching correlations estimation (MCE)

From the closed formula (15) for the covariance γ and the relationship (30) between $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$, we have a mapping $(\boldsymbol{\beta},v^2)\mapsto\gamma(t)$, for each t. Since

$$\boldsymbol{\rho}^{(T)}:=(\rho(1),\rho(2),\ldots,\rho(T))^{\intercal}=(\gamma(1),\gamma(2),\ldots,\gamma(T))^{\intercal}/\gamma(0)$$

does not depend on v^2, these equations determine a map $\mathcal{C} : (\boldsymbol{\beta}, T) \mapsto \boldsymbol{\rho}^{(T)} = \mathcal{C}(\boldsymbol{\beta}, T)$ for each T. After choosing a value of T and obtaining an estimate $\boldsymbol{\rho}_e^{(T)}$ of $\boldsymbol{\rho}^{(T)}$ based on the empirical covariances of x, we propose as a first estimate of $\boldsymbol{\beta}$, the vector $\breve{\boldsymbol{\beta}}_T$ such that all the components of the corresponding $\boldsymbol{\kappa}$ have positive real parts, and such that the Euclidean norm $\|\boldsymbol{\rho}_e^{(T)} - \mathcal{C}(\breve{\boldsymbol{\beta}}_T, T)\|$ reaches its minimum. The procedure resembles the estimation by the *method of moments*. The components of $\boldsymbol{\rho}_e^{(T)}$ for the series $(x_j)_{j=1,2,\ldots,n}$ are computed as

$$\rho_{e,h} = \gamma_{e,h}/\gamma_{e,0}, \qquad \gamma_{e,h} = \frac{1}{n}\sum_{j=1}^{n-h} x_j x_{j+h}.$$

6.3. Maximum likelihood estimation (MLE) in the Gaussian case

In this case $x(t) = \mathcal{OU}_{\boldsymbol{\kappa}} \sigma W(t)$, where $W(t)$ is standard Wiener process. Assume that $x(t)$ is observed at times $0, \tau, 2\tau, \ldots, n\tau$. By choosing τ the time unit of measure, as in Section 4, we assume without loss of generality that our observations are $\boldsymbol{x} = (x(0), x(1), \ldots, x(n))^\intercal$.

The likelihood L of the vector $\boldsymbol{x}$ is given by

$$\log L(\boldsymbol{x}; \boldsymbol{\beta}, \sigma^2) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log(\det(\boldsymbol{\Gamma}(\boldsymbol{\beta}, \sigma^2)) - \tfrac{1}{2}\boldsymbol{x}^\intercal(\boldsymbol{\Gamma}(\boldsymbol{\beta}, \sigma^2))^{-1}\boldsymbol{x}$$

where $\boldsymbol{\Gamma}$ has components $\Gamma_{h,j} = \gamma(|h-j|)$ $(h, j = 0, 1, \ldots, n)$. The Kalman filter associated to the dinamical state-space model in Section 4 provides an efficient alternative to compute the likelihood.

From these elements, a numerical optimization leads to obtain the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\hat{\sigma}^2$ of σ^2. If required, the estimations $\hat{\boldsymbol{\kappa}}$ follow by solving the analogue of the polynomial equation (30) written in terms of the estimators:

$$\prod_{j=1}^{p}(1 + \hat{\kappa}_j z) = 1 - \sum_{j=1}^{p}\hat{\beta}_j z^j.$$

The optimization for large n and the solution of the algebraic equation for large p require a considerable computation effort, but there are efficient programs to perform both operations, such as `optim` and `polyroot` in `R` (R Core Team, 2015). An alternative when the observed process is not assumed to be centred, is to maximize the log-likelihood of $\Delta\boldsymbol{x} = (x(1) - x(0), x(2) - x(1), \ldots, x(n) - x(n-1))$ given by

$$\log L(\boldsymbol{x}; \boldsymbol{\beta}, \sigma^2) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log(\det(\boldsymbol{V}(\boldsymbol{\beta}, \sigma^2)) - \tfrac{1}{2}\Delta\boldsymbol{x}^\intercal(\boldsymbol{V}(\boldsymbol{\beta}, \sigma^2))^{-1}\Delta\boldsymbol{x}$$

with $\boldsymbol{V}(\boldsymbol{\beta},\sigma^2)$ equal to the $n\times n$ matrix with components

$$V_{h,j} = 2\gamma(|h-j|) - \gamma(|h-j|+1) - \gamma(|h-j|-1)$$

that reduce to $2(\gamma(0)-\gamma(1))$ at the diagonal $h=j$.

The optimization procedures require an initial guess about the value of the parameter to be estimated. The estimators obtained by matching correlations described in the previous section can be used for that purpose.

6.4. The Gaussian case: examples

When Λ is a Wiener process W, the OU process of order p belongs to a subclass with $p+1$ parameters of the classical family of the $2p$-parameters Gaussian ARMA$(p,p-1)$

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta_0\epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_{p-1}\epsilon_{t-p+1}$$

where $\phi_1,\ldots,\phi_p$ and $\theta_0,\ldots,\theta_q$ are parameters and ϵ_t is a Gaussian noise with variance 1. The parameters $\boldsymbol{\kappa},\sigma^2$ determine the Gaussian likelihood of $\mathcal{OU}_{\boldsymbol{\kappa}}\sigma W$, and are estimated by the values $\hat{\boldsymbol{\kappa}}$ and $\hat{\sigma}^2$ that maximize that likelihood.

We have observed in several examples that the covariances of the process with the maximum likelihood estimators as parameters, follow closely the empirical covariances of the series. We have simulated the sample paths for the Wiener-driven OU(p) for different values of the parameters.

In the examples below we present simulated series $x(j), j=0,1,2,\ldots,n$ obtained from an OU process x for $n=300$ and three different values of the parameters and computed the MC and ML estimators $\check{\boldsymbol{\beta}}_T$, and $\hat{\boldsymbol{\beta}}$. The value of T for the MC estimation has been arbitrarily set equal to the integral part of $0.9\cdot n$, but the graphs of $\check{\boldsymbol{\beta}}_T$ for several values of T show in each case that after T exceeds a moderate threshold, the estimates remain practically constant. One of such graphs is included below (see Figure 2). It is of interest to perform further comparisons of these two methodologies for parameter estimation. A recent antecedent of this kind of comparisons and its importance can be found in Nieto, Orbe and Zarraga (2014).

The simulations show that the correlations of the series with the estimated parameters are fairly adapted to each other and to the empirical covariances. The departure from the theoretical covariances of x can be ascribed to the simulation intrinsic randomness.

Our first two examples describe OU(3) processes with arbitrarily (and randomly) chosen parameters and the third one imitates the behaviour of Series A that appears in Section 6.5.

Example 1. A series $(x_h)_{h=0,1,\dots,n}$ of $n = 300$ observations of the $\mathrm{OU}_{\boldsymbol{\kappa}}$ process x of order $p = 3$, $\boldsymbol{\kappa} = (0.9, 0.2+0.4i, 0.2-0.4i)$ and $\sigma^2 = 1$ was simulated, and the parameters $\boldsymbol{\beta} = (-1.30, -0.56, -0.18)$ and $\sigma^2 = 1$ were estimated by means of matching correlations:

$$\check{\boldsymbol{\beta}}_T = (-1.9245, -0.6678, -0.3221),$$

with $T = 270$; and maximum likelihood:

$$\hat{\boldsymbol{\beta}} = (-1.3546, -0.6707, -0.2355)$$

and $\hat{\sigma}^2 = 0.8958$. The corresponding estimators for $\boldsymbol{\kappa}$ are $\check{\boldsymbol{\kappa}} = (1.6368, 0.1439 + 0.4196i, 0.14389 - 0.4196i)$ and $\hat{\boldsymbol{\kappa}} = (0.9001, 0.2273 + 0.4582i, 0.2273 - 0.4582i)$.

The following table summarizes the different estimations of this OU(3) process.

original $\boldsymbol{\beta}$	-1.30	-0.56	-0.18	$\sigma^2 = 1$
original $\boldsymbol{\kappa}$	0.9	$0.2+0.4i$	$0.2-0.4i$	$\sigma^2 = 1$
MCE $\check{\boldsymbol{\beta}}_T$	-1.9245	-0.6678	-0.3221	
$\check{\boldsymbol{\kappa}}$	1.6368	$0.1439+0.4196i$	$0.14389-0.4196i$	
MLE $\hat{\boldsymbol{\beta}}$	-1.3546	-0.6707	-0.2355	$\hat{\sigma}^2 =$
$\hat{\boldsymbol{\kappa}}$	0.9001	$0.2273+0.4582i$	$0.2273-0.4582i$	0.8958

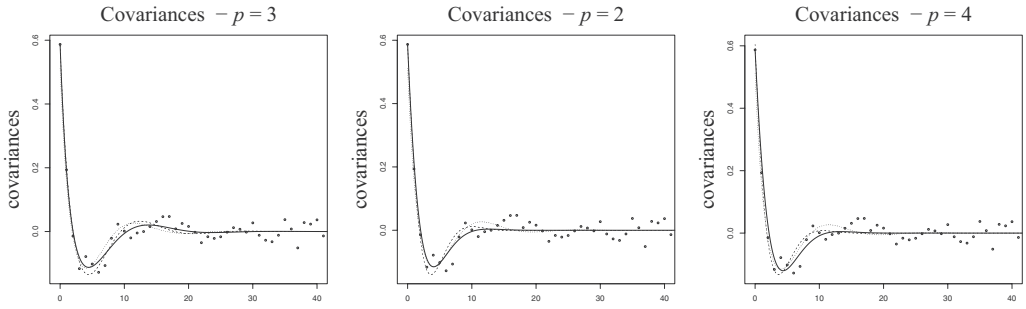


Figure 1: *Empirical covariances (∘) and covariances of the MC (—) and ML (- - -) fitted OU models, for $p = 3$, 2 and 4 corresponding to Example 1. The covariances of $\mathrm{OU}_{\boldsymbol{\kappa}}$ are indicated with a dotted line.*

Figure 1 describes the theoretical, empirical and estimated covariances of x under the assumption $p = 3$, that is, the actual order of x. The results obtained when the estimation is performed for $p = 2$ and $p = 4$ are also shown. Figure 2 shows that the MC estimates of $\boldsymbol{\beta}$ become stable for T moderately large, and close to the already indicated estimations for $T = 270$ (the horizontal lines).

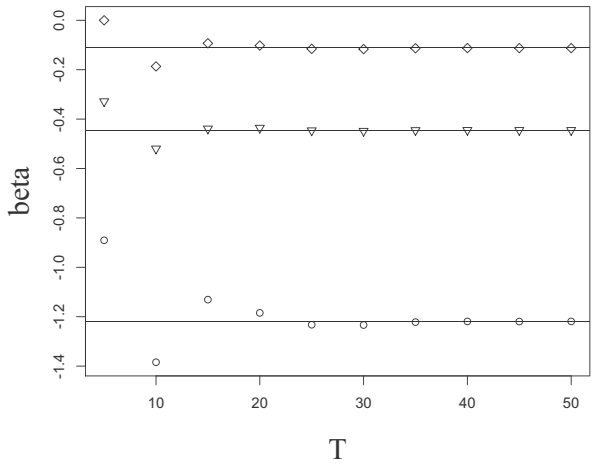


Figure 2: *The MC estimations $\check{\beta}_1(\circ)$, $\check{\beta}_2(\triangledown)$ and $\check{\beta}_3(\diamond)$ for different values of T, corresponding to Example 1. The horizontal lines indicate the estimations for $T = 270$.*

The coefficients ϕ_1, ϕ_2, ϕ_3 of the ARMA(3,2) model (26) satisfied by the series $(x(h))_{h=0,1,\ldots,300}$ are obtained by computing the product $\prod_{j=1}^{3}(1-\mathrm{e}^{-\kappa_j}B) = 1-\phi_1 B - \phi_2 B^2 - \phi_3 B^3 = 1 - 1.9148B + 1.2835B^2 - 0.2725B^3$.

As for the coefficients $\theta_0, \theta_1, \theta_2$, the first step is to compute the function

$$J(z) = 0.2995z^{-2} - 1.1943z^{-1} + 1.7904 - 1.1943z + 0.2995z^2,$$

then obtain the roots $\rho_1 = 1.1443 - 0.1944i$, $\rho_2 = 1.1443 + 0.1944i$, $\rho_3 = 0.8494 - 0.1443i$, $\rho_4 = 0.8494 + 0.1443i$ of the equation $z^2 J(z) = 0$, ordered by decreasing moduli, discard the last two, and write the function $\theta(z) = \theta_0 + \theta_1 z + \theta_2 z^2$ defined in (29):

$$\theta_0 \prod_{j=1}^{2}(1 - B/\rho_j) = \theta_0(1 - 1.6988z + 0.7423z^2).$$

Solve $\theta_0^2(1 + (-1.6988)^2 + 0.74229^2) = 1.7904$ to have $\theta_0 = 0.6352$, and hence $\theta(B) = 0.6352 - 1.0791B + 0.4715B^2$.

Example 2. The process $x = \mathrm{OU}_{(0.04,0.21,1.87)}$ is analysed as in Example 1. The resulting estimators are $\check{\boldsymbol{\beta}}_T = (-2.0611, -0.7459, -0.0553)$, $T = 270$, $\check{\boldsymbol{\kappa}} = (1.6224, 0.3378, 0.1009)$, $\hat{\boldsymbol{\beta}} = (-1.8253, -0.7340, -0.0680)$, $\hat{\sigma}^2 = 0.7842$, $\hat{\boldsymbol{\kappa}} = (1.3015, 0.3897, 0.1342)$.

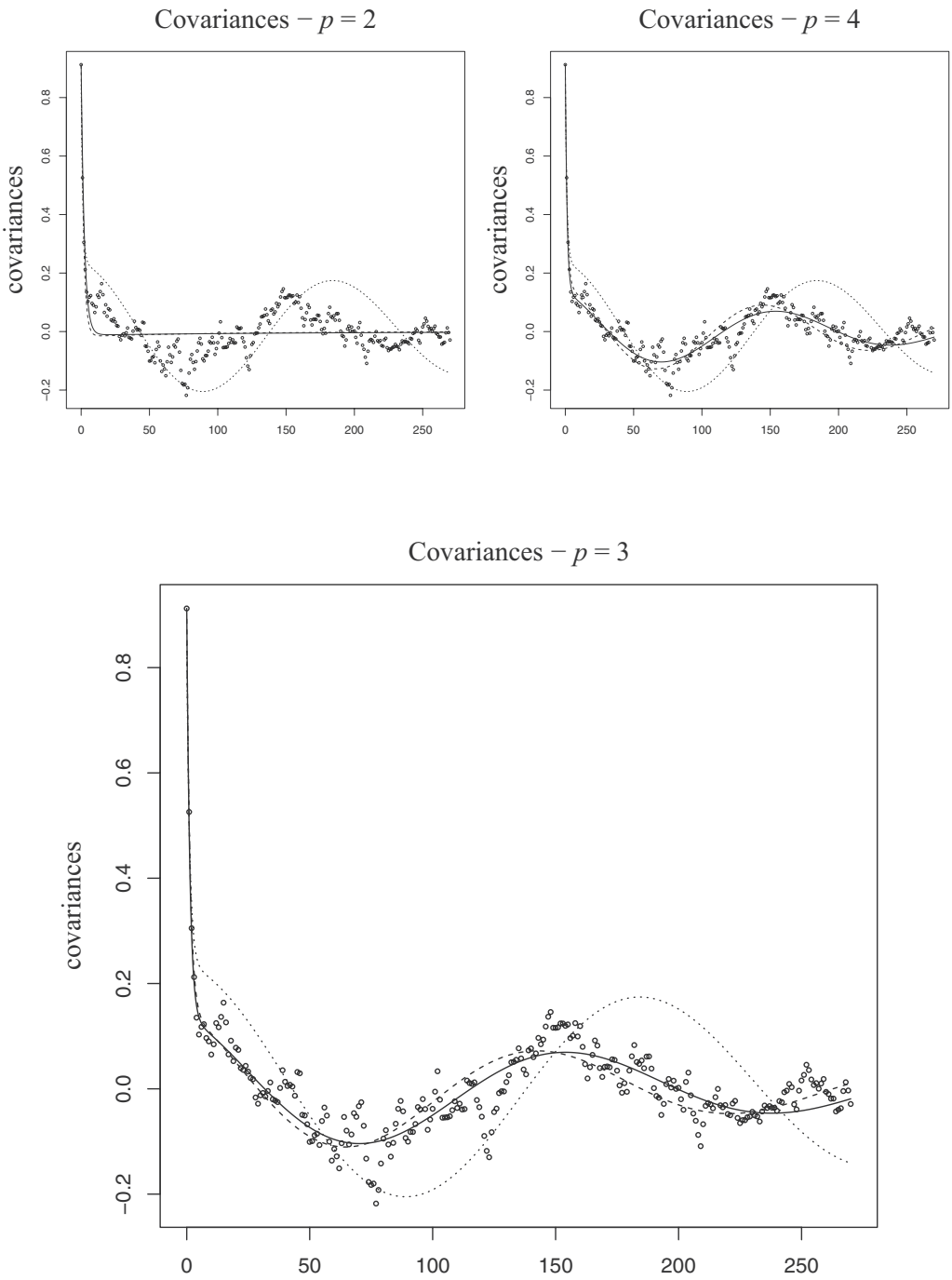


Figure 3: *Empirical covariances (○) and covariances of the MC (—) and ML (- - -) fitted OU models, for $p = 2, p = 4$ and $p = 3$, the actual value of the parameter, corresponding to Example 3. The covariances of $OU_{\boldsymbol{\kappa}}$ are indicated with a dotted line.*

The associated ARMA(3,2) model is

$$(1-1.9255B+1.05185B^2-0.1200B^3)x=(0.4831-0.9044B+0.4230B^2)\epsilon.$$

Example 3. The parameter $\boldsymbol{\kappa}=(0.83,0.0041,0.0009)$ used in the simulation of the OU process x treated in the present example is approximately equal to the parameter $\hat{\boldsymbol{\kappa}}$ obtained by ML estimation with $p=3$ for Series A in Section 6.5.1. A graphical presentation of the estimated covariances is given in Figure 3. The associated ARMA(3,2) model is

$$(1-2.4311B+1.8649B^2-0.4339B^3)x=(0.6973-1.3935B+0.6962B^2)\epsilon$$

The description of the performance of the model is complemented by comparing in Figure 4 the simulated values of the process in 400 equally spaced points filling the interval (199,201) with the predicted values for the same interval, based on the OU(3) model and the assumed observed data $x(0),x(2),x(3),\ldots,x(200)$. Also a confidence band limited by the predicted values plus and minus twice their standard deviation (2-st.-dev. confidence band) is included in the graph, in order to describe the precision of the predicted values.

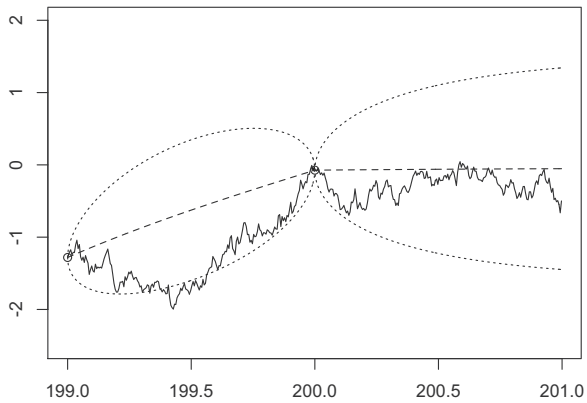


Figure 4: *Estimated interpolation and prediction of* $x(t)$ *for* $199<t<200$ *and* $200<t<201$*, respectively (- - -), 2-st.-dev. confidence bands based on* $(x(i))_{i=0,1,\ldots,200}$ $(\cdots)$*, and a refinement of the simulation of* $x(t)$ *on* $199<t<200$*.*

6.5. Applications to real data

In this section we present experimental results on two real data sets. We fit OU(p) processes for small values of p and also some ARMA processes. In each case we have observed that we can find an adequate value of p for which the empirical covariances are well approximated by the covariances of the adjusted OU(p) model. This is not the case

for the ARMA models adjusted by maximum likelihood, in all examples. We present a detailed comparison of both methodologies for the first example.

The first data set is taken from Box, Jenkins, and Reinsel (1994), and correspond to equally spaced observations of continuous-time processes that might be assumed to be stationary. The second one is a series obtained by choosing one in every 100 terms of a high frequency recording of oxygen saturation in blood of a newborn child. The data were obtained by a team of researchers of Pereira Rossell Children Hospital in Montevideo, Uruguay, integrated by L. Chiapella, A. Criado and C. Scavone. Their permission to analyse the data is gratefully acknowledged by the authors.

6.5.1. Box, Jenkins and Reinsel "Series A"

The Series A is a record of $n = 197$ chemical process concentration readings, taken every two hours, introduced with that name and analysed in (Box, Jenkins, and Reinsel, 1994, Ch. 4)[1]. Box et al. suggest an ARMA(1,1) as a model for this data, and subsets of AR(7) are proposed in (Cleveland, 1971) and (McLeod and Zhang, 2006). Figure 5 shows that these models fit fairly well the autocovariances for small lags, but fail to capture the structure of autocorrelations for large lags present in the series. On the other hand, the approximations obtained with the OU(p) processes, for $p = 3,4$ reflect both the short and long dependences, as shown in Figure 6.

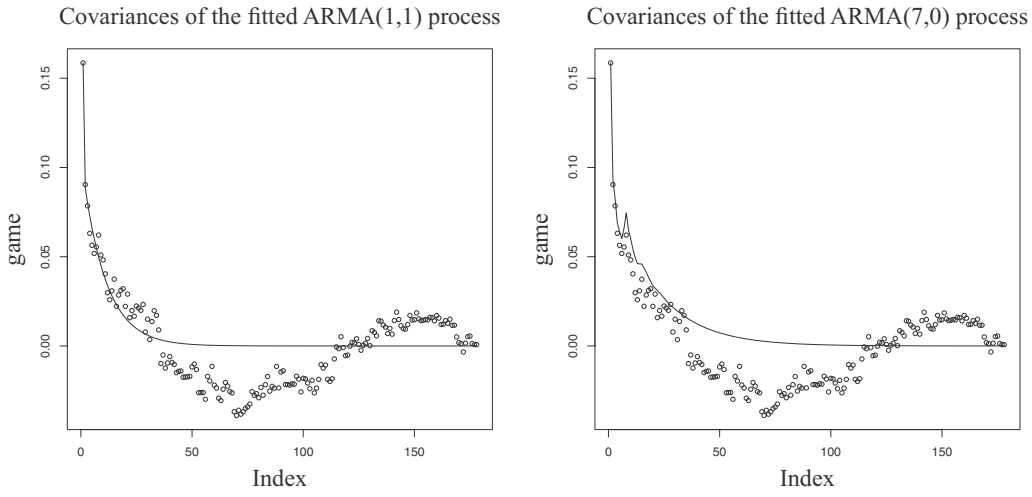


Figure 5: *Empirical covariances (∘) and covariances of the ML (—) fitted models ARMA(1,1) and AR(7) for Series A.*

1. see also http://rgm2.lab.nig.ac.jp/RGM2/tfunc.php?rd_id=FitAR:SeriesA

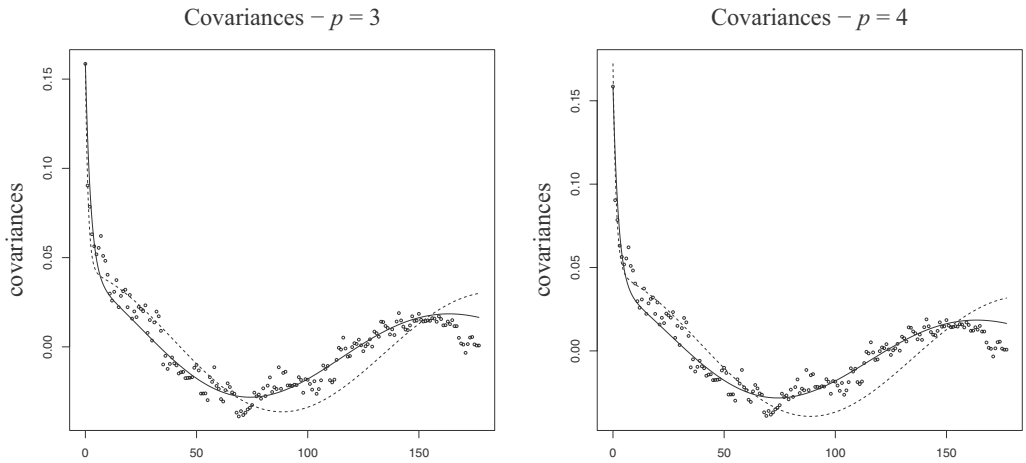


Figure 6: *Empirical covariances (∘) and covariances of the MC (—) and ML (- - -) fitted OU(p) models, for $p = 3,4$ corresponding to Series A.*

It is interesting to consider jointly the ARMA(3,2) model (31) fitted to the original data by maximum likelihood (computed also with the `R` function `arima`) and the ARMA(3,2) model (32) obtained by the procedure described in Section 5, corresponding to the OU(3) process also fitted to the data by maximum likelihood. The estimated parameters of this OU process are

$$\hat{\boldsymbol{\kappa}} = (0.8293, 0.0018 + 0.0330i, 0.0018 - 0.0330i) \text{ and } \hat{c} = 0.4401$$

and the ARMA(3,2) processes are respectively

$$(1-0.7945B-0.3145B^2+0.1553B^3)x=0.3101(1-0.4269B-0.2959B^2)\epsilon \qquad (31)$$

and

$$(1-2.4316B+1.8670B^2-0.4348B^3)x=0.4401(1-1.9675B+0.9685B^2)\epsilon. \qquad (32)$$

The autocorrelations of both ARMA models, shown in Figure 7, together with the empirical correlations of the series were computed by means of the `R` function `ARMAacf`, although the ones corresponding to (32) could have been obtained as the restrictions to integer lags of the covariance function for continuous-time described in Section 3.2. It is worth to notice that the autocorrelations of (31) do not approach the empirical correlations, indicated by circles, as much as the correlations of (32). The logarithms of the likelihoods of (31) and (32) are $\ell' = -49.23$, and $\ell'' = -50.95$, respectively. But since the number of parameters of the second model (which is four) is smaller than the number of parameters of the complete family of ARMA(3,2) processes (six), the Akaike

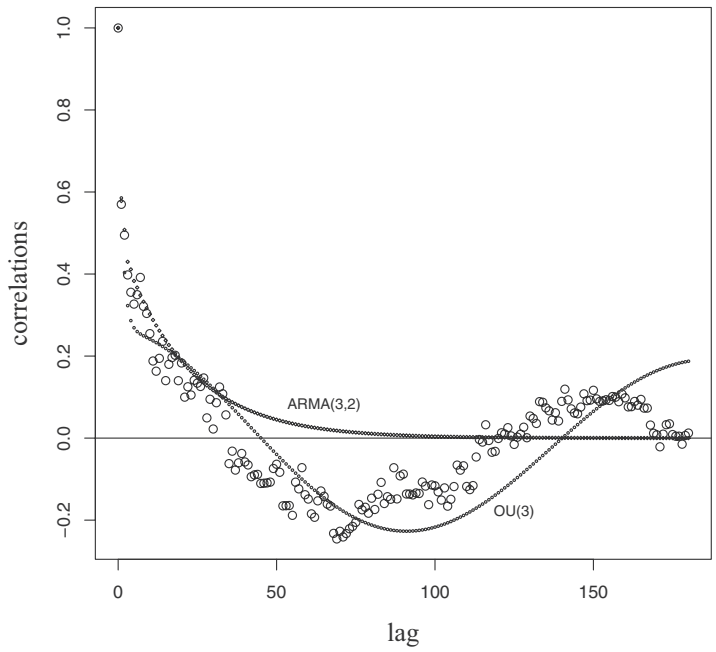


Figure 7: *Empirical correlations (○) of Series A, and autocorrelations of models (31) and (32) fitted by maximum likelihood from the family of all ARMA(3,2) and the restricted family of ARMA(3,2) derived from OU(3).*

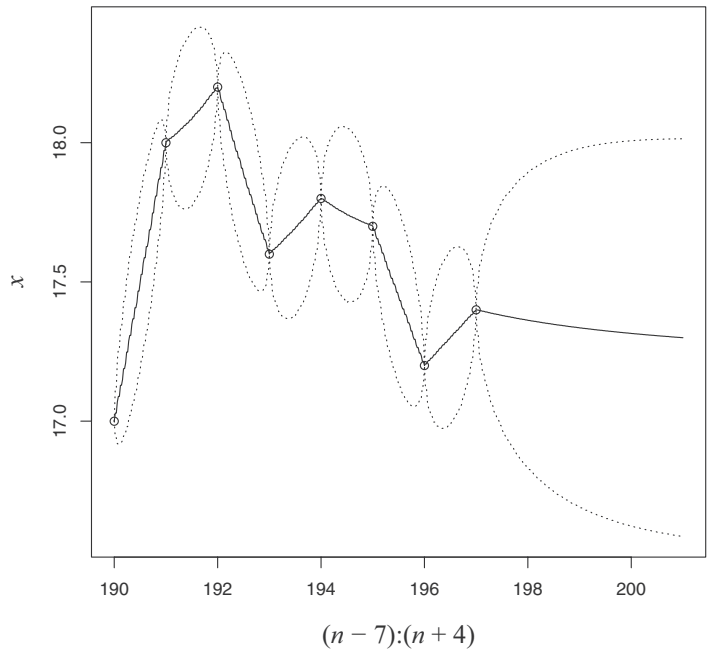


Figure 8: *Confidence bands for interpolated and extrapolated values of Series A for continuous domain.*

information criterion (AIC) of the parsimonious OU model is $8-2\ell'' = 109.90$, slightly better than the AIC of the unrestricted ARMA model, equal to $12-2\ell' = 110.46$.

Finally we show in Figure 8 the predicted values of the continuous parameter process $x(t)$, for t between $n-7$ and $n+4$ (190-201), obtained as the best linear predictions based on the last 90 observed values, and on the correlations given by the fitted OU(3) model. The upper and lower lines are two standard deviation confidence limits for each value of the process.

6.5.2. Oxygen saturation in blood

The oxygen saturation in blood of a newborn child has been monitored during 17 hours, and measures taken every two seconds. We assume that a series $x_0, x_1, \ldots, x_{304}$ of measures taken at intervals of 200 seconds is observed, and fit OU processes of orders $p = 2,3,4$ to that series.

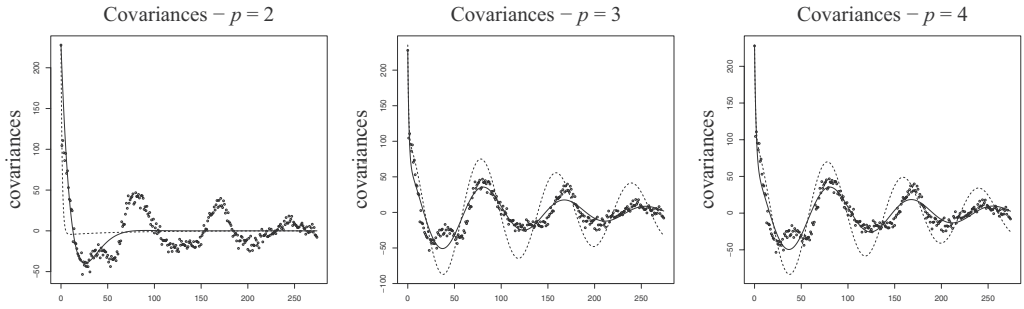


Figure 9: *Empirical covariances (∘) and covariances of the MC (—) and ML (- - -) fitted OU(p) models for $p = 2,3,4$ corresponding to the series of oxygen saturation in blood.*

Again the empirical covariances of the series and the covariances of the fitted OU(p) models for $p = 2,3,4$ are plotted (see Figure 9) and the estimated interpolation and extrapolation are shown in Figure 10. In the present case, the actual values of the series for integer multiples of 1/100 of the unit measure of 200 seconds are known, and plotted in the same figure.

6.6. Estimating the shape of the Lévy noise

There are various methods proposed in the literature to estimate the parameters of Lévy driven Ornstein–Uhlenbeck processes; in particular, the Lévy-Khinchin triplet comprised of two real numbers and a measure. For example, Valdivieso, Schoutens, and Tuerlinckx (2009) propose a maximum likelihood estimation methodology based on the inversion of the characteristic function of the Lévy process and the use of the discrete fast Fourier transform. Jongbloed, van der Meulen, and van der Vaart (2005) propose a nonparametric estimation based on a preliminary estimator of the characteristic function. Both methods require a large amount of information and intensive computation.

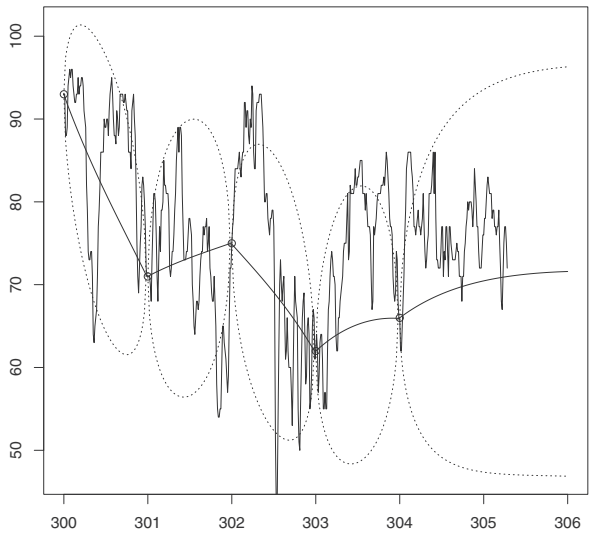


Figure 10: *Partial graph showing the five last values of the series of O_2 saturation in blood at integer multiples of the 200 seconds unit of time (∘), interpolated and extrapolated predictions (—), 2-st.-dev. confidence bands (⋯), and actual values of the series.*

We propose a naive method of estimating the parameters of the Lévy driven Ornstein–Uhlenbeck process that works in general situations when the maximum likelihood function is not known or difficult to approximate. These estimators are easy to compute, but also require a large amount of data to attain high accuracy.

Our method of estimation resembles the methods described in (Yu, 2004) consisting on matching the characteristic function derived from the model and the empirical characteristic function derived from the data.

Given a Lévy process $\Lambda(t)$, the characteristic function of $\Lambda(t)$ is $\mathrm{E}\,\mathrm{e}^{iu\Lambda(t)} = (\mathrm{E}\,\mathrm{e}^{iu\Lambda(1)})^t$, and is usually written as $\mathrm{E}\,\mathrm{e}^{iu\Lambda(1)} = \mathrm{e}^{\psi_\Lambda(iu)}$. The function $\psi_\Lambda(iu) = \log \mathrm{E}\,\mathrm{e}^{iu\Lambda(1)}$ is called *characteristic exponent* and has the form

$$\psi_\Lambda(iu) = aiu - \frac{\sigma^2}{2}u^2 + \int_{|x|<1} (\mathrm{e}^{iux} - 1 - iux)d\nu(x) + \int_{|x|\geq 1} (\mathrm{e}^{iux} - 1)d\nu(x)$$

where $\nu(\{0\}) = 0$, $\int_{|x|<1} x^2 d\nu(x) < \infty$, $\int_{|x|\geq 1} d\nu(x) < \infty$. The Lévy-Khinchin triplet is (σ^2, a, ν).

Assume that the admissible exponents belong to a parametric class $\Psi = \{\psi_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^d$, and obtain the value of θ for which a chosen quadratic distance between the exponential of $\psi_\theta(iu)$ and the empirical characteristic function of the residuals is minimum.

In order to ease notation, let us consider the case of an OU(p) model with parameter $\boldsymbol{\kappa}$ of pairwise different components; either $\boldsymbol{\kappa}$ is known or it is estimated by maximum likelihood or matching correlation methods. The innovation in each component ξ_j is

$$\eta_j(t) = \int_{t-1}^{t} \mathrm{e}^{-\kappa_j(t-s)} d\Lambda(s),$$

so that the innovation of x_κ is

$$\eta(t) = \int_{t-1}^{t} g(t-s) d\Lambda(s) \quad \text{where} \quad g(t) = \sum_{j=1}^{p} K_j \mathrm{e}^{-\kappa_j t}.$$

Hence, if we denote $\eta := \eta(1)$, we have

$$\eta \sim \int_0^1 g(1-s) d\Lambda(s) \sim \int_0^1 g(s) d\Lambda(s)$$

and its characteristic exponent is therefore

$$\psi_\eta(iu) = \log \mathrm{E}\mathrm{e}^{iu\eta} = \log \mathrm{E}\mathrm{e}^{iu\int_0^1 g(s)d\Lambda(s)} = \int_0^1 \psi_\Lambda(iug(s))ds$$

Example 4. Consider the estimation of a noise sum of a Poisson process plus a Gaussian term. Let us assume that the noise is given by

$$\Lambda(t) = \sigma W(t) + a(N(t) - \lambda t)$$

where W is a standard Wiener process and N is a Poisson process with intensity λ. The family of possible noises depends on the three parameters (σ, λ, a). In this case, the characteristic exponent has a simple form:

$$\psi_{\Lambda(1)}(iu) = -\frac{\sigma^2 u^2}{2} + \lambda(\mathrm{e}^{iua} - iua - 1),$$

hence

$$\psi_\eta(iu) = \int_0^1 \left(-\frac{\sigma^2 u^2 g^2(s)}{2} + \lambda(\mathrm{e}^{iug(s)a} - iug(s)a - 1) \right) ds$$

Defining $g_h = \int_0^1 g^h(s)ds$, we have

$$\psi_\eta(iu) = -\frac{\sigma^2 u^2 g_2}{2} + \lambda \left(-\frac{u^2 g_2 a^2}{2} - i\frac{u^3 g_3 a^3}{6} + \frac{u^4 g_4 a^4}{24} + \dots \right)$$

Then we propose to estimate the parameters by equating the coefficients of u^2, u^3, u^4 in $\psi_\eta(iu)$ with the corresponding ones in the logarithm of the empirical characteristic function of the residuals.

Assuming that the mean of the residuals $r_1, r_2, \ldots, r_n$ is zero, their empirical characteristic function is

$$\frac{1}{n}\sum_{h=1}^{n} \mathrm{e}^{iur_h} = 1 - \frac{1}{2}u^2 R_2 - \frac{1}{6}iu^3 R_3 + \frac{1}{24}u^4 R_4 + \ldots$$

where $R_m = \frac{1}{n}\sum_{h=1}^{n} r_h^m$. Then the logarithm has the expansion

$$\log \frac{1}{n}\sum_{h=1}^{n} \mathrm{e}^{iur_h} = -\frac{1}{2}u^2 R_2 - \frac{1}{6}iu^3 R_3 + \frac{1}{24}u^4 R_4 - \frac{1}{8}u^4 R_2^2 + \ldots$$

Consequently, the estimation equations are

$$\begin{aligned} (\sigma^2 + \lambda a^2) g_2 &= R_2, \\ \lambda a^3 g_3 &= R_3, \\ \lambda a^4 g_4 &= R_4 - 3R_2^2 \end{aligned}$$

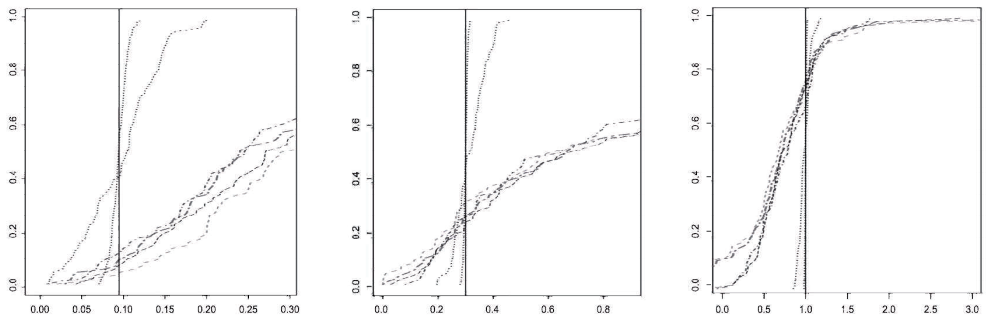
from which the estimators follow:

$$\tilde{a} = \frac{R_4 - 3R_2^2}{R_3}\frac{g_3}{g_4}, \qquad \tilde{\lambda} = \frac{R_3^4}{(R_4 - 3R_2^2)^3}\frac{g_4^3}{g_3^4},$$

$$\tilde{\sigma}^2 = \frac{R^2}{g_2} - \frac{R_3^2}{(R_4 - 3R_2^2)}\frac{g_4}{g_3^2}.$$

Figure 11 shows the empirical c.d.f. of 90 estimators of the parameters obtained from simulated series of 200 terms. The residuals were obtained by applying a Kalman filter to the space state formulation, starting from the actual value of $\boldsymbol{\kappa}$ used at the simulation (-·-), that in practical situations is unknown, and from matching correlations estimation (– –) and by maximum likelihood estimation (–·–).

The estimators are not sharp at all, but the ones obtained by the same procedure applied directly on the unfiltered noise Λ (– –) are equally rough. Larger series (of size 10000 and 1000000) produce sharper estimates, also shown in the figures by dotted lines.



c.d.f. of 90 estimators of σ *c.d.f. of 90 estimators of λ* *c.d.f. of 90 estimators of c*

Figure 11: *Estimation of the parameters of the noise (σ –left panel–, λ –center–, a –right–) from 90 replications of* $\{x_{\boldsymbol{\kappa}}(t) : t = 0,1,\ldots,200\}$, $\boldsymbol{\kappa} = (0.01 \pm 0.1i, 0.2)$, *driven by* $\Lambda(t) = 0.1W(t) + N_{0.3}(t) - 0.3t$. *Normality is rejected in all cases.*

7. Conclusions

We have proposed a family of continuous-time stationary processes, based on p iterations of the linear operator that maps a second order Lévy process onto an Ornstein-Uhlenbeck process. These operators have some nice properties, such as being commutative, and their p-compositions decompose as a linear combination of simple operators of the same kind. We remark that this result, stated in Theorem 1, is independent of the process onto which the operators $\mathcal{OU}_{\boldsymbol{\kappa}}$ act on. We have reduced the present scope of the applications envisaged by applying the operators only to Lévy processes, but other choices deserve consideration, for example, the results of applying the same operators to fractional Brownian motions.

An OU(p) process depends on $p+1$ parameters that can be easily estimated by either maximum likelihood (ML) or matching correlations (MC) procedures. MC estimators provide a fair estimation of the covariances of the data, even if the model is not well specified. When sampled on equally spaced instants, the OU(p) family can be written as a discrete time state-space model; i.e., a VARMA model in a space of dimension p. As a consequence, the families of OU(p) models are a parsimonious subfamily of the ARMA$(p, p-1)$ processes in the Gaussian case. Furthermore, the coefficients of the ARMA can be deduced from those of the corresponding OU(p). We have shown examples for which the ML-estimated OU model is able to capture features of the empirical autocorrelations at large lags that the ML-estimated ARMA model does not (see for instance Figure 7). This leads to recommend the inclusion of OU models as candidates to represent stationary series, either in discrete time or continuous-time.

References

Barndorff-Nielsen, O.E. (2001). Superposition of Ornstein-Uhlenbeck type processes. *Theory of Probability and Its Applications*, 45, 175–194.

Bergstrom, A.R. (1984). Continuous time stochastic models and issues of aggregation over time. *Handbook of Econometrics*, II, 1145–1212.

Bergstrom, A.R. (1996). Survey of continuous-time econometrics. In *Dynamic Disequilibrium Modeling: Theory and Applications: Proceedings of the Ninth International Symposium in Economic Theory and Econometrics*, volume 9, page 1. Cambridge University Press.

Box, G.E.P. Jenkins, G.M. and Reinsel, G.C. (1994). *Time Series Analysis, Forecasting and Control*. Prentice Hall.

Brockwell, P.J. (2004). Representations of continuous-time ARMA processes. *Journal of Applied Probability*, 41, 375–382.

Brockwell, P.J. (2009). Lévy–driven continuous–time ARMA processes. In *Handbook of Financial Time Series*, pages 457–480. Springer.

Cleveland, W.S. (1971). The inverse autocorrelations of a time series and their applications. *Technometrics*, 14, 277–298.

Doob, J.L. (1944). The elementary Gaussian processes. *Annals of Mathematical Statistics*, 15, 229–282.

Durbin, J. (1961). Efficient fitting of linear models for continuous stationary time-series from discrete data. *Bulletin of the International Statistical Institute*, 38, 273–282.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Eliazar, I. and Klafter, J. (2009). From Ornstein-Uhlenbeck dynamics to long-memory processes and Fractional Brownian motion. *Physical Review E*, 79, 021115.

Granger, C.W.J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14, 227–238.

Granger, C.W.J. and Morris, M.J. (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A*, 139, 246–257.

Jongbloed, G., van der Meulen, F.H. and van der Vaart, A.W. (2005). Nonparametric inference for Lévy-driven Ornstein-Uhlenbeck processes. *Bernoulli*, 11, 759–791.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

McLeod, A.I. and Zhang, Y. (2006). Partial autocorrelation parameterization for subset autoregression. *Journal of Time Series Analysis*, 27, 599–612.

Nieto, B., Orbe, S. and Zarraga, A. (2014). Time-Varying Market Beta: Does the estimation methodology matter? *SORT*, 31, 13–42.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distribution*, volume 68 of *Cambridge Studies in Advance Mathematics*. Cambridge University Press.

Thornton, M.A. and Chambers, M.J. (2013). Continuous-time autoregressive moving average processes in discrete time: representation and embeddability. *Journal of Time Series Analysis*, 34, 552–561.

Uhlenbeck, G.E. and Ornstein, L.S. (1930). On the theory of the Brownian motion. *Physical Review*, 36, 823–841.

Valdivieso, L., Schoutens, W. and Tuerlinckx, F. (2009). Maximum likelihood estimation in processes of Ornstein-Uhlenbeck type. *Statistical Inference for Stochastic Processes*, 12, 1–19.

Yu, J. (2004). Empirical characteristic function estimation and its applications. *Econometric Reviews*, 23, 93–123.

Appendix A: Proofs of Theorem 1 and its corollaries

Parts (i) and (iii) are obtained by direct computation of the integrals, (ii) follows from (i) by finite induction, as well as (iv) from (iii).

From the continuity of the integrals with respect to the parameter κ, the power $\mathcal{OU}_\kappa^p$ satisfies

$$\mathcal{OU}_\kappa^p = \lim_{\delta\downarrow 0}\prod_{j=1}^{p}\mathcal{OU}_{\kappa+j\delta} = \lim_{\delta\downarrow 0}\sum_{j=1}^{p}K'_j(\delta,\kappa,p)\mathcal{OU}_{\kappa+j\delta} \tag{33}$$

with

$$K'_j(\delta,\kappa,p) = \frac{1}{\prod_{1\le l\le p, l\neq j}\left(1-\frac{\kappa+l\delta}{\kappa+j\delta}\right)}.$$

On the other hand, by (i),

$$\prod_{h=1}^{q}\mathcal{OU}_{\kappa_h}^{p_h} = \lim_{\boldsymbol{\delta}\downarrow 0}\prod_{h=1}^{q}\prod_{j=1}^{p_h}\mathcal{OU}_{\kappa_h+j\delta_h} = \lim_{\boldsymbol{\delta}\downarrow 0}\sum_{h=1}^{q}\sum_{j=1}^{p_h}K''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa})\mathcal{OU}_{\kappa_h+j\delta_h} \tag{34}$$

where $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_q)$,

$$K''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa}) = \frac{1}{\prod_{\substack{1\le h'\le q, 1\le j'\le p_h,\\ (h',j')\neq(h,j)}}\left(1-\frac{\kappa_{h'}+j'\delta_{h'}}{\kappa_h+j\delta_h}\right)} = K'''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa})K'_j(\delta_h,\kappa_h,p_h),$$

and

$$K'''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa}) = \frac{1}{\prod_{\substack{1\le h'\le q,\\ h'\neq h}}\prod_{j'=1}^{p_{h'}}(1-(\kappa_{h'}+j'\delta_{h'})/(\kappa_h+j\delta_h))} \to K_h(\boldsymbol{\kappa}) \text{ as } \boldsymbol{\delta}\downarrow 0$$

For the h-th term in the right-hand side of (34), we compute

$$\begin{aligned}\lim_{\boldsymbol{\delta}\downarrow 0}\sum_{j=1}^{p_h}K''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa})\mathcal{OU}_{\kappa_h+j\delta_h} &= \lim_{\boldsymbol{\delta}\downarrow 0}\sum_{j=1}^{p_h}K'''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa})K'_j(\delta_h,\kappa_h,p_h)\mathcal{OU}_{\kappa_h+j\delta_h}\\ &= \lim_{\boldsymbol{\delta}\downarrow 0}\sum_{j=1}^{p_h}(K'''_{h,j}(\boldsymbol{\delta},\boldsymbol{\kappa})-K_h(\boldsymbol{\kappa}))K'_j(\delta_h,\kappa_h,p_h)\mathcal{OU}_{\kappa_h+j\delta_h}\\ &+\ K_h(\boldsymbol{\kappa})\lim_{\boldsymbol{\delta}\downarrow 0}\sum_{j=1}^{p_h}K'_j(\delta_h,\kappa_h,p_h)\mathcal{OU}_{\kappa_h+j\delta_h} = K_h(\boldsymbol{\kappa})\mathcal{OU}_{\kappa_h}^{p_h}\end{aligned}$$

by Equation (33) since, in addition, each term in the first sum tends to zero. This ends the verification of (v).

Corollary 1 is an immediate consequence of (iv) and (v), and Corollary 2 follows by applying (i) to compute

$$\begin{aligned}\mathcal{OU}_{\lambda+i\mu}\mathcal{OU}_{\lambda-i\mu} &= \frac{\lambda+i\mu}{2i\mu}\mathcal{OU}_{\lambda+i\mu} - \frac{\lambda-i\mu}{2i\mu}\mathcal{OU}_{\lambda-i\mu} \\ &= \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-s)} \left[\tfrac{\lambda+i\mu}{2i\mu}(\cos(\mu(t-s)) + i\sin(\mu(t-s))) \right. \\ &\quad \left. -\tfrac{\lambda-i\mu}{2i\mu}(\cos(\mu(t-s)) - i\sin(\mu(t-s)))\right] d\Lambda(s) \\ &= \int_{-\infty}^{t} \mathrm{e}^{-\lambda(t-s)}(\cos(\mu(t-s)) + \tfrac{\lambda}{\mu}\sin(\mu(t-s)))d\Lambda(s). \qquad \blacksquare\end{aligned}$$

Appendix B: Derivation of a state-space model

The form of equations (22) for a state-space representation of the OU(p) process in the general case can be derived by considering three special cases:

1. *When the components of $\boldsymbol{\kappa}$ are all different*. This case is treated in Section 4.
2. *When the components of $\boldsymbol{\kappa}$ are all equal*. Let κ denote the common value of the components of $\boldsymbol{\kappa}$. The state of the system is described by the vector

$$\boldsymbol{\xi}_{\kappa,p} = (\xi_\kappa^{(0)}, \xi_\kappa^{(1)}, \ldots, \xi_\kappa^{(p-1)})^\intercal,$$

with components $\xi_\kappa^{(h)}(t) = \displaystyle\int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} d\Lambda(s)$.
Each of these terms can be written as the sum

$$\xi_\kappa^{(h)}(t) = \mathrm{e}^{-\kappa} \int_{-\infty}^{t-1} \mathrm{e}^{-\kappa(t-1-s)} \frac{(-\kappa(t-1-s+1))^h}{h!} d\Lambda(s) + \eta_{\kappa,h}(t) \tag{35}$$

where $\eta_{\kappa,h}(t) = \displaystyle\int_{t-1}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} d\Lambda(s)$.
The first term in the right-hand side of (35) is equal to

$$\begin{aligned}&\mathrm{e}^{-\kappa} \sum_{j=0}^{h} \frac{(-\kappa)^{h-j}}{(h-j)!} \int_{-\infty}^{t-1} \mathrm{e}^{-\kappa(t-1-s)} \frac{(-\kappa(t-1-s))^j}{j!} d\Lambda(s) \\ &= \mathrm{e}^{-\kappa} \sum_{j=0}^{h} \frac{(-\kappa)^{h-j}}{(h-j)!} \xi_\kappa^{(j)}(t-1)\end{aligned}$$

and therefore, by introducing the matrix

$$\boldsymbol{A}_{\kappa,p} = \mathrm{e}^{-\kappa} \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ \frac{(-\kappa)}{1!} & 1 & 0 & \dots & 0 & 0 \\ \frac{(-\kappa)^2}{2!} & \frac{(-\kappa)}{1!} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{(-\kappa)^{p-2}}{(p-2)!} & \frac{(-\kappa)^{p-3}}{(p-3)!} & \frac{(-\kappa)^{p-4}}{(p-4)!} & \dots & 1 & 0 \\ \frac{(-\kappa)^{p-1}}{(p-1)!} & \frac{(-\kappa)^{p-2}}{(p-2)!} & \frac{(-\kappa)^{p-3}}{(p-3)!} & \dots & \frac{(-\kappa)}{1!} & 1 \end{pmatrix}$$

we may write

$$\boldsymbol{\xi}_{\kappa,p}(t) = \boldsymbol{A}_{\kappa,p}\boldsymbol{\xi}_{\kappa,p}(t-1) + \boldsymbol{\eta}_{\kappa,p}$$

where $\boldsymbol{\eta}_{\kappa,p}(t) = (\eta_{\kappa,0}(t), \eta_{\kappa,1}(t), \dots, \eta_{\kappa,p-1}(t))^\top$ is a vector of centered innovations (independent of the σ-algebra generated by $\{\Lambda(s) : s \le t-1\}$) with covariance matrix $\boldsymbol{B}_{\kappa,\kappa,p}$ obtained with $\kappa_1 = \kappa_2$ and $p_1 = p_2$ from the general expression of the $p_1 \times p_2$ matrix $B_{\kappa_1,\kappa_2,p_1,p_2} = ((b_{\kappa_1,\kappa_2,h_1,h_2}))_{1\le h_1\le p_1, 1\le h_2\le p_2}$, where

$$\begin{aligned} b_{\kappa_1,\kappa_2,h_1,h_2} &= \mathrm{E}\,\eta_{\kappa_1,h_1}(t)\bar{\eta}_{\kappa_2,h_2}(t) \\ &= v^2 \int_{t-1}^{t} \mathrm{e}^{-(\kappa_1+\bar{\kappa}_2)(t-s)}(-\kappa_1)^{h_1}(-\bar{\kappa}_2)^{h_2}(t-s)^{h_1+h_2}ds \\ &= v^2 \int_0^1 \mathrm{e}^{-(\kappa_1+\bar{\kappa}_2)y}(-\kappa)^{h_1}(-\bar{\kappa})^{h_2}y^{h_1+h_2}dy. \end{aligned}$$

The equation $x(t) = \boldsymbol{K}_p^\top \boldsymbol{\xi}_{\boldsymbol{\kappa},p}(t)$, with $\boldsymbol{K}_p^\top = (\binom{p-1}{0}, \binom{p-1}{1}, \dots, \binom{p-1}{p-1})$ completes the description of the system state dynamics.

3. *The vector κ has components $\kappa_1 = \lambda + \mu i$ and $\kappa_2 = \lambda - \mu i$, $\mu \neq 0$, each repeated p_1 times*. A description involving imaginary processes is immediate from the previous case. The equations

$$\begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t) \end{pmatrix} = \begin{pmatrix} A_{\kappa_1,p_1} & 0 \\ 0 & A_{\kappa_2,p_1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t-1) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t-1) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_{\kappa_1,p_1} \\ \boldsymbol{\eta}_{\kappa_2,p_1} \end{pmatrix}$$

$$x(t) = (\boldsymbol{K}_{p_1}^\top, \boldsymbol{K}_{p_1}^\top) \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t) \end{pmatrix}$$

hold, and $\mathrm{Var}\begin{pmatrix} \boldsymbol{\eta}_{\kappa_1,p_1} \\ \boldsymbol{\eta}_{\kappa_2,p_1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_{\kappa_1,\kappa_1,p_1,p_1} & \boldsymbol{B}_{\kappa_1,\kappa_2,p_1,p_1} \\ \boldsymbol{B}_{\kappa_2,\kappa_1,p_1,p_1} & \boldsymbol{B}_{\kappa_1,\kappa_1,p_1,p_1} \end{pmatrix}$.

A description in terms of real processes is obtained by multiplying the first equation by the matrix

$$\boldsymbol{C}_{p_1} = \begin{pmatrix} I_{p_1} & I_{p_1} \\ -iI_{p_1} & iI_{p_1} \end{pmatrix}$$

(I_p denotes the $p \times p$ identity matrix), because the vectorial process $\boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t) \end{pmatrix}$ has real components. The new equations are

$$\boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t) \end{pmatrix} = \boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{\eta}_{\kappa_1,p_1} \\ \boldsymbol{\eta}_{\kappa_2,p_1} \end{pmatrix} + \left(\boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{A}_{\kappa_1,p_1} & 0 \\ 0 & \boldsymbol{A}_{\kappa_2,p_1} \end{pmatrix} \boldsymbol{C}_{p_1}^{-1} \right) \times \left(\boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t-1) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t-1) \end{pmatrix} \right)$$

and

$$\boldsymbol{x}(t) = \left((\boldsymbol{K}_{p_1}^{\top}, \boldsymbol{K}_{p_1}^{\top}) C_{p_1}^{-1} \right) \times \left(\boldsymbol{C}_{p_1} \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_1}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_1}(t) \end{pmatrix} \right)$$

General case, real processes

Let us assume that $\kappa_1, \ldots, \kappa_q$ are distinct components of $\boldsymbol{\kappa}$, each repeated $p_1, \ldots, p_q$ times. We assume in addition that the imaginary components are $\kappa_1, \kappa_2 = \bar{\kappa}_1, \ldots, \kappa_{2c-1}$, $\kappa_{2c} = \bar{\kappa}_{2c-1}$ and the remaining $\kappa_{2c+1}, \ldots, \kappa_q$ are real. With this notation, $p_{2h-1} = p_{2h}$ for $h = 1, 2, \ldots, c$. We make intensive use of the notations introduced in previous cases to write

$$\boldsymbol{\xi}(t) = \boldsymbol{A}\boldsymbol{\xi}(t-1) + \boldsymbol{\eta}(t), \tag{36}$$

$$\boldsymbol{x}(t) = K^{\top}\boldsymbol{\xi}(t)$$

with

$$\boldsymbol{\xi}(t) = \begin{pmatrix} \boldsymbol{\xi}_{\kappa_1,p_2}(t) \\ \boldsymbol{\xi}_{\kappa_2,p_2}(t) \\ \boldsymbol{\xi}_{\kappa_3,p_4}(t) \\ \boldsymbol{\xi}_{\kappa_4,p_4}(t) \\ \cdots \\ \boldsymbol{\xi}_{\kappa_{2c-1},p_{2c}}(t) \\ \boldsymbol{\xi}_{\kappa_{2c},p_{2c}}(t) \\ \boldsymbol{\xi}_{\kappa_{2c+1},p_{2c+1}}(t) \\ \boldsymbol{\xi}_{\kappa_{2c+2},p_{2c+2}}(t) \\ \cdots \\ \boldsymbol{\xi}_{\kappa_q,p_q}(t) \end{pmatrix}, \boldsymbol{\eta}(t) = \begin{pmatrix} \boldsymbol{\eta}_{\kappa_1,p_2}(t) \\ \boldsymbol{\eta}_{\kappa_2,p_2}(t) \\ \boldsymbol{\eta}_{\kappa_3,p_4}(t) \\ \boldsymbol{\eta}_{\kappa_4,p_4}(t) \\ \cdots \\ \boldsymbol{\eta}_{\kappa_{2c-1},p_{2c}}(t) \\ \boldsymbol{\eta}_{\kappa_{2c},p_{2c}}(t) \\ \boldsymbol{\eta}_{\kappa_{2c+1},p_{2c+1}}(t) \\ \boldsymbol{\eta}_{\kappa_{2c+2},p_{2c+2}}(t) \\ \cdots \\ \boldsymbol{\eta}_{\kappa_q,p_q}(t) \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} A_{\kappa_1,p_2} & 0 & \dots & 0 \\ 0 & A_{\kappa_2,p_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{\kappa_q,p_q} \end{pmatrix}$$

and

$$\boldsymbol{K}^\top = (K^\top_{\kappa_1,p_1}, K^\top_{\kappa_2,p_2}, \dots, K^\top_{\kappa_q,p_q}).$$

The real version, when the process $\boldsymbol{\xi}$ has imaginary components is obtained by multiplying (36) by the matrix

$$\boldsymbol{C} = \begin{pmatrix} C_{p_2} & 0 & \dots & 0 & 0 \\ 0 & C_{p_4} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & C_{p_{2c}} & 0 \\ 0 & 0 & \dots & 0 & I_{p_{2c+1}+\dots+p_q} \end{pmatrix} \tag{37}$$

thus obtaining

$$\boldsymbol{C\xi}(t) = (\boldsymbol{CAC}^{-1}) \times (\boldsymbol{C\xi}(t-1)) + \boldsymbol{C\eta}(t), \tag{38}$$

$$\boldsymbol{x}(t) = (\boldsymbol{K}^\top \boldsymbol{C}^{-1}) \times (\boldsymbol{C\xi}(t)). \tag{39}$$

Appendix C: Identification of the ARMA

In order to find the coefficients of the ARMA with the same autocovariances as $x(t)$ in the general case, we need the following technical results.

Lemma 1 *For each positive integer p,* $\sum_{j=1}^{p} j^{p-1} \prod_{l \neq j} \frac{1}{j-l} = 1$ *and for* $h = 0, 1, \dots, p-2$, $\sum_{j=1}^{p} j^h \prod_{l \neq j} \frac{1}{j-l} = 0$.

Proof: The polynomial $G(z) = \sum_{j=1}^{p} \left(\frac{1}{j}\right)^{p-1-h} \prod_{l \neq j} \frac{1-lz}{1-l/j}$ has degree $p-1$ and coincides for p different values of the variable, namely $z = 1/j$, $j = 1, 2, \dots, p$, with the polynomial z^{p-1-h}, also of degree not greater than $p-1$ for $h = 0, 1, \dots, p-1$. Therefore, both polynomials are identical, and hence $G(0) = 0$ for $h < p-1$ and $G(0) = 1$ for $h = p-1$. ∎

Lemma 2 *The power series* $g(z,n) = \sum_{h=0}^{\infty} z^h h^n$, $|z| < 1, n = 0,1,2,\ldots$ *has the sum*

$$\sum_{h=0}^{n} \alpha_{n,h}(1-z)^{-h-1}$$

with coefficients determined by $\alpha_{0,0} = 1$ *and the recurrence relations*

$$\alpha_{n+1,h} = h\alpha_{n,h-1} - (h+1)\alpha_{n,h}, h = 0,1,\ldots,n+1, n = 0,1,2,\ldots, \alpha_{n,n+1} = 0.$$

In particular, $\alpha_{n,0} = (-1)^n$. ∎

As an intermediate step from the case described in Section 5.1 and building to the general case, let us approach the OU(p) process x with parameter equal to the p-vector with equal components $\boldsymbol{\kappa} = (\kappa,\kappa,\ldots,\kappa)^\intercal$ as the limit of $x_\delta = \mathcal{OU}_{\boldsymbol{\kappa}(\delta)}\Lambda$, $\boldsymbol{\kappa}(\delta) = (\kappa(1+\delta),\kappa(1+2\delta),\ldots,\kappa(1+p\delta))^\intercal$ when δ tends to zero. From the results in Section 5.1 we use the representation

$$x_\delta = \sum_{j=1}^{p} K_j \xi_j, \quad K_j = \frac{(1+j\delta)^{p-1}}{\delta^{p-1}} \prod_{l \neq j} \frac{1}{j-l} \tag{40}$$

in terms of the vector

$$\boldsymbol{\xi} = (\xi_1,\xi_2,\ldots,\xi_p)^\intercal, \ \ \xi_j(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(1+j\delta)(t-s)} d\Lambda(s)$$

that satisfies $\boldsymbol{\xi} = \mathrm{diag}(\mathrm{e}^{-\kappa(1+j\delta)})B\boldsymbol{\xi} + \boldsymbol{\eta}$ where B is the backshift operator defined in Section 2 and

$$\eta_j(t) = \int_{t-1}^{t} \mathrm{e}^{-\kappa(1+j\delta)(t-s)} d\Lambda(s)$$

and introduce the power expansions

$$\xi_j(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)} \sum_{h=0}^{\infty} \frac{(j\delta)^h(-\kappa(t-s))^h}{h!} d\Lambda(s) = \sum_{h=0}^{\infty} (j\delta)^h \xi_\kappa^{(h)}(t)$$

with $\xi_\kappa^{(h)}(t) = \int_{-\infty}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} d\Lambda(s)$ and the similar expansion for the innovations

$$\eta_j(t) = \sum_{h=0}^{\infty} (j\delta)^h \eta_\kappa^{(h)}(t) \text{ with } \eta_\kappa^{(h)}(t) = \int_{t-1}^{t} \mathrm{e}^{-\kappa(t-s)} \frac{(-\kappa(t-s))^h}{h!} d\Lambda(s). \tag{41}$$

We write now the ARMA model

$$\prod_{j=1}^{p}(1-\mathrm{e}^{-\kappa(1+j\delta)}B)x_\delta = \sum_{j=1}^{p}\prod_{l\neq j}(1-\mathrm{e}^{-\kappa(1+l\delta)}B)K_j\eta_j$$

and notice that the limit when $\delta \to 0$ of the left-hand side is $(1-\mathrm{e}^{-\kappa}B)^p x$.

In order to take limits at the right-hand side, we replace K_j by its expression in (40), expand $\prod_{l\neq j}(1-\mathrm{e}^{-\kappa(1+l\delta)}B)$ as the product of the series

$$\prod_{l=1}^{p}(1-\mathrm{e}^{-\kappa(1+l\delta)}B) = \sum_{\nu=0}^{\infty}a_\nu\delta^\nu \tag{42}$$

independent of j and

$$(1-\mathrm{e}^{-\kappa(1+j\delta)}B)^{-1} = \sum_{h=0}^{\infty}(\mathrm{e}^{-\kappa(1+j\delta)}B)^h = \sum_{\mu=0}^{\infty}b_\mu(j\delta)^\mu \tag{43}$$

with coefficients independent of j and substitute the expansion (41) for η_j thus obtaining the series

$$\sum_{j=1}^{p}\left(\sum_{\nu=0}^{\infty}a_\nu\delta^\nu \times \sum_{\mu=0}^{\infty}b_\mu(j\delta)^\mu \times (1+j\delta)^{p-1}\prod_{l\neq j}\frac{1}{j-l} \times \sum_{h=0}^{\infty}(j\delta)^h\eta_\kappa^{(h)}\right)$$

divided by δ^{p-1}. After ordering this series by increasing powers of δ, it may be noticed that the terms in δ raised to a power smaller than $p-1$ vanish, because their coefficient include a factor $\sum_{j=1}^{p} j^h \prod_{l\neq j}\frac{1}{j-l}$ with $h \in \{0,1,\ldots,p-2\}$ that is equal to zero as established in Lemma 1 below. Therefore, the limit when $\delta \to 0$ of the series divided by δ^{p-1} is the coefficient of δ^{p-1} in the series. Unless the term a_0 of the first factor is taken, the power of j appearing in the coefficient of δ^{p-1} will be smaller than $p-1$ and again Lemma 1 leads to conclude that the coefficient vanishes. Therefore, since the same lemma establishes that $\sum_{j=1}^{p} j^{p-1}\prod_{l\neq j}\frac{1}{j-l} = 1$, the required limit is the linear combination of moving averages

$$a_0 \sum_{\mu+i+h=p-1}\binom{p-1}{i}b_\mu\eta_\kappa^{(h)} \tag{44}$$

where it remains to make explicit the dependence with respect to the backshift operator B.

From (42) it follows immediately that $a_0 = (1-\mathrm{e}^{-\kappa}B)^p$, while from (43) we get $b_\mu j^\mu \mu! = \left[\frac{\partial^\mu}{\partial\delta^\mu}\sum_{h=0}^{\infty}\mathrm{e}^{-\kappa h}B^h\mathrm{e}^{-hj\delta}\right]_{\delta=0} = (-j)^\mu\sum_{h=0}^{\infty}(\mathrm{e}^{-\kappa}B)^h h^\mu$ and hence

$$b_\mu = \frac{(-1)^\mu}{\mu!}\sum_{\nu=0}^{\infty}(\mathrm{e}^{-\kappa}B)^\nu\nu^\mu.$$

Now we apply Lemma 2 (stated at the end of this section) such that, with the coefficients $\alpha_{\mu,\nu}$ there defined, leads us to write

$$\sum_{\nu=0}^{\infty}(\mathrm{e}^{-\kappa}B)^\nu\nu^\mu = \sum_{\nu=0}^{\mu}\alpha_{\mu,\nu}(1-\mathrm{e}^{-\kappa}B)^{-\nu-1}$$

and therefore (44) is equal to the moving average of order at most $p-1$

$$\sum_{\mu+i+h=p-1}\binom{p-1}{i}\frac{(-1)^\mu}{\mu!}\sum_{\nu=0}^{\mu}\alpha_{\mu,\nu}(1-\mathrm{e}^{-\kappa}B)^{p-\nu-1}\eta_\kappa^{(h)}. \tag{45}$$

Let us observe finally that the order of the moving average is actually $p-1$. The term in B^{p-1} corresponds to $\nu=0$ and reduces to

$$\sum_{\mu+i+h=p-1}\binom{p-1}{i}\frac{(-1)^\mu}{\mu!}\alpha_{\mu,0}(-1)^{p-1}e^{-(p-1)\kappa}B^{p-1}\eta_\kappa^{(h)}.$$

At least the term in $B^{p-1}\eta_\kappa^{(p-1)}$ with coefficient $(-1)^{p-1}e^{-(p-1)\kappa}$ does not vanish. On the other hand, neither the term with lag zero in $\eta_\kappa^{(p-1)}$ vanishes, because its coefficient is $\alpha_{0,0}=1$.

General case. We now join the previous results for the general case with parameter $\boldsymbol{\kappa}$, a p-vector with p_j components equal to κ_j, $j=1,2,\ldots,q$, with $\kappa_1,\ldots,\kappa_q$ all different of each other and $\sum_{j=1}^q p_j = p$. We use the result of Theorem 1(1) and conclude that $x=\mathcal{OU}_{\boldsymbol{\kappa}}(\Lambda)$ has the same second-order moments as the ARMA$(p,p-1)$ model

$$\prod_{j=1}^{q}(1-\mathrm{e}^{-\kappa_j}B)^{p_j}x = \sum_{j=1}^{q}K_j\prod_{l\neq j}(1-\mathrm{e}^{-\kappa_l}B)^{p_l}\mathrm{MA}_j \tag{46}$$

with MA_j the moving average of order p_j-1 given by Equation (45).

SORT 40 (2) July-December 2016, 303-320

Modelling extreme values by the residual coefficient of variation

Joan del Castillo[1,*] and Maria Padilla[2]

Abstract

The possibilities of the use of the coefficient of variation over a high threshold in tail modelling are discussed. The paper also considers multiple threshold tests for a generalized Pareto distribution, together with a threshold selection algorithm. One of the main contributions is to extend the methodology based on moments to all distributions, even without finite moments. These techniques are applied to euro/dollar daily exchange rates and to Danish fire insurance losses.

MSC: 62E20, 62E15, 62M02.

Keywords: Statistics of extremes, heavy tails, high quantile estimation, value at risk.

1. Introduction

Fisher and Tippett (1928) and Gnedenko (1943) show that, under regularity conditions, the limit distribution for the normalized maximum of a sequence of independent and identically distributed (iid) random variables (r.v.) is a member of the generalized extreme value (GEV) distribution with a cumulative distribution function

$$H_\xi(x) = \exp\{-(1+\xi x)^{-1/\xi}\}, \quad (1+\xi x) > 0,$$

where ξ is called *extreme value index*. This family of continuous distributions contains the Fréchet distribution ($\xi > 0$), the Weibull distribution ($\xi < 0$), and the Gumbell distribution ($\xi = 0$, as a limit case), see McNeil et al. (2005) and Gomes and Guillou (2015).

* *Corresponding author:* castillo@mat.uab.cat

[1] Department of Mathematics, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Spain.

[2] Department of Mathematics. Universitat Autònoma de Barcelona. 08193, Cerdanyola del Vallès (Barcelona). Spain.

Received: March 2016
Accepted: August 2016

The Pickands–Balkema–DeHaan theorem, see Embrechts et al. (1997) and McNeil et al. (2005), initiated a new way of studying extreme value theory via distributions above a threshold, which use more information than the maximum data grouped into blocks. This theorem is a very widely applicable result that essentially says that the generalized Pareto distribution (GPD) is the canonical distribution for modelling excess losses over high thresholds. The cumulative distribution function of GPD(ξ,ψ) is

$$F(x) = 1-(1+\xi x/\psi)^{-1/\xi}, \tag{1}$$

where $\psi > 0$ and ξ are scale and shape parameters. For $\xi > 0$ the range of x is $x > 0$, being in this case the usual Pareto distribution. The limit case $\xi = 0$ corresponds to the exponential distribution. For $\xi < 0$ the range of x is $0 < x < \psi/|\xi|$ and GPD has bounded support. The GPD has mean $\psi/(1-\xi)$ and variance $\psi^2/[(1-\xi)^2(1-2\xi)]$ provided $\xi < 1/2$.

Let X be a continuous non-negative r.v. with distribution function $F(x)$. For any threshold, $t > 0$, the r.v. of the conditional distribution of threshold excesses $X - t$ given $X > t$, denoted as $X_t = \{X - t \mid X > t\}$, is called the *residual distribution* of X over t. The cumulative distribution function of X_t, $F_t(x)$, is given by

$$1 - F_t(x) = (1-F(x+t))/(1-F(t)). \tag{2}$$

The quantity $M(t) = E(X_t)$ is called the *residual mean* and $V(t) = \mathrm{Var}(X_t)$ the *residual variance*. The *residual coefficient of variation* (CV) is given by

$$\mathrm{CV}(t) \equiv \mathrm{CV}(X_t) = \sqrt{V(t)}/M(t), \tag{3}$$

like the usual CV, the function $\mathrm{CV}(t)$ is independent of scale, that is, if λ is a positive constant then $\mathrm{CV}(\lambda X_t) = \mathrm{CV}(X_t)$.

The residual distribution of a GPD is again GPD and for any threshold $t > 0$, the shape parameter ξ is invariant, in fact

$$\mathrm{GPD}_t(\xi,\psi) = \mathrm{GPD}(\xi,\psi+\xi t). \tag{4}$$

Note that the residual CV is independent of the threshold and the scale parameter, since it is given by

$$\mathrm{CV}(t) = c_\xi = \sqrt{1/(1-2\xi)}. \tag{5}$$

Gupta and Kirmani (2000) show that the residual CV characterizes the distribution in univariate and bivariate cases, provided that a finite second moment exists. In the case of GPD, the residual CV is constant and it is a one to one transformation of the extreme value index suggesting its use to estimate this index.

Castillo et al. (2014) suggest a new tool to identify the tail of a distribution based on the residual CV, henceforth called CV-plot, as an alternative to the *mean excess plot* (ME-plot), a commonly used diagnostic tool in risk analysis to justify fitting a GPD, see Ghosh and Resnick (2010), Embrechts et al. (1997) and Davison and Smith (1990). What is important here is the fact that for a GPD distribution with $\xi < 1$, the residual mean function $t \to M(t)$ is linear with positive, negative or zero slope depending on whether $0 < \xi < 1$, $\xi < 0$ or $\xi = 0$.

Given a sample $\{x_k\}$ of size n of positive numbers, we denote the ordered sample $\{x_{(k)}\}$, so that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The CV-plot is the function $cv(t)$ of the sample coefficient of variation of the *threshold excesses* $(x_j - t)$ for the *exceedances* $\{x_j : x_j > t\}$ given by

$$t \to cv(t) = \frac{sd\{x_j - t \mid x_j > t\}}{\text{mean}\{x_j - t \mid x_j > t\}}, \tag{6}$$

in practice $t = x_{(k)}$ are the order statistics, where, k $(1 \leq k \leq n)$ is the size of the subsample removed. This tool has been applied to financial and environmental datasets, see Castillo and Serra (2015).

The CV-plot has two advantages over ME-plot: first, ME-plot depends on a scale parameter and CV-plot does not; second, linear functions are defined by two parameters and the constants by only one. So the uncertainty is reduced from three to one single parameter.

A unconscientious use of some measures of variation can lead to wrong conclusions, see Albrecher et al. (2010). A serious problem with the residual coefficient of variation is the fact that the proposed method only works when the extreme value index is smaller than 0.25 (otherwise its variance is not finite). To fix this, some transformations that relate light-heavy tails are introduced in Section 2.

Section 3 extends some results of Castillo et al. (2014) from the exponential distribution to all GPD when the extreme value index is below 0.25. Moreover, multiple threshold tests together with a threshold selection algorithm, designed in a way that avoids subjectivity, are also achieved. In Section 4, these techniques are applied first to euro/dollar daily exchange rates and validated with out of sample observations. Secondly, the approach developed in Section 2, is illustrated using the Danish fire insurance dataset, a highly heavy-tailed, infinite-variance model.

2. Transformations of heavy-light tails

The transformations introduced to this section make it possible to estimate the extreme value index using methods based on moments in situations where moments are not finite.

A distribution function F is said to be in the maximum domain of attraction of H_ξ, written $F \in D(H_\xi)$, if under appropriate normalization the block maxima of an iid se-

quence of r.v. with distribution F converge to H_ξ. For a r.v. X with distribution function F is also written $X \in D(H_\xi)$. A positive function L on $(0,\infty)$ *slowly varies* at ∞ if

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1, \quad t > 0.$$

Regularly varying functions can be represented by power functions multiplied by slowly varying functions, i.e. $h(x) \in \mathrm{RV}_\rho$ if and only if $h(x) = x^\rho L(x)$.

Gnedenko proved, see McNeil et al. (2005, Theorems 7.8 and 7.10), that the maximum domain of attraction of a Fréchet distribution, with shape parameter $\xi > 0$, is characterized in terms of the tail function, $\overline{F}(x) = 1 - F(x)$, by

$$F \in D(H_\xi) \Leftrightarrow \overline{F} \in \mathrm{RV}_{-1/\xi} \quad (\xi > 0).$$

Similarly the maximum domain of attraction of a Weibull distribution, with shape parameter $\xi < 0$, is characterized by

$$F \in D(H_\xi) \Leftrightarrow \overline{F}(x_+ - 1/x) \in \mathrm{RV}_{1/\xi} \quad (\xi < 0),$$

where $x_+ = \sup\{x : F(x) < 1\}$.

The following result of practical importance is embedded in the previous characterizations, and which to our knowledge it has not been pointed out.

Corollary 1 *Let X be a continuous r.v. with cumulative distribution function F.*

(1) If $X \in D(H_\xi), X > 0$, with $\xi > 0$, then $X^* = -1/X \in D(H_{-\xi})$.

(2) If $X \in D(H_\xi)$ with $\xi < 0$, then $X^* = -1/(X - x_+) \in D(H_{-\xi})$, where $x_+ = \sup\{x : F(x) < 1\}$.

Proof. (1) The cumulative distribution function of X^* is $F^*(x) = F(-1/x)$ and $x_+ = \sup\{x : F^*(x) < 1\} = 0$. By assumption $\overline{F}(x) = x^{-1/\xi}L(x)$ with L slowly varying at ∞, hence $\overline{F^*}(x_+ - 1/x) = \overline{F}(x) = x^{1/(-\xi)}L(x)$ and $X^* \in D(H_{-\xi})$.

(2) Since the translation of a v.a. does not alter the domain of attraction, we can assume $x_+ = 0$ without loss of generality. The tail function of X^* is now $\overline{F^*}(x) = \overline{F}(-1/x) = x^{1/\xi}L(x)$. Hence, $\overline{F^*}(x) \in \mathrm{RV}_{1/\xi}$ and $X^* \in D(H_{-\xi})$. ∎

Corollary 1 provides an asymptotic tool and it is related to an exact result in the GEV model: X has Fréchet distribution if and only if $-1/X$ has Weibull distribution with the same extreme value index, but with the sign changed. However, the corresponding result is not true in GPD, as we discuss below.

For a r.v. X, the Pickands–Balkema–DeHaan theorem shows that $X \in D(H_\xi)$ if and only if the limiting behaviour of the residual distribution of X over t, X_t, is like a GPD

with the same parameter ξ, see McNeil et al. (2005, Theorem 7.20). According to the Pickands–Balkema–DeHaan theorem, Corollary 1 can be interpreted as follows.

Corollary 2 *Let X be a continuous positive r.v. such that the limiting behaviour of the residual distribution of X over a threshold is* GPD *with parameter $\xi > 0$ ($\xi < 0$), then the limiting behaviour of the residual distribution of $-1/X$ ($-1/(X-x_+)$) over a threshold is* GPD *with parameter $-\xi$.*

Corollary 2 enables determination of the extreme value index for heavy tailed distributions using light tail models and vice versa. For instance ME-plot and CV-plot can be used to determine the extreme value index in really heavy tailed distributions, see the example 4.2 in Section 4. These asymptotic results can be improved on GPD for practical aplications.

The GPD(ξ,ψ) distributions are standardized so that all their observations take positive values. The supports of the distributions are $(0,\sigma)$, where $\sigma=\infty$ for $\xi \geq 0$ and $\sigma = \psi/|\xi|$ for $\xi < 0$. The GPD distributions can be extended to include a location parameter by $Y = X + \mu$. The behaviour of X near σ is the same as that of Y near $\sigma+\mu$. The transformation $X^* = -1/X$ is also associated with the origin at zero, but can be generalized to $Y = -1/(X+c)$, provided $c \geq 0$, or $c \leq -\sigma$, and the transformations is monotonous increasing on $(0,\sigma)$. The following result examines these transformations on GPD.

Theorem 3 *Let X be a r.v. with* GPD(ξ,ψ) *distribution in $(0,\sigma)$ and $c \geq 0$ or $c \leq -\sigma$, then $Y = -1/(X+c)$ has distribution* GPD *with location parameter if and only if $c = \psi/\xi$. Then $Z = Y + 1/c$ has* GPD$(-\xi,\xi^2/\psi)$ *distribution.*

Proof: From (1) the distribution function of Y is

$$F_Y(y) = F(x(y)) = 1 - \left(1 - \frac{\xi}{\psi}\left(\frac{cy+1}{y}\right)\right)^{-1/\xi} = 1 - \left(\frac{\psi y}{y(\psi-\xi c)-\xi}\right)^{1/\xi}, \quad (7)$$

where $-1/c < y < -1/(\sigma+c)$. The denominator of the right term of (7) is a constant if and only if $c = \psi/\xi$. In this case the distribution function of Z is

$$F_Z(z) = F_Y(y(z)) = 1 - (1-\psi z/\xi)^{1/\xi} = 1 - (1-\xi z/(\xi^2/\psi))^{1/\xi},$$

where $0 < z < \sigma_z$, $\sigma_z = \xi/\psi$ for $\xi > 0$ and $\sigma_z = \infty$ for $\xi < 0$. Hence, Z has GPD$(-\xi,\xi^2/\psi)$ distribution and Y has GPD distribution with location parameter $-1/c$. ∎

The following result establishes the essential content of the Theorem 3 avoiding the location parameter.

Corollary 4 *Let $\xi > 0$, $\psi > 0$ and $c = \psi/\xi$, then a r.v. X has* GPD(ξ,ψ) *distribution if and only if $Z = X/(c(X+c))$ has* GPD(ξ_z,ψ_z) *distribution with $\xi_z = -\xi$, $\psi_z = \xi^2/\psi$ and the support $(0,\xi/\psi)$.*

Proof: In one sense, this is proved by Theorem 3, because $c > 0$ and $Z = X/(c(X+c)) = -1/(X+c) + 1/c$.

The converse is also a consequence of Theorem 3, because the inverse of the above transformation is

$$X = c^2 Z/(1-cZ) = Z/(c_2(Z+c_2)) = -1/(Z+c_2) + 1/c_2$$

where $c_2 = -1/c = -\xi/\psi$. The support of Z is $(0,\psi_z/|\xi_z|) = (0,\xi/\psi)$ and $Z + c_2 < 0$ (equivalently $c_2 \leq -\xi/\psi$), then X is a monotonous increasing function of Z and Theorem 3 proves the result. ∎

3. Multiple threshold test

In this Section, the asymptotic distribution of the residual coefficient of variation for GPD is studied as a random process indexed by the threshold. This provides pointwise error limits for CV-plot and the multiple thresholds test that really reduce the multiple testing problem, hence, the p-values are clearly defined. These results generalize and summarize some of those of Castillo et al. (2014) on the the exponential distribution. Multiple test are often used on testing extreme value copulas, see Bahraoui et al. (2014).

Theorem 5 *Let $\{X_j\}$ be a sample of size n of iid* GPD(ξ,ψ) *distributed r.v., with $\xi < 1/4$. Then $\sqrt{n}(cv(t) - c_\xi)$, where $cv(t)$ and c_ξ were respectively defined in* (6) *and* (5)*, converges in finite-dimensional distributions to a Gaussian process with zero mean and covariance function given by*

$$\rho_0(s,t) = \exp(\min(s,t)/\psi),$$

for $\xi = 0$, and

$$\begin{aligned}\rho_\xi(s,t) = (((\psi+\xi s)/\psi)^{1/\xi})(1-\xi)^2(&6\xi^4t^2 + 12\psi\xi^3t + 8\xi^3st - 9\xi^3t^2 + 6\psi^2\xi^2 \\ &+ 8\psi\xi^2 s - 10\psi\xi^2 t - 2\xi^2 st + 3\xi^2t^2 - \psi^2\xi - 2\psi\xi s + 4\psi\xi t + \psi^2) \\ &/((1-3\xi)(1-2\xi)^2(1-4\xi)(\psi+\xi s)^2)\end{aligned}$$

for $\xi \neq 0$ and $s \leq t$.

Proof: See Appendix A. ∎

Pointwise error limits of the CV-plot under GPD follow from the next result.

Corollary 6 *Given a sample* $\{X_j\}$ *of a* GPD(ξ,ψ) *distribution* $(\xi < 1/4)$ *and a fixed threshold* t*, the asymptotic distribution of the residual* CV *is*

$$\sqrt{n(t)}(cv(t) - c_\xi) \xrightarrow{d} N(0, \sigma_\xi^2). \tag{8}$$

where c_ξ *is in* (5)*,* $n(t) = \sum_{j=1}^{n} 1_{(X_j > t)}$ *and*

$$\sigma_\xi^2 = \frac{(1-\xi)^2(6\xi^2 - \xi + 1)}{(1-2\xi)^2(1-3\xi)(1-4\xi)}.$$

Proof: The proof follows directly from Corollary 2 in Castillo et al. (2014). The asymptotic variance is given by $\sigma_\xi^2 = \rho_\xi(0,0)$, where the covariance function is in Theorem 5. The Theorem 5 can be applied to the threshold excesses $\{X_j - t \mid X > t\}$, replacing n with $n(t)$ and $cv(0)$ with $cv(t)$. From (4) the threshold excesses are again GPD with the same parameter ξ and the CV does not depend on ψ. ■

From the last result the asymptotic confidence intervals of the CV-plot for exponential distribution are obtained taking $c_0 = 1$ and $\sigma_0^2 = 1$ and for uniform distribution taking $c_{-1} = 1/\sqrt{3}$ and $\sigma_{-1}^2 = 8/45$.

Corollary 6 needs a fixed value ξ and a fixed threshold t. However, in order to have a consistent test in GPD, $\mathrm{CV}(t) = c_\xi$ must be checked for all of threshold t, in accordance with the characterization by Gupta and Kirmani (2000). For instance, the absolute value of the Student t_4 distribution has CV equal to 1 and can not be distinguished from the exponential distribution with a direct application of Corollary 6.

3.1. Exact null hypothesis test

In order to test whether a sample $\{x_j\}$ of size n of non-negative numbers, is distributed as a GPD with parameter ξ, a set of thresholds $th = \{0 = t_0 < t_1 < \cdots < t_m\}$ will be selected to test the null hypothesis

$$H_0 : \mathrm{CV}(t_k) = c_\xi, \quad k = 0, 1, \ldots, m.$$

Hence, if H_0 is accepted and m is large enough, say 20 or 50, it will be more reasonable to assume that the sample comes from a distribution GPD(ξ,ψ) than from applying Corollary 6 to a single threshold.

Let us denote $D_t(\xi) \equiv \sqrt{n(t)}(cv(t) - c_\xi)$, from Corollary 6, $D_t^2(\xi)/\sigma_\xi^2$ has asymptotic distribution χ_1^2 under the null hypothesis of GPD $(\xi < 0.25)$. Let us denote

$$T_{th}(\xi) = \sum_{k=0}^{m} D_{t_k}^2.$$

The distribution of $T_{th}(\xi)$ is independent from the scale parameter ψ under the null hypothesis of GPD. Then, its asymptotic expectation is $(m+1)\sigma_\xi^2$ and $T_{th}(\xi)/(m+1)$ is an estimator of the asymptotic variance σ_ξ^2, when ξ is known or estimated.

Given a sample $\{x_j\}$ of size n of non-negative numbers, $Q_n(p)$ denotes the inverse of the empirical distribution function,

$$Q_n(p) = \inf[x : F_n(x) \geq p]. \tag{9}$$

From a set of probabilities $\{0 = p_0 < p_1 < \cdots < p_m\}$ let $qu = \{0 = q_0 < q_1 < \cdots < q_m\}$ be the corresponding empirical quantiles of the sample, $q_k = Q_n(p_k)$, that will be used like the previous thresholds. Let us denote

$$T_{qu}(\xi) = \sum_{k=0}^{m} D_{q_k}^2.$$

$T_{qu}(\xi)$ is a multiple thresholds invariant statistic when the sample is multiplied by a positive number while maintaining the set of probabilities, since the empirical CV is invariant. This first condition ensures that the test results do not depend on units used for the observations.

A second desirable condition is to select the set of probabilities that determine the statistic $T_{qu}(\xi)$ so that the corresponding thresholds are approximately equally spaced. This can be achieved for the exponential distribution by taking $0 < p < 1$, $p_k = 1 - p^k$, $(k = 0, \ldots, m)$ and q_k as the corresponding quantiles. Since for a random variable X, distributed as an exponential with expected value μ, its quantile function is $Q(p) = \mu \log(1/p)$ and $\Pr\{X > (\mu \log(1/p))k\} = p^k$. Selecting the probabilities this way, $q_k = Q_n(p_k) \approx x_{(n-np^k)}$, $n(q_k) \approx n\,p^k$ and $T_{qu}(\xi)$ becomes

$$T_m(\xi) = n \sum_{k=0}^{m} p^k (cv(q_k) - c_\xi)^2. \tag{10}$$

In applications, given the number of single tests that will be included in the multivariant test, m, we choose the value of p, which determines the distance between the quantiles, such that $n\,p^m \approx n_s$, where n_s is the sample size such that for smaller subsamples CV is not accurate enough. Hence, given m, $p = (n_s/n)^{1/m}$ is suggested. In this paper $n_s \approx 8$ is used in numerical algorithms. Note that this way $T_m(\xi)$ depends only on ξ and m and the researcher chose only the number of thresholds used in the analysis, essentially eliminating subjectivity. These multiple thresholds tests generalize those developed by Castillo et al. (2014) for $\xi = 0$ and $p = 1/2$.

The asymptotic distribution of $T_m(\xi)$ is easily calculated from Theorem 5, following the steps suggested by Castillo et al. (2014), whenever $\xi < 0.25$. However, taking into

account the different values of the extreme value index and the diverse small sample sizes, it is easier in practice to calculate the p-value for $T_m(\xi)$ using simulation methods, which are especially simple in this case. Assuming GPD for simulations, only the sample size, the number of thresholds, m, and ξ are needed. Since the distribution does not depend on scale, parameter $\psi = 1$ will be used.

3.2. Composite null hypothesis test

In most cases the parameter ξ is unknown and its estimate should be incorporated in the statistic $T_m(\xi)$ (see the R code in Appendix B). The method for estimating ξ leads to slight variations in the statistic, leading to essentially equivalent inference whenever we use the same estimation method in simulations to obtain the p-value. The null hypothesis is now that the sample comes from a distribution in which all $(m+1)$ residual CV are equal.

$$H_0 : \mathrm{CV}(q_0) = \cdots = \mathrm{CV}(q_m), \quad k = 0, 1, \ldots, m.$$

The alternative hypothesis is that the residual CV are equal from a threshold q_r $(0 < r \leq m)$ to the threshold q_m.

The most recommended estimation method is maximum likelihood estimation (MLE), although in GPD it is only asymptotically efficient provided $-0.5 < \xi$, see Davison and Smith (1990). For this distribution, the CV is a one-to-one transformation of ξ, see (5), and the empirical CV of the residual sample, $\mathrm{CV}(t)$, provides an alternative method of estimation. It is asymptotically normal whenever $\xi < 0.25$, see Corollary 6. The multiple thresholds tests (10) suggest estimating ξ as the value such that c_ξ achieves the minimum $T_m(\xi)$, namely

$$\tilde{c}_\xi = \sum_{k=0}^{m} p^k cv(q_k) / \sum_{k=0}^{m} p^k = (1-p) \sum_{k=0}^{m} p^k cv(q_k) / (1-p^{m+1}), \tag{11}$$

and reversing (5) provides $\tilde{\xi}$; standard errors of this estimator are readily provided by simulation. The main advantage of this method is that under the alternative hypothesis it is a better estimator than CV or MLE, since the sample is only GPD over a threshold q_r. Since the main interest is in samples that are not GPD, but in the tail, and results are often used in small samples with $\xi < 0$, the estimation method (11) is included in (10). Hence, the statistics for composite null hypothesis, that only depends on m, is $T_m = T_m(\tilde{\xi})$ given by

$$T_m = n \sum_{k=0}^{m} p^k (cv(q_k) - \tilde{c}_\xi)^2. \tag{12}$$

The R code for T_m used in the algorithms is in Appendix B.

3.3. Threshold Selection Algorithms

To select the number of extremes used in applying the peaks over a high threshold method, threshold selection algorithms are developed in this section to estimate the point above which the GPD distribution can be used to estimate the extreme value index for a set of extreme events, $\{x_j\}$, of size n. For this purpose the previous statistical tests will be adapted.

Note that in the T_m calculation the number of thresholds m is the only parameter that must be fixed by the researcher. This determines the thresholds (quantiles) where the CV is calculated, $\{0 = q_0 < q_1 < \cdots < q_m\}$, which are fixed throughout the procedure. Then, by simulation of GPD, the associated p-value is calculated (running 10^4 samples). After that, we accept or reject the null hypothesis with the estimated shape parameter using all the thresholds.

If the hypothesis is rejected, the threshold excesses $\{x_j - q_1\}$ are calculated for the sub-sample $\{x_j \geq q_1\}$. The previous steps are repeated, but removing one threshold, to accept or reject the null hypothesis that the sample comes from a GPD. At every stage only statistics associated to thresholds $k = r, \ldots, m$, where $0 \leq r \leq m$, are calculated:

$$T_m^r(\tilde{\xi}) = n \sum_{k=r}^{m} p^k (cv(q_k) - \tilde{c}_\xi)^2. \tag{13}$$

In summary, the **steps of the general algorithm** are

(1) Given m find p such that $np^m \approx n_s$, where n_s is the smaller sample size used to calculate CV (here $n_s = 8$ is used, but it can be modified).

(2) Calculate $\{0 = p_0 < p_1 < \cdots < p_m\}$, where $p_k = 1 - p^k$, and $\{0 = q_0 < q_1 < \cdots < q_m\}$, where $q_k = Q_n(1 - p^k)$, $k = 1, \ldots, m$.

(3) Estimate $\tilde{\xi}$ minimizing the value of $T_m(\xi)$ with the specific values in the previous steps.

(4) Calculate by simulation of GPD the p-value associated to the minimum $T_m(\tilde{\xi})$ and accept or reject the null hypothesis with the estimated shape parameter using all the thresholds (starting with $q_0 = 0$).

(5) If the hypothesis is rejected, compute the threshold excesses $\{x_j - q_1\}$ for the sub-sample $\{x_j \geq q_1\}$ and repeat the previous steps with $\{p_1 < \cdots < p_m\}$ and $\{q_1 < \cdots < q_m\}$, to accept or reject the null hypothesis that the sample comes from a GPD, but removing a threshold.

(6) Continue the process for the next value in the index of thresholds while the hypothesis is rejected.

Several authors recommend giving a prominent role to the exponential distribution in the model GPD, see Castillo and Serra (2015). The usual method for doing this is to consider the exponential models as the null hypothesis testing against GPD, see Kozubowski et al. (2009). Alternatively, one can consider the Akaike or Bayesian information criteria for model selection, see Clauset et al. (2009). The previous algorithm can be adapted to the case when $\xi = 0$ (or any known parameter) skipping step-3.

4. Fitting GPD to empirical data

In this Section, the methods developed previously are applied to two classic examples. The first one, the euro/dollar daily exchange rates between 1999 and 2005, is analyzed in the literature using distributions with heavy tails, when these models are not appropriate. Our methodology clearly shows this fact, see Figure 1. In addition, the analysis is validated with *out of sample* observations between 2005 and 2014, including the financial crisis of 2007-08.

For the second example, the Danish fire insurance dataset, the fitted model is a highly heavy-tailed, infinite-variance model. Hence, the methodology developed in Section 2 is needed to avoid unconscientious use of measures of variation that can lead to wrong connclusions Albrecher et al. (2010).

4.1. EUR/USD daily exchange rates

Gomes and Pestana (2007), introduce a new semi-parametric quantile estimation method based on an adequate bias-corrected Hill estimator. To illustrate their method it is applied to the analysis of log-returns of the euro/dollar (EUSD) daily exchange rates, from January 4, 1999 through November 17, 2005 (1,794 observations). The paper gives the estimations of the tail index $\hat{\xi} = 0.279$ (Hill estimator) and $\hat{\xi} = 0.247$ (bias-corrected) for the positive log returns of EUSD.

It should be mentioned that the Hill method always provide estimators with $\xi > 0$, as in this case. Hence, previously, this hypothesis has to be checked. Figure 1 shows the CV-plots (6) for the positive and negative (with the sign changed) log-returns of EUSD. In both cases there is empirical evidence that the residual CV is lower than 1. Since in GPD CV < 1 is equivalent to $\xi < 0$, this suggests light tails where some researchers assume heavy tails. This qualitative approach can be confirmed with the multiple thresholds tests.

Applying T_m, where $m = 20$, to the 900 positive log-returns of EUSD, the estimate of CV given by (11) is $\tilde{c}_\xi = 0.861$, which corresponds to $\tilde{\xi} = -0.174\ (0.031)$ assuming GPD. The statistic is $T_m = 6.435$ with a p-value of 0.421. Hence, the null hypothesis of GPD is not rejected for the entire sample and the previous estimation of ξ is validated (in the first step of the algorithm). The result is similar for the 874 negative log-returns and $m = 20$. Here $\tilde{c}_\xi = 0.868$ is obtained, which corresponds to $\tilde{\xi} = -0.163\ (0.032)$

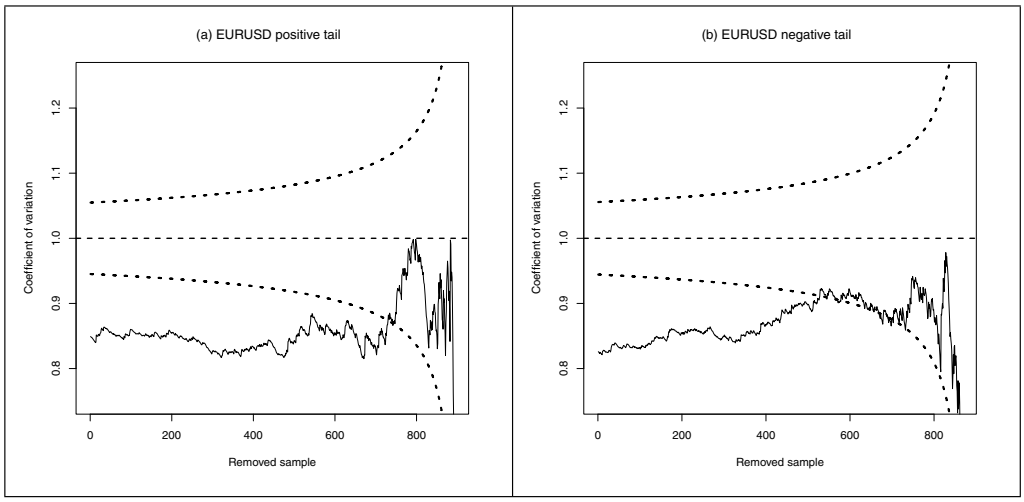


Figure 1: *Residual empirical* CV *for positive and negative tails of EUR/USD daily exchange rates from January 4, 1999 through November 17, 2005. The dotted lines correspond to the asymptotic confidence intervals (*90%*) under exponentially* (CV $= 1$).

assuming GPD. The statistic is $T_m = 6.120$ with a p-value of 0.449. The null hypothesis of GPD is not rejected for the entire sample and the previous estimation of ξ is also validated. The results are fairly coincidental for $m = 10$, 20, 30 and 50, in both cases.

Despite the evidence of light tails on this previous sample, it is better to follow the recommendation of testing exponentiality at the tails. This approach is also compatible with the assumption of heavy tails in a wider sense ($\xi \geq 0$) including the exponential as a boundary point, see Castillo and Serra (2014). Applying T_m to all positive log-returns of EUSD, with $m = 20$ and $\xi = 0$, the null hypothesis of exponentiality is rejected (p-value 0.01). Taking $p = (n_s/n)^{1/m}$ the sample is reduced by $(1-p) = 21\%$ in each step, then for thresholds 0.134, 0.249 and 0.376, the null hypothesis is also rejected (p-values 0.017,0.026, and 0.057). Finally, exponentiality is not rejected over the threshold $t_p = 0.516$ with a p-value 0.133. For negative log-returns with $m = 20$ and $\xi = 0$, the exponentiality is rejected in the first three steps and not rejected over $t_n = 0.411$ with a p-value 0.126.

The main objective of statistics of extremes lies in the estimation of quantities related to extreme events that may happen in the future. Hence, the real challenge is to compare the results in out of sample observations. To this end, from the previous analysis, the *value at risk* at a level α (VaR$_\alpha$), the quantile so that the chance of exceedance of that value is equal to α, is estimated by the *peak-over-threshold* method, using the empirical sample in the interval $(0,t)$, up to the estimated threshold, and the exponential distribution over threshold t. For $\alpha = 0.05$, 0.01 and 0.001, the quantiles of positive log-returns of EUSD are 1.316, 1.937 and 2.824; for the negative log-returns they are 1.352, 2.010 and 2.950.

Then, daily exchange rates, from November 18, 2005 through January, 14, 2014 (2,128 observations), including the financial crisis of 2007-08, are used as out of sample observations to assessing the predictive ability of the estimation of quantiles under the first dataset.

Using these 2,128 out of sample observations (the second dataset), the number of empirical exceedances of the last VaR_α estimations (under the first dataset, at 5%, 1%, 0.1%) are 42, 13, and 2, for the 1,080 positive log-returns (expected values 54.0, 10.8 and 1.1); and 47, 11 and 0, for the 1023 negative log-returns (expected values 51.2, 10.2 and 1.0). These results are fairly satisfactory and it can be concluded that the EUR / USD exchange has daily log-returns with exponential tails, including the financial crisis of 2007-08.

4.2. Danish fire insurance data

An interesting aspect of this article is the combination of the results of sections 2 and 3 when applying the peaks over threshold technique for tails in any maximum domain of attraction, even without finite moments. This approach is illustrated here using a classical example analyzed in several books and articles.

The Danish fire insurance data are a well-studied set of losses to illustrate the basic ideas of extreme value theory. The dataset consists of $2,156$ fire insurance losses over one million Danish kroner from 1980 to 1990 inclusive, see Embrechts et al. (1997, Example 6.2.9), Resnick (1997) and McNeil et al. (2005, Example 7.23).

In this example the authors agree to assume iid observations and a heavy tailed model. They also agree to set the threshold at $t = 10$ million Danish kroner, the exceedances over the threshold, denoted $\{x_j\}$, are $n_{10} = 109$. Fitting a GPD to $\{x_j\}$ by MLE, the parameter estimates in McNeil et al. (2005) are $\hat{\xi} = 0.50$ and $\hat{\psi} = 7.0$ with standard errors 0.14 and 1.1, respectively. Thus the fitted model is a very heavy-tailed, infinite-variance model and the methods in Section 3 cannot be applied directly. However, they can be used through the results shown in Section 2.

First of all, let us suppose we want to use CV to check whether the above data correspond to a GPD distribution with the estimated extreme value index. Applying Theorem 3 with $c = \hat{\psi}/\hat{\xi} = 14$, let $z_j = -1/(x_j + c) + 1/c$ be, then the set $\{z_j\}$ has light tails and the same extreme value index with the sign changed, provided that the estimated parameters are the true parameters. The CV of $\{z_j\}$ is $cv = 0.697$ which provides a new estimation of ξ, solving (5) by $\xi_z = (cv^2 - 1)/(2cv^2) = -0.530$, then, according to Theorem 3, $\tilde{\xi} = -\xi_z = 0.53$, not far from the parameter estimation in McNeil et al. (2005) , 0.50, since his standard error was 0.14. Alternatively, the multiple thresholds statistic T_m, from (13), can be used to check $\xi = 0.5$. The corresponding CV under GPD is $c_\xi = 0.707$. Taking $m = 20$, we get $T_m = 4.89$ with a p-value 0.421 (by simulation with 10^4 samples), not rejecting the null hypothesis.

Now consider the problem of choosing the threshold to estimate the extreme value index. In this example, most researchers use a visual observation of the ME-plot on

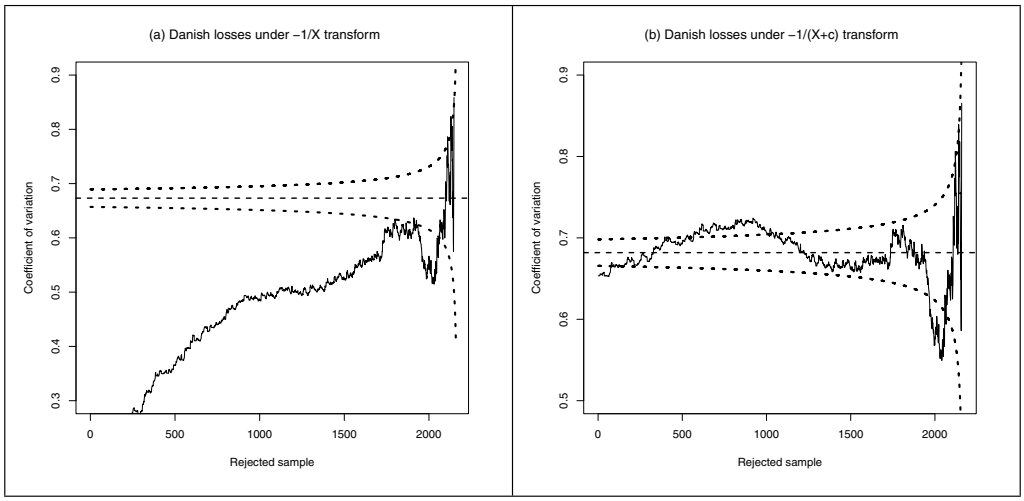


Figure 2: *Residual empirical* CV *for The Danish fire insurance losses under transformation of the data. (a): Dataset, transformed by* $X^* = -1/X$ *. (b): Dataset, transformed by* $Z = -1/(X + \psi/\xi)$*. The dotted lines correspond to the asymptotic confidence intervals (*90%*) under the estimated parameter, the dashed line is its* CV.

the full Danish dataset. The algorithm in Section 3.3 with the transformations from Section 2, comes to similar solutions automatically and opens up new perspectives.

Figure 2 shows the CV-plots of the full Danish dataset, transformed according to the Corollary 2, plot (a), and Theorem 3, plot (b). The first, corresponding to the transformation $X^* = -1/X$, shows an increasing CV and the second, corresponding to $Z = -1/(X+c)+1/c$, shows a stabilized CV close to a constant, indicating that the original dataset is close to a GPD, which is also shown by ME-plot.

Applying the algorithm of Section 3.3 with $m = 20$ after transformation X^*, constant residual CV is rejected in the first 11 steps (each one reduces the sample size by $(1-p) = 24\%$). Step 12, for the last 106 observations, accepts constant residual CV (p-value $= 0.269$) with estimates $\tilde{c}_\xi = 0.673$ and $\tilde{\xi} = 0.603$. The estimated threshold is approximately the same ($t = 10.2$ instead of 10), while the extreme value index is different but within the confidence interval.

The algorithm in Section 3.3, with $m = 20$ after transformation Z with $c = = 0.932/0.611 = 1.524$, rejects constant residual CV in the first three steps. Step 4, for the last 951 observations, accepts constant residual CV (p-value $= 0.167$) with estimates $\tilde{c}_\xi = 0.675$ and $\tilde{\xi} = 0.599$. The number of observations is much higher, the extreme value index being very close to that obtained with the transformation X^* and within the confidence interval. The p-value remains similar in the following steps up until the 12th, where it jumps up to 0.474. The number of observations is again 106 and the estimation $\tilde{\xi} = 0.548$, close to 0.50.

The conclusions from using the new methodology to analyze this dataset are the following. First, the results obtained by previous investigators are validated, in particular

GPD can be accepted with parameter $\xi = 0.5$, for the 109 larger observations see McNeil et al. (2005). This also shows the consistency of the presented methodology with other common techniques.

Moreover, from examining the extreme value index it is now known that for the 951 larger observations GPD can also be accepted, where the MLE parameter estimate is $\hat{\xi} = 0.680$, with standard error 0.055 ($\tilde{\xi} = 0.599$ obtained by T_m is within the confidence interval). The estimated extreme value index is now much more accurate because the sample size is much larger. We also note that the tails are heavier than was assumed, which means that higher risks should be considered.

Acknowledgement

This work was supported by the Spanish Ministry of Economy and Competitiveness under Grants: Applied Stochastic Processes, MTM 2012-31118 and Statistical modelling of environmental, technological and health risks, MTM2015-69493-R.

References

Albrecher, H., Ladoucette, S. and Teugels, J. (2010). Asymptotics of the sample coefficient of variation and the sample dispersion. *Journal of Statistical Planning and Inference*, 140, 358–368.

Bahraoui, Z., Bolance, C. and Perez-Marin, A. (2014). Testing extreme value copulas to estimate the quantile. *SORT*, 38, 89–102.

Castillo, J. D., Daoudi, J. and Lockhart, R. (2014). Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics*, 41, 382–393.

Castillo, J. D. and Serra, I. (2015). Likelihood inference for generalized Pareto distribution. *Computational Statistics and Data Analysis*, 83, 116–128.

Clauset, A., Shalizi C. and Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review*, 51, 661–703.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B*, 393–442.

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

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, 180–190, Cambridge University Press.

Ghosh, S. and Resnick, S. (2010). A discussion on mean excess plot. *Stochastic Processes and their Applications*, 120, 1492–1517.

Gilleland, E., Ribatet, M. and Stephenson, A. G. (2013). A software review for extreme value analysis. *Extremes*, 16, 103–119.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44, 423–453.

Gomes, M. I. and Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: A review. *International Statistical Review*, 83, 263–292.

Gomes, M. I. and Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal of the American Statistical Association*, 102, 280–292.

Gupta, R. C. and Kirmani, S. N. U. A. (2000). Residual coefficient of variation and some characterization results. *Journal of Statistical Planning and Inference*, 91, 23–31.

Kozubowski, T. J., Panorska, A. K., Qeadan, F., Gershunov, A. and Rominger, D. (2009). Testing exponentiality versus Pareto distribution via likelihood ratio. *Communications in Statistics-Simulation and Computation* 38, 118–139.

McNeil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance, New Jersey (2005).

R Core Team (2010). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*. Vienna, Austria.

Resnick, S. I. (1997). Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin*, 27, 139–151.

Appendix A: Proof of Theorem 5

The residual CV in (3) can be expressed in terms of the moments of the truncated distribution. Let X be a continuous non-negative r.v. with distribution function $F(x)$, let $X1_{(X>t)}$ denote the r.v. which is equal to X if $X > t$ and equal to zero otherwise. Let $\mu_0(t) = \Pr\{X > t\}$ and $\mu_k(t) = E[X^k 1_{(X>t)}]$, $k > 0$. Throughout this paper we assume that $\mu_0(t) > 0$ for all t. Note that

$$\mu_k(t) = \mu_0(t)E(X^k \mid X > t), \tag{14}$$

hence, in particular, the residual mean and the residual variance are

$$M(t) = \mu_1(t)/\mu_0(t) - t,\ V(t) = \mu_2(t)/\mu_0(t) - (\mu_1(t)/\mu_0(t))^2,$$

and the residual CV

$$\mathrm{CV}(t) = \sqrt{\mu_2(t)\mu_0(t) - \mu_1(t)^2}/(\mu_1(t) - t\mu_0(t)).$$

Let $\{X_j\}$ be a sample of independent and identically distributed (iid) r.v.s of size n. Let $n(t) = \sum_{j=1}^{n} 1_{(X_j>t)}$ be the number of exceedances over a threshold, t. By the law of large numbers, $n(t)/n$ converges to $\mu_0(t)$. The *empirical* CV *of the conditional exceedances* is given by

$$cv(t) = cv_n(t) = \frac{n(t)}{\sum_{j=1}^{n}(X_j - t)1_{(X_j>t)}} \times \left[\frac{\sum_{j=1}^{n} X_j^2 1_{(X_j>t)}}{n(t)} - \left(\frac{\sum_{j=1}^{n} X_j 1_{(X_j>t)}}{n(t)}\right)^2\right]^{1/2}, \tag{15}$$

see (6) for a simpler expression when the r.v. are observed.

Then $cv_n(t)$ is a consistent estimator of $\mathrm{CV}(t)$ by the law of large numbers, assuming F has a finite second moment.

From Theorem 1 in Castillo et al. (2014),

$$\sqrt{n}(cv_n(t) - c_\xi) = a'(t)W(t) + O_p(1/\sqrt{n})$$

where

$$\mathrm{cov}(W(s), W(t)) \equiv M(s,t) = (\mu_{i+j}(t) - \mu_i(s)\mu_j(t))_{i,j=0,1,2},$$

and $\mu_k(t)$ are the moments of the truncated distribution (14).

$$a'(t) = (\mu_0(\mu_1 - t\mu_0), 2\mu_0(t\mu_1 - \mu_2), (-2t\mu_1^2 + t\mu_0\mu_2 + \mu_1\mu_2)) \\ /(2(\mu_1 - t\mu_0)^2\sqrt{\mu_2\mu_0 - \mu_1^2}),$$

where for simplicity dependence on t is dropped for $\mu_k = \mu_k(t)$ in the last expression. Then, the covariance function is

$$\rho_\xi(s,t) = a(s)'M(s,t)\, a(t),$$

using the conditional moments of GPD and some algebra, the result of the theorem holds.

Appendix B: R code for T_m

The following R code for T_m is used in the algorithms, see R Development Core Team (2010). See Gilleland et al. (2013) for a review of the currently available software on the generalized Pareto distribution and estimation of the extremal index.

```
#Statistic Tm of a sample given the number of thresholds m.
Tm<-function(m,sample){sam<-sample-min(sample);
     n<-length(sam);ns<-8;
     p<-round(exp(log(ns/n)/m),digits=2);
     Ws<-Ps<-Qs<-Cs<-numeric(m+1);
     for(k in 1:(m+1)){Ws[k]<-p^(k-1)};
     Ps<-1-Ws;Qs<-as.vector(quantile(sam,Ps));
     for(k in 1:(m+1))
     {Cs[k]<-sd(sam[sam>=Qs[k]]-Qs[k])/mean(sam[sam>=Qs[k]]-Qs[k])};
     cx<-(1-p)*sum(Ws*Cs)/(1-p^(m+1));xi<-(cx^2-1)/(2*cx^2);
     tm<-n*sum(Ws*(Cs-cx)^2);list(CV=cx,Tm=tm,Xi=xi)}
```

SORT 40 (2) July-December 2016, 321-348

Using robust FPCA to identify outliers in functional time series, with applications to the electricity market

Juan M.Vilar, Paula Raña and Germán Aneiros

Abstract

This study proposes two methods for detecting outliers in functional time series. Both methods take dependence in the data into account and are based on robust functional principal component analysis. One method seeks outliers in the series of projections on the first principal component. The other obtains uncontaminated forecasts for each data set and determines that those observations whose residuals have an unusually high norm are considered outliers. A simulation study shows the performance of these proposed procedures and the need to take dependence in the time series into account. Finally, the usefulness of our methodology is illustrated in two real datasets from the electricity market: daily curves of electricity demand and price in mainland Spain, for the year 2012.

MSC (2010): 62M10, 62H25, 62M20.

Keywords: Functional data analysis, functional principal component analysis, functional time series, outlier detection, electricity demand and price.

1. Introduction

Functional data analysis (FDA) is a branch of Statistics that analyses data providing information about curves, surfaces or any other mathematical object varying over a continuum. The continuum is often time, but it may also be spatial location, wavelength, etc. These curves are defined by a functional form and are called functional data.

Over the last two decades there has been growing research on FDA and most statistical techniques have been generalized to the functional context. This includes linear regression models (Cardot, Ferraty, and Sarda, 1999; Li and Hsing, 2007; García-Portugués, González-Manteiga, and Febrero-Bande, 2014), nonparametric smoothing

Departamento de Matemáticas, Universidade da Coruña, Spain.
juan.vilar@udc.es, paula.rana@udc.es, ganeiros@udc.es
Received: May 2015
Accepted: June 2016

methods (Ferraty and Vieu, 2002; Delsol, Ferraty, and Vieu, 2011; Shang, 2014), classification (Cuevas, Febrero, and Fraiman, 2007; Baíllo, Cuesta-Albertos, and Cuevas, 2011; Sguera, Galeano, and Lillo, 2014), dimension reduction (Boente and Fraiman, 2000; Hall, Müller, and Wang, 2006) and bootstrap methods (González-Manteiga and Martínez-Calvo, 2011; Ferraty, van Keilegom, and Vieu, 2012). In addition, FDA has been successfully applied in a wide range of fields such as climatology (Besse, Cardot, and Stephenson, 2000), chemometrics (Ferraty and Vieu, 2002), environmetrics (Aneiros-Pérez et al., 2004), demography (Hyndman and Ullah, 2007), social sciences (Ocaña, Aguilera, and Escabias, 2007) and the electricity market (Aneiros et al., 2013 and 2016). Of course, the above references form a non-exhaustive list of recent methodological and practical presentations related to FDA. See the monographs by Ramsay and Silverman (2005) and Ferraty and Vieu (2006) for parametric and nonparametric methods, respectively. For a recent state of the art on FDA, see Ferraty and Romain (2011), Horváth and Kokoszka (2012) and Cuevas (2014).

Procedures for detecting functional outliers have also been proposed over recent years despite the fact that the functional nature of the data makes outliers hard to detect. As a matter of fact, a rigorous definition of functional outlier remains to be given. Throughout this paper, we define a functional outlier as an observation (functional datum) that has been generated by a stochastic process with a distribution different from the vast majority of the remaining observations, which are assumed to be identically distributed (note that this is the definition given in Febrero, Galeano, and González-Manteiga, 2008; Hyndman and Shang, 2010). The first papers that have addressed outlier identification in the context of functional data are Hyndman and Ullah (2007) and Febrero, Galeano, and González-Manteiga (2007, 2008). Hyndman and Ullah (2007) proposed a method for robust estimation of functional principal components, which is the basis of their methodology for forecasting functional time series. As a by-product, they constructed a method for detecting outliers based on the integrated squared error between each functional datum and its projection into a given number of robust principal components. The procedure in Febrero, Galeano, and González-Manteiga (2007) (Febrero, Galeano, and González-Manteiga, 2008) performs a distance-based (depth-based) test statistic for each curve, where the critical value is obtained with a bootstrap method. Several procedures for detecting outliers in functional data have been proposed from these works. They are generally based on functional principal components analysis (Hyndman and Shang, 2010; Sawant, Billor, and Shin, 2012; Yu, Zou, and Wang, 2012), functional depths (Sun and Genton, 2011; Gervini, 2012; Arribas-Gil and Romo, 2014) or random projections (Fraiman and Svarc, 2013). All of these papers deal with independent functional data.

This paper addresses the problem of outlier detection in functional time series coming from a real-valued continuous time stochastic process. Specifically, to define the functional time series, $\{\chi_i\}_{i=1}^n$, which are going to be used along this paper, we consider a real-valued continuous time stochastic process $\{\chi(t)\}_{t\in R}$. Then, we assume that such process is seasonal with seasonal length τ and we regard that it is observed on the in-

terval $(a,b]$ with $b = a + n\tau$. We define the functional time series $\{\chi_i\}_{i=1}^n$ in terms of $\{\chi(t)\}_{t\in R}$ as:

$$\chi_i(t) = \chi(a + (i-1)\tau + t) \text{ with } t \in [0,\tau).$$

As in the case of finite-dimensional data, dependence affects functional outlier detection (see Raña, Aneiros, and Vilar, 2015). This is clearly illustrated in Figure 1, which shows the sequential graph of a simulated functional time series contaminated with four outliers (left panel) and the corresponding curves (right panel). Looking closely at the left panel in Figure 1, one may suspect the possible presence of such four outliers; however, the same may not be said when observing the right panel in Figure 1. Local trends induced from the dependence structure could mask the presence of outliers; so, in functional time series, an observation could be an outlier despite being inside the range of the vast majority of the data. It therefore seems reasonable to believe that this kind of outlier cannot be detected by applying outlier detection procedures designed for independent data. To the best of our knowledge, the only paper that has addressed outlier detection in functional time series is Raña, Aneiros, and Vilar (2015). These authors suggested adapting the procedure in Febrero, Galeano, and González-Manteiga (2008) to the functional time series setting by considering bootstrap techniques that take into account the dependence between functional data (instead of standard bootstrap).

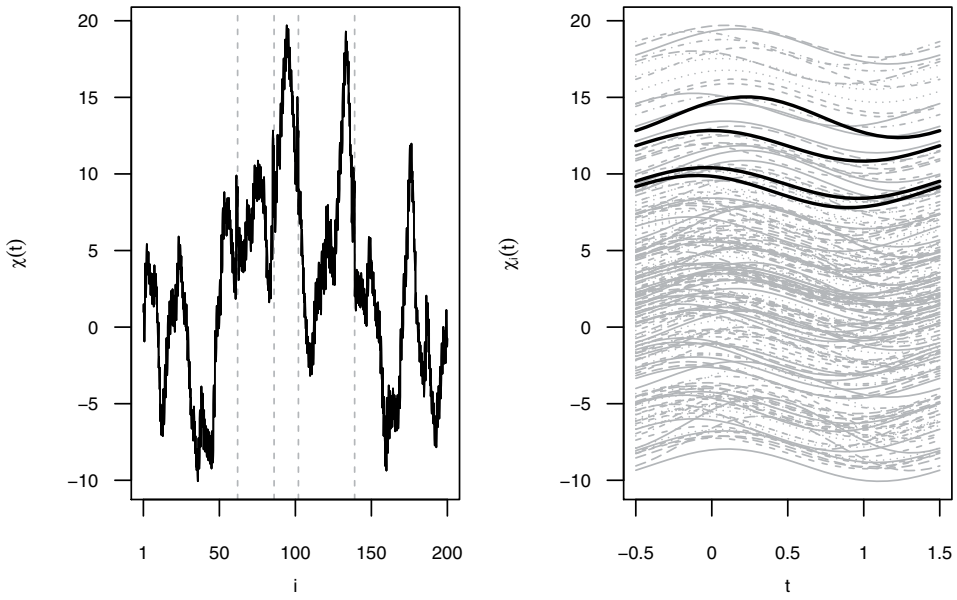


Figure 1: *Left panel: functional time series (i denotes the temporal index) contaminated with four outliers; the vertical dashed lines indicate the positions where the outliers emerged. Right panel: the corresponding curves $\chi_i(t)$ (the black curves are the outliers).*

This article proposes two procedures for detecting outliers in functional time series. Both methods are based on robust functional principal component analysis and use ideas developed by Hyndman and Ullah (2007) to forecast functional time series.

The remainder of this paper is organized as follows. Section 2 presents basic ideas on principal component analysis. The proposed methodology is described in Section 3. Its behavior is evaluated and compared with other approaches through a simulation study in Section 4. In Section 5 our methods are applied to electricity demand and price curves. Finally, Section 6 concludes with a discussion.

2. Functional principal component analysis

Since our proposed procedures for detecting functional outliers will be constructed based on functional principal component analysis (FPCA), this section presents a brief review on this topic. The interested reader can find a more complete review in Hall (2011). Without loss of generality, we assume that the considered functional random variable has zero mean.

Principal component analysis (PCA) is a standard approach to explore variability in multivariate data, $\mathbf{X} \in \mathbb{R}^d$. This approach specifies the d directions, $\{\mathbf{v}_k\}_{k=1}^d \in \mathbb{R}^d$, that maximize the variance along each component, subject to the orthonormal condition. Reducing the dimension is especially important when data belong to infinite dimensional spaces, this being the case of functional data. In this article we focus on curves observed in $[a,b]$ $(-\infty < a < b < \infty)$ and square integrable. Then, if χ denotes a functional random variable, the aim of FPCA is to find the functions $\phi_k : [a,b] \to \mathbb{R}$ such that the variance of

$$\beta_k = \int_a^b \phi_k(t)\chi(t)dt \tag{1}$$

is maximized subject to the constraints

$$\int_a^b \phi_k^2(t)dt = 1 \text{ and } \int_a^b \phi_k(t)\phi_j(t)dt = 0 \ (k \neq j). \tag{2}$$

The functional principal components $\phi_k(\cdot)$ can also be defined as the orthonormal functions verifying

$$\int_a^b \mathbf{C}(t,s)\phi_k(s)ds = \lambda_k\phi_k(t) \ (t \in [a,b],\ k = 1,2,\dots), \tag{3}$$

where $\mathbf{C}(t,s)$ denotes the covariance between $\chi(t)$ and $\chi(s)$. Finally, dimension reduction is performed by considering the approximation

$$\chi(\cdot) \approx \sum_{k=1}^{K} \beta_k \phi_k(\cdot), \tag{4}$$

where $K < \infty$ and $\sum_{k=1}^{K} \lambda_k$ is close to $\sum_{k=1}^{\infty} \lambda_k$ (we have assumed that $\lambda_k > \lambda_{k+1}$, $k = 1,2,\ldots$). For details, see e.g. Ramsay and Silverman (2005).

Functional principal components, $\phi_k(\cdot)$, depend on the unknown covariance operator $\mathbf{C}(\cdot,\cdot)$. Assuming that one has observations $\{\chi_i\}_{i=1}^{n}$ identically distributed from the functional random variable χ, estimates for $\phi_k(\cdot)$ can be obtained by using

$$\widehat{\mathbf{C}}(t,s) = \frac{1}{n}\sum_{i=1}^{n}(\chi_i(t) - \overline{\chi}(t))(\chi_i(s) - \overline{\chi}(s)), \text{ where } \overline{\chi}(t) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(t),$$

instead of $\mathbf{C}(t,s)$ in (3). See Horváth and Kokoszka (2012) for the consistency of $\widehat{\mathbf{C}}$ and of the corresponding eigenfunctions and eigenvalues, under either independent curves or weakly dependent functional time series.

It is worth noting that, apart being used for dimension reduction, FPCA can also be used as a tool for outlier detection. Nevertheless, as noted in the previous paragraph, the estimation of functional principal components is based on the estimated covariance operator $\widehat{\mathbf{C}}(\cdot,\cdot)$, which is known to be sensitive to outliers. Thus, if the goal is to construct an approach based on principal components to identify functional outliers, robust FPCA should be considered. In this way, Hyndman and Ullah (2007) propose estimating the functional principal components by means of the functions $\widehat{\phi}_k(\cdot)$ that maximize the variance of the scores

$$z_{i,k} = w_i \int_a^b \phi_k(t)\chi_i(t)dt \tag{5}$$

subject to the constraints (2). The weights w_i are computed as

$$w_i = \begin{cases} 1 & \text{if} \quad v_i < S + \lambda\sqrt{S} \\ 0 & \quad \text{otherwise} \end{cases}$$

where

$$v_i = \int_a^b (\chi_i(t) - \sum_{k=1}^{K} \widetilde{\beta}_{i,k}\widetilde{\phi}_k(t))^2 dt \tag{6}$$

with $\widetilde{\phi}_k(\cdot)$ being initial (highly robust) projection-pursuit estimates of $\phi_k(\cdot)$ obtained from the RAPCA algorithm (see Hubert, Rousseeuw, and Verboven, 2002) considering equal weights w_i in (5), while $\widetilde{\beta}_{i,k} = \int_a^b \widetilde{\phi}_k(t)\chi_i(t)dt$. In addition, S is the median of

$\{v_1,\ldots,v_n\}$ and $\lambda > 0$ is a tuning parameter to control the degree of robustness. Once the robust estimates $\widehat{\phi}_k(\cdot)$ are obtained, the coefficients corresponding to the curve χ_i are constructed as

$$\widehat{\beta}_{i,k} = \int_a^b \widehat{\phi}_k(t)\chi_i(t)dt. \tag{7}$$

As a by-product, Hyndman and Ullah (2007) proposed an outlier detection method (the ISE method): the curve χ_i is detected as outlier if $w_i = 0$. For other FPCA-based procedures to identify outliers, see e.g. Hyndman and Shang (2010) and Sawant, Billor, and Shin (2012).

3. Outlier detection in functional time series

As noted in Section 1, the dynamics in the data should be taken into account to detect outliers in functional time series. In other words, methods based only on the set of curves and not on the dependence structure among them, cannot detect the outliers that remain hidden among all of the curves (note that these outliers make sense in time series).

We propose two procedures to detect outliers in functional time series. Both proposals are based on the suggestions of Hyndman and Ullah (2007) for obtaining robust forecasting in functional time series. We establish our methods in the following subsections.

3.1. Method based on projections

Our first proposal detects outliers on the first K robust principal component scores and then map the detected outliers into the functional space.

Specifically, the method based on projections proposes to detect outliers in functional time series with the following algorithm:

- Step 1. Perform robust FPCA and construct the corresponding time series of coefficients $\{(\widehat{\beta}_{i,1},\ldots,\widehat{\beta}_{i,K})\}_{i=1}^n$.
- Step 2. Identify outliers in the series constructed in Step 1 by means of a time-series outlier detection method.
- Step 3. Establish the set of outliers as $\mathscr{O} = \{\chi_i : i \in \mathscr{I}\}$, where $\mathscr{I} = \{i : (\widehat{\beta}_{i,1},\ldots,\widehat{\beta}_{i,K})$ was detected as outlier in Step 2$\}$.

The key points in this method are the use of robust FPCA together with procedures to detect outliers in time series. Given that the estimated functional principal components $\widehat{\phi}_k$ are not affected by the outliers, the corresponding projections $\widehat{\beta}_{i,k}$ reflect the main

features of the datum χ_i. Thus, we may expect that if a curve is an outlier, its projection on the directions of maximum variance (the first principal components) will also be an outlier.

In practice, both a robust FPCA and a time-series outlier detection method must be fixed to implement our proposal. On the one hand, the robust FPCA proposed in Hyndman and Ullah (2007) could be considered (for a brief exposition, see last paragraph in Section 2). On the other hand, it is worth being noted that the principal component scores $\widehat{\beta}_{i,k}$ and $\widehat{\beta}_{i,l}$ are uncorrelated for $k \neq l$. Thus, as suggested in Hyndman and Ullah (2007), each univariate time series $\{\widehat{\beta}_{i,k}\}_{i=1}^n$, $k = 1, \ldots, K$, can be studied independently. In this way, we propose to use some univariate time-series outlier detection method to identify outliers in each of such scalar time series, and, in Step 2, consider that $(\widehat{\beta}_{i,1}, \ldots, \widehat{\beta}_{i,K})$ is an outlier if some of its components was detected as outlier in the univariate study (for the univariate time-series outlier detection method based on ARIMA models used in this paper, see Section 11.2 in Cryer and Chan, 2008). Another alternative would be to use a multivariate time-series outlier detection method (see, for instance, Tsay, Peña, and Pankratz, 2000).

3.2. Method based on errors

Unlike the previous method, our second procedure takes the whole of each curve into account. Using techniques for robust forecasting in functional time series, it constructs a non-contaminated version for each curve, which is compared with the corresponding original curve. A curve is considered an outlier if it is substantially different from its uncontaminated version.

Specifically, this method proposes to detect outliers in functional time series with the following algorithm:

- Step 1. Perform robust FPCA and construct the corresponding time series of coefficients $\{(\widehat{\beta}_{i,1}, \ldots, \widehat{\beta}_{i,K})\}_{i=1}^n$.
- Step 2. Fit a robust model to the time series constructed in Step 1.
- Step 3. Obtain the fitted values $\{(\widehat{\beta}^*_{i,1}, \ldots, \widehat{\beta}^*_{i,K})\}_{i=1}^n$ from the model constructed in Step 2.
- Step 4. Construct the residual curves $\{\chi_i - \widehat{\chi}_i\}_{i=1}^n$ and compute some kind of norm $\{u_i\}_{i=1}^n$ for such curves. We have denoted

$$\widehat{\chi}_i = \sum_{k=1}^{K} \widehat{\beta}^*_{i,k} \widehat{\phi}_k.$$

- Step 5. Identify "abnormally high values" in $\{u_i\}_{i=1}^n$, and set the functional outliers as $\mathscr{O} = \{\chi_i : i \in \mathscr{J}\}$, where $\mathscr{J} = \{i : u_i \text{ was identified as abnormally high}\}$.

As in the method based on projections, robust FPCA plays a main role (together with robust modelling of nonfunctional time series). Note that, because the fitted values obtained in Step 3 are not contaminated by the outliers, $\widehat{\chi}_i$ can be seen as the "expected value" of the functional time series at time i when no contamination is present. Thus, an abnormally high value for u_i suggests that χ_i is an outlier.

Note that our proposal can be seen as an extension in different ways of the ISE method proposed in Hyndman and Ullah (2007) (for a brief exposition, see last paragraph in Section 2). Clearly, our main contribution is related to the dependence in the functional time series: our procedure takes the dependence among the sample into account (see Step 3) to construct the coefficients associated to each functional data χ_i ($\widehat{\beta}^*_{i,k}$ in Step 4 above), while the method in Hyndman and Ullah (2007) does not do so (see $\widetilde{\beta}_{i,k}$ in (6)). As it will be clearly shown in the simulation study to be presented in the next Section 4, this seemingly minor modification will greatly improve the performance of the method when applied on functional time series.

In practice, several choices must be done to implement our algorithm. As in the method based on projections, we suggest to consider the robust FPCA proposed in Hyndman and Ullah (2007) and construct univariate models instead of multivariate ones (see Section 3.1). Specifically, we suggest to fit, for each series $\{\widehat{\beta}_{1,k},\ldots,\widehat{\beta}_{n,k}\}$, $k=1,\ldots,K$, the univariate robust ARIMA models studied in Cryer and Chan (2008) (for details, see Section 11.2 in the cited reference). As for the norm to be used to construct the set $\{u_i\}_{i=1}^n$ in Step 4, one might consider, for instance, the L_1-norm or the L_2-norm (or even the squared of the L_2-norm, as in Hyndman and Ullah, 2007). Finally, we suggest to consider that u_i is high enough to be considered as abnormally high if $u_i > q_{0.75} + 1.32(q_{0.75} - q_{0.25})\}$ (q_p denotes the quantile of order p of $\{u_1,\ldots,u_n\}$). Actually, this is the rule given by the classical boxplot; that is, under normality, the probability of detecting no outliers is 0.993, when no outliers are actually present (note that the usual constant factor 1.5 was changed to 1.32 because low values are not considered outliers).

3.3. Tuning parameter

As common to all FDA procedures using FPCA, the proposed methods depend on the number of principal components considered, K. In practice, the value of K must be specified. Hyndman and Ullah (2007) suggest choosing K to minimize the integrated squared forecast error (ISFE), while Hyndman and Booth (2008) find that the forecasts are insensitive to the choice of K, provided K is large enough. Then, Hyndman and Booth (2008) recommend using a value K that is apparently larger than actually required by the components. In the cited works of Hyndman and Ullah (2007) and Hyndman and Booth (2008), and also in Liebl (2013), in the study of different applications using this technique they use a value of K which explains, at least, 98% of the variability.

We have carried out sensitivity studies for the values of K in our methods, using the dependent simulated data considered in the next section (Models 1, 2 and 3). On the

one hand, our findings agree with the general suggestion given in Hyndman and Booth (2008): to consider a larger than necessary value K (for instance, a value explaining at least 98% of the variability). On the other hand, to detect "shape outliers" (that arise when they are within the range of the rest of the data but differ from them in shape; see Hyndman and Shang, 2010) by means of the method based on projections (PB), the recommendation is to select a value K even higher (for instance, explaining at least the 99.9% of variability). To justify this very high value we argue that (i) the PB method only uses scores (and not the whole of the curve), (ii) the first scores inform about the possible presence of "magnitude outliers" (that arise when they lie outside the range of the majority of the data; see Hyndman and Shang, 2010) and (iii) the scores of higher order inform about the possible presence of shape outliers.

4. Simulation study

A simulation study was conducted to compare the performance of our methods with other methods available in the statistical literature.

On the one hand, three main models were constructed to generate functional time series. They are the superposition of a deterministic signal and random noise. Noise in main Models 1, 2, and 3 was the superposition of a scalar AR(1) process and functional AR(1)-, MA(1)- and ARMA(1,1)-type processes, respectively. On the other hand, another main model (Main Model 0) was constructed in the same way, but considering independent noise instead of dependent one. Note that main Models 1, 2 and 3 are favourable to methods that take dependence in the sample into account, while Main Model 0 is favourable to methods designed for independent data. From each main model, two contaminated models were constructed by randomly adding either three magnitude outliers or three shape outliers.

More specifically, we considered the following main models:

- Main Model 0:

$$\zeta_i(t) = \cos(\pi t)(1-c) + a_i(t) \quad \text{if} \quad -n+1 \leq i \leq n.$$

- Main Model 1:

$$\zeta_i(t) = \begin{cases} \cos(\pi t) & \text{if} \quad i = -n+1 \\ \cos(\pi t)(1-c) + \rho\zeta_{i-1}(t) + a_i(t) + b_i & \text{if} \quad -n+1 < i \leq n. \end{cases}$$

- Main Model 2:

$$\zeta_i(t) = \cos(\pi t)(1-c) + \theta a_{i-1}(t) + a_i(t) + b_i \quad \text{if} \quad -n+1 \leq i \leq n.$$

- Main Model 3:

$$\zeta_i(t) = \begin{cases} \cos(\pi t) & \text{if} \quad i = -n+1 \\ \cos(\pi t)(1-c) + \rho\zeta_{i-1}(t) + \theta a_{i-1}(t) + a_i(t) + b_i & \text{if} \quad -n+1 < i \leq n. \end{cases}$$

In the processes above we have denoted $a_i(t) = X_i \sin(\pi t)$ with X_i being i.i.d. Gaussian variables with mean 0 and standard deviation 0.3, while $\{b_i\}$ is a scalar Gaussian AR(1) process with correlation coefficient $d = 0.8$ and standard deviation $(1-d^2)^{-1/2}$. $c = 0.8$ and $t \in [-0.5, 1.5]$ were considered.

Values ρ and θ manage the dependence strength in the functional time series. We consider two options, one with low dependence (LD, $\rho = 0.5$ and $\theta = -0.5$) and other with high dependence (HD, $\rho = 0.8$ and $\theta = 0.8$).

Then, given each main model, ζ_i, methods were applied on the following contaminated models to detect outliers:

- Contaminated model with magnitude outliers:

$$\chi_i(t) = \zeta_i(t) + k1_{\{i \in \{I_j\}\}}, \ 1 \leq i \leq n.$$

- Contaminated model with shape outliers:

$$\chi_i(t) = \zeta_i(t) + k\cos(3\pi t)1_{\{i \in \{I_j\}\}}, \ 1 \leq i \leq n.$$

Note that k is a contamination size while $1_{\{\cdot\}}$ and I_j denote the indicator function and i.i.d. random variables with discrete uniform distribution on $\{1, \ldots, n\}$, respectively. The curves χ_i were discretized on a grid $\{t_j\}$ of 30 equispaced points in $[-0.5, 1.5]$. Note also that in the simulation process we generate curves corresponding to the double of the sample size n. That is, we simulate the curves $\{\zeta_i(t)\}$, where $-n+1 \leq i \leq n$, but we use only the last half of the curves, $\{\zeta_i(t) : 1 \leq i \leq n\}$, for the contaminated models. The first n realizations are not used in order to avoid the impact of the initial values. The number of outliers introduced in the models was $j = 0.02n$ (that is, 2% of the curves). Value of k was 0.75 for Contaminated Model 0, in which dependence does not affect, and 5 for contaminated Models 1, 2 and 3. It is worth noting here that the contamination size, k, considered in this study is low compared with other existing simulation studies (see, e.g. Sun and Genton, 2011).

Figure 2 shows curves simulated from these four contaminated models. First row corresponds to the Model 0 (no dependence), and the other three rows to the Models 1, 2 and 3 (functional time series), respectively. Last three models are shown for the HD case. We can see in the figure the difference between the data simulated from Model 0 and from Models 1, 2 and 3: in the case of functional time series, outliers are almost always hidden within the rest of the curves.

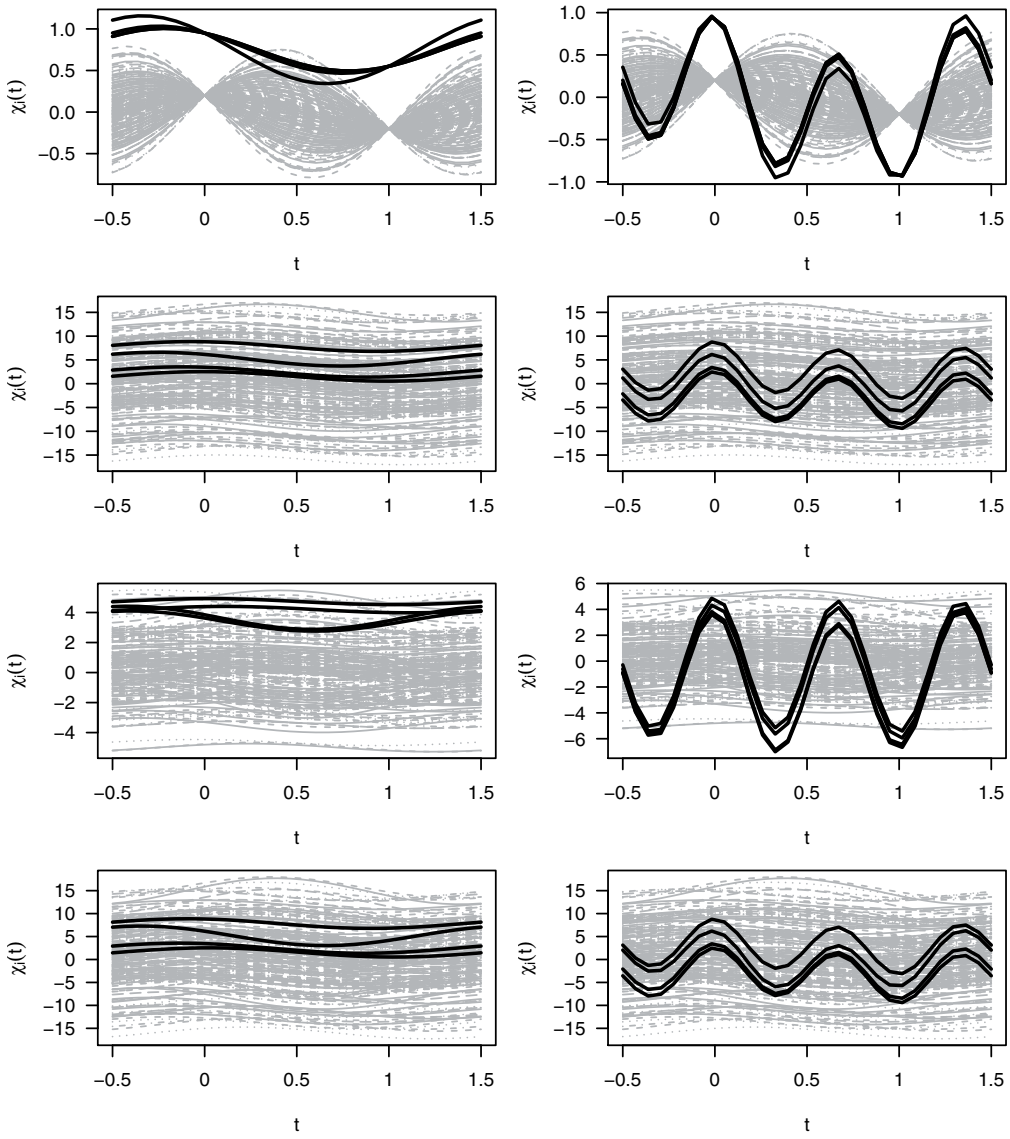


Figure 2: *Left panels: from top to bottom, curves ($\chi_i(t)$) generated from contaminated Models 0, 1, 2 and 3, respectively (the black curves are magnitude outliers). Right panels: from top to bottom, curves ($\chi_i(t)$) generated from contaminated Models 0, 1, 2 and 3, respectively (the black curves are shape outliers).*

We applied the following four procedures on each generated sample in addition to the proposed projections-based (PB) and errors-based (EB) methods.

- Functional highest density region boxplot (HDR). This graphical method is based on the bivariate HDR boxplot (Hyndman, 1996) applied to the first two robust principal component scores (for details, see Hyndman and Shang 2010). The coverage probability $(1 - \alpha_0)$ of the outer region must be prespecified.

- Integrated squared error (ISE). This is the method proposed in Hyndman and Ullah (2007) (for a brief exposition, see last paragraph in Section 2). Both the parameter that controls the degree of robustness (λ) and the quantity of principal components (K) need to be prespecified.
- Depth-based trimming (DBT). Performs a test statistic. A curve is considered an outlier if its depth is lower than a cutoff. The cutoff is determined by a bootstrap method based on trimming the sample (for details, see Febrero, Galeano, and González-Manteiga, 2008). The functional depth, the signification level (α_1), the proportion of potential outliers (α_2), the parameter used to obtain smoothed bootstrap samples (γ) and the number of bootstrap samples (B) must be prespecified.
- DBT for dependent data. Adapts the DBT procedure to the functional time series setting. For that, to determine the cutoff, bootstrap techniques for dependent data are used instead of standard bootstrap. In addition, the functional boxplot of Sun and Genton (2011) is used to trim the sample in the first stage (for details, see Raña, Aneiros, and Vilar, 2015). The functional depth, the signification level (α_1), the bootstrap technique and the number of bootstrap samples (B) need to be prespecified.

Note that the methods HDR, ISE and DBT are designed to detect outliers in samples of independent curves, even if they were also applied to functional time series. Meanwhile, DBT-MBB, PB and EB are specifically address to deal with the problem of outlier detection in the context of functional time series. Along this simulation study, we will compare the performance of the cited methods in situations of both independent and dependent data.

$M = 500$ simulations were run for each model. The percentage of correctly identified outliers p_c (100 times the number of correctly identified outliers over the number of outliers in the sample, or sensitivity) and the percentage of false positives p_f (100 times the number of wrongly identified outliers over the number of non-outlying curves in the sample, or false detection percentage) were computed for each simulation and for each method considered.

Routines fboxplot and foutliers, available in the R package rainbow, were used to detect outliers from the HDR and ISE procedures, respectively. $\alpha_0 = 0.01$ was considered in the HDR method while values for λ and K in the ISE approach were chosen following the suggestion given in Hyndman and Ullah (2007): $\lambda = 3.29$ and K being the value minimizing the ISFE. The method DBT is implemented in the routine Outliers.fdata, available in the R package fda.usc. We considered $\alpha_1 = \alpha_2 = 0.01$, $\gamma = 0.05$ and $B = 200$, while the functional depth was the h-modal depth (Cuevas, Febrero, and Fraiman, 2006), as recommended in Febrero, Galeano, and González-Manteiga (2008). Moving blocks bootstrap (Künsch, 1989) was used in the DBT-for-dependent-data procedure (DBT-MBB) while α_1 and B took the same values as in the DBT method. The functional h-modal depth was also considered in this case.

The choices related to the procedures PB and EB were done following the recommendations given in Section 3.1 and 3.2, respectively. The tuning parameter K was chosen as suggested in Section 3.3. Specifically, we chose $K = 1$ for magnitude outliers (for both methods PB and EB). In the case of shape outliers, we chose $K = 3$ for the PB method and $K = 1$ for the EB. This election explains more than 98% of the variability (in some cases, even with only the first component, it explains around 99.5%), increasing until 99.9% when we use PB method to detect shape outliers. This choice agrees with the guidelines given in Section 3.3 about the requirement of more components when dealing with shape outlier detection and the PB method. In the case of Model 0, due to the simplest performance of the data, it is enough to take $K = 1$ for the EB and $K = 2$ for the PB method and both kind of outliers. The signification level used to detect scalar outliers in the PB method (Step 2) was $\alpha_3 = 0.01$. In the case of the norm to be used in Step 4 of the EB procedure, both the L_1-norm and the L_2-norm were considered. Because similar results were obtained, we only show the corresponding ones to the L_2-norm.

Our first simulation study employs $n = 200$ and the results are reported in tables 1, 2 and 3. These tables show the mean and standard deviation of the values of both p_c and p_f obtained from the two proposed procedures (PB and EB) and the other four considered methods (HDR, ISE, DBT and DBT-MBB) when they are applied to the different contaminated models. In Table 1 the Model 0 is considered (independent data), which is contaminated with magnitude or shape outliers. In Tables 2 and 3, the Models 1, 2 and 3 (dependent data) and the two cases of dependence (low and high dependence) are considered (see Table 2 for contamination with magnitude outliers and Table 3 for the case of shape outliers).

Table 1: *Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Model 0 contaminated with magnitude or shape outliers.*

Model 0				
	Magnitude outliers		Shape outliers	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	40.60 (14.21)	0.19 (0.29)	40.50 (14.24)	0.19 (0.29)
ISE	100.00 (0.00)	0.00 (0.02)	100.00 (0.00)	0.00 (0.00)
DBT	87.00 (26.12)	0.64 (0.48)	84.15 (27.46)	0.61 (0.47)
DBT-MBB	99.80 (4.47)	2.83 (1.44)	99.80 (4.47)	2.83 (1.44)
PB	95.40 (10.44)	0.02 (0.10)	95.15 (10.98)	0.02 (0.10)
EB	100.00 (0.00)	2.14 (1.23)	95.75 (15.61)	2.10 (1.22)

Table 2: *Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Models 1, 2 and 3 (with low or high dependence) contaminated with magnitude outliers.*

Model 1

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	16.50 (15.91)	0.68 (0.32)	9.85 (14.04)	0.82 (0.29)
ISE	22.00 (22.01)	15.33 (2.69)	25.25 (22.74)	21.07 (3.65)
DBT	26.30 (23.28)	1.19 (0.87)	10.55 (15.67)	1.14 (1.17)
DBT-MBB	30.45 (24.04)	2.07 (1.78)	13.10 (16.93)	2.38 (2.24)
PB	70.65 (35.95)	0.31 (0.45)	62.05 (38.65)	0.56 (0.59)
EB	88.55 (17.12)	3.71 (1.24)	84.10 (20.81)	4.07 (1.26)

Model 2

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	28.25 (16.92)	0.44 (0.35)	30.25 (17.08)	0.40 (0.35)
ISE	24.30 (23.73)	14.60 (2.51)	27.00 (23.39)	16.64 (2.79)
DBT	66.75 (25.49)	0.81 (0.69)	67.60 (24.83)	0.76 (0.69)
DBT-MBB	73.15 (24.58)	1.60 (1.17)	73.40 (24.00)	1.55 (1.24)
PB	67.60 (37.26)	0.07 (0.20)	68.40 (37.24)	0.07 (0.18)
EB	91.60 (14.83)	3.17 (1.29)	91.60 (15.08)	3.23 (1.28)

Model 3

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	14.70 (15.31)	0.72 (0.31)	10.75 (14.27)	0.80 (0.29)
ISE	19.60 (21.78)	13.31 (2.72)	30.65 (24.04)	28.96 (3.40)
DBT	26.30 (23.55)	1.20 (0.86)	10.65 (15.44)	1.13 (1.15)
DBT-MBB	30.30 (24.35)	2.08 (1.67)	12.85 (16.78)	2.35 (2.03)
PB	69.30 (36.30)	0.33 (0.47)	60.95 (38.36)	0.61 (0.64)
EB	88.45 (16.76)	3.63 (1.24)	84.20 (20.28)	3.81 (1.17)

Table 3: *Mean and standard deviation (in parentheses) of the percentage of correctly and falsely identified outliers in Models 1, 2 and 3 (with low or high dependence) contaminated with shape outliers.*

Model 1

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	16.15 (19.20)	0.69 (0.39)	13.90 (18.77)	0.74 (0.38)
ISE	100.00 (0.00)	14.58 (2.74)	100.00 (0.00)	20.15 (3.59)
DBT	95.75 (16.47)	0.21 (0.41)	64.25 (38.55)	0.61 (0.91)
DBT-MBB	99.40 (6.69)	0.70 (1.16)	56.05 (37.45)	0.58 (1.06)
PB	95.20 (10.71)	0.04 (0.17)	95.00 (11.19)	0.04 (0.24)
EB	100.00 (0.00)	2.58 (1.26)	100.00 (0.00)	2.59 (1.29)

Model 2

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	10.75 (15.20)	0.80 (0.31)	9.10 (14.23)	0.83 (0.23)
ISE	100.00 (0.00)	13.85 (2.51)	100.00 (0.00)	15.95 (2.67)
DBT	96.40 (15.57)	0.38 (0.45)	98.30 (11.06)	0.33 (0.42)
DBT-MBB	100.00 (0.00)	1.81 (1.36)	100.00 (0.00)	1.67 (1.35)
PB	95.20 (10.71)	0.03 (0.16)	95.15 (10.63)	0.04 (0.16)
EB	100.00 (0.00)	2.61 (1.38)	100.00 (0.00)	2.64 (1.34)

Model 3

	Low dependence		High dependence	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	17.35 (19.26)	0.67 (0.39)	4.65 (11.40)	0.93 (0.23)
ISE	100.00 (0.00)	12.52 (2.62)	100.00 (0.00)	27.86 (3.23)
DBT	94.40 (19.15)	0.22 (0.42)	49.60 (38.54)	0.79 (1.04)
DBT-MBB	99.40 (6.69)	0.47 (1.23)	41.80 (34.50)	0.94 (1.32)
PB	95.05 (10.82)	0.05 (0.20)	94.95 (11.45)	0.14 (0.31)
EB	100.00 (0.00)	2.51 (1.26)	100.00 (0.00)	2.06 (1.19)

Several conclusions can be drawn from these results. First of all, we look at Contaminated Model 0 in Table 1, which considers independent data. Under that situation, ISE method gets the best result for both kind of outliers. On the contrary, HDR presents poor results with the lowest sensitivity, but also its false detection rate is low. Looking at the pair of DBT and DBT-MBB method, we can see an improvement with the second option, even if dependence is not affecting this data. p_c is much better for the DBT-MBB method, compared to the DBT, but also the p_f is higher. Note that DBT-MBB not only adapts DBT to work with functional time series (by taking dependence into account), but also improves the method itself by changing some other aspects. This is why we can see different results even when they are applied to independent data. Our both proposals, PB and EB, are very competitive in this situation, even compared to methods designed to work with independent data. They maintain high and low values for p_c and p_f, respectively. Their sensitivity is greater than 95% and there is no big difference between magnitude and shape outliers. We can see that PB detects less outliers than EB but also its false detection rate is lower.

Now, we focus on the simulated models that include dependence structure; that is, contaminated Models 1, 2 and 3. The role of this analysis is two-fold: to illustrate the performance of the two proposed procedures and to show the need to take into account the dependence in the functional time series. We restrict first to the magnitude outliers under both situations of low dependence (LD) and high dependence (HD), which results are given in Table 2. In general we do not observe major differences in the behaviour of the proposed methods (PB and EB) when the dependence scenario changes (LD or HD), and we can see that the best results are achieved by the methods that take into account dependence (DBT-MBB, PB and EB). Results are analysed below in a deeper way. HDR and ISE methods lose their effectiveness in detecting outliers when dealing with dependent data. We may highlight the large p_f (around 20%) of the ISE method, indicating a high volatility in its behaviour. We look now at the pair DBT and DBT-MBB methods (remember that DBT-MBB adapts DBT to work with functional time series). It is true that DBT-MBB gets always higher p_c, which clearly indicates that taking dependence in the data into account is outstanding. Both methods are also better than HDR and ISE in most of the cases. Despite of getting worse p_c than ISE when dealing with Models 1 and 3 under high dependence, they get significantly lower p_f. Both methods (DBT and DBT-MBB) also show a sharp difference between dependence scenarios for the Models 1 and 3, in which the outlier detection rate decreases as the dependence structure becomes more relevant.

All the methods analysed above are overcome by our two proposals PB and EB. Both options achieve high sensitivity, greater than DBT-MBB (excepting Model 2 in which DBT-MBB overcomes PB) and far away from the other considered methods that not take into account dependence. PB method holds lower sensitivity than EB, but also lower false detection rate. To obtain a trade-off between high sensitivity and low false detection rate, in general, the proposed EB seems to be a good choice for magnitude outlier detection under the considered dependence scenarios.

Table 3 shows the results when the models are contaminated with shape outliers. HDR still performance very similar to the magnitude outliers case, however ISE methods shows an improvement by detecting all the shape outliers (at the expense of a large false detection rate). DBT and DBT-MBB behaves also similarly to the magnitude outliers case, with a remarkable difference in the levels of p_c. They achieve now very high sensitivity with low dependence (around $95-100\%$) but under high dependence they provide low values, around $40-60\%$ for Models 1 and 3.

Proposed methods PB and EB show high sensitivity (95% and 100%, respectively) and low false detection rate (0.05% and 2.5%), being very stable for the three simulated models. As in Table 2 for magnitude outliers, also with shape outliers there is no major differences between both dependence scenarios (LD and HD). In summary, even if both methods obtain very good results for shape outlier detection under dependence, EB seems to be a better choice due its great success detecting all the outliers.

A second simulation study is developed in order to study the influence of the sample size (n) over the analysed methods for outlier detection. In this case we restrict to Models 1, 2 and 3 (simulated functional time series) contaminated with magnitude outliers. Table 4 shows the mean of the percentage of correctly and falsely identified outliers (p_c and p_f, respectively) when the sample size varies within the values $n = 100, 200, 300$ and 400. These results are obtained under the scenario of high dependence (HD) and the number of outliers introduced in each sample follows the same rule as the previous results (including $j = 0.02n$ outliers; that is, 2% of the curves).

Results given by the two proposed methods (PB and EB) in Table 4 clearly overcome the rest of the methods included in the comparison (HDR, ISE, DBT and DBT-MBB) in almost all the situations (except when one considers $n = 100$ in Model 2). That is, for the three contaminated models and the different values of the sample size n (except the combination Model 2, $n = 100$), PB and EB get the best performing. On the one hand, HDR, ISE, DBT and DBT-MBB show poor results with very low p_c and also, in the case of ISE method, very high false detection rate. DBT-MBB gets always better results than DBT, showing again the importance of taking dependence in the data into account. Actually, both DBT and DBT-MBB are very competitive for the Model 2, specially with the lowest sample size $n = 100$, but they are overcome by PB and EB as n increases.

On the other hand, also HDR, ISE, DBT and DBT-MBB remains stable when the sample size varies. Indeed, we can see a slight decrease in the p_c and increase in p_f as long as the sample size n increases. On the opposite, for the three contaminated models, proposed methods PB and EB clearly improve the sensitivity (p_c) meanwhile the false detection rate (p_f) decreases slightly as n increases. The reason for this is that PB and EB methods are based on fitting univariate time series (of the coefficients given by FPCA) as a previous step to the outlier detection. Therefore, by increasing the sample size n the fit of the univariate time series is improved and, accordingly, also the outlier detection with PB and EB methods improves.

Table 4: *Mean of the percentage of correctly and falsely identified outliers in Models 1, 2 and 3, with high dependence, contaminated with magnitude outliers and for* $n = 100, 200, 300$ *and* 400.

Model 1								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	12.30	0.77	9.85	0.82	7.43	0.87	7.28	0.87
ISE	26.10	22.44	25.25	21.07	22.57	21.10	23.13	20.94
DBT	12.10	0.34	10.55	1.14	9.23	1.60	9.63	1.86
DBT-MBB	12.80	1.13	13.10	2.38	11.80	2.99	12.03	2.98
PB	29.20	0.56	62.05	0.56	80.23	0.42	87.90	0.28
EB	58.80	4.52	84.10	4.07	90.83	3.99	93.60	4.00
Model 2								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	34.20	0.33	30.25	0.40	28.57	0.44	26.50	0.48
ISE	32.60	17.10	27.00	16.64	22.50	16.51	21.50	16.45
DBT	63.70	0.47	67.60	0.76	68.63	0.90	68.87	1.01
DBT-MBB	73.60	1.66	73.40	1.55	72.30	1.52	72.08	1.54
PB	34.50	0.10	68.40	0.07	84.93	0.04	89.43	0.04
EB	73.70	3.60	91.60	3.23	95.70	3.35	96.25	3.27
Model 3								
	n=100		n=200		n=300		n=400	
Method	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$	$\widehat{p}_c$	$\widehat{p}_f$
HDR	15.80	0.70	10.75	0.80	8.70	0.84	8.40	0.85
ISE	31.90	29.08	30.65	28.96	30.10	28.74	29.80	28.84
DBT	12.00	0.32	10.65	1.13	9.37	1.61	9.58	1.85
DBT-MBB	12.70	1.18	12.85	2.35	11.17	2.78	11.40	2.77
PB	29.80	0.60	60.95	0.61	81.10	0.40	87.00	0.29
EB	54.90	3.68	84.20	3.81	90.67	3.78	93.18	3.81

5. Applications in the electricity market

Nowadays, in many countries all over the world, the production and sale of electricity is traded under competitive rules in free markets. The agents involved in this market (namely, system operators, regulatory agencies, producers and consumers) are greatly interested in the study of electricity load and price. Since electricity cannot be stored, the demand must be satisfied instantaneously and producers need to anticipate future demands to avoid overproduction. So good forecasting of electricity demand is very important for the agents in the market. On the other hand, if reliable predictions of electricity price are available to producers and consumers, they can develop their bidding strategies and establish a pool-bidding technique to achieve a maximum benefit. Consequently, the prediction of electricity demand and price pose significant concerns to this sector. In recent years, these concerns have been addressed from a functional perspective. Regression models with functional covariates (and even functional response) have been used to forecast electricity demand and price. Some related papers are Antoniadis, Paparoditis, and Sapatinas (2006), Antoch et al. (2010), Vilar, Cao, and Aneiros (2012), Cho et al. (2013), Lielb (2013) and Aneiros et al. (2016). It is well known that the presence of outliers affects the accuracy of forecasts obtained from regression models. Thus, outlier detection represents a first step in any descriptive analysis of a dataset, prior to any type of modelling or prediction method. In that sense, depending on the objective of the study, one of the following strategies can be used once the outliers are identified: if outliers come from gross errors, they are subsequently removed from the sample. In another case, robust prediction methods or complex models that take into account the existence of outliers (for instance, introducing dummy variables) can be used.

In the next two sections, the proposed methods for detecting outliers in functional time series, projections-based (PB) and errors-based (EB) methods, are applied on daily curves of electricity demand and price. The corresponding tuning parameters were selected in a similar way as in the simulation study.

5.1. Case study: electricity demand

We are interested in outlier detection in time series of electricity demand curves. Data collect hourly electricity demand in the Spanish mainland electricity market on Mondays, . . . , Fridays in the year 2012. They are available at http://www.omie.es, the official website of Operador del Mercado Ibérico de Energía. These hourly data present a trend. Thus, by subtracting the trend (estimated by means of a kernel regression) we obtained the corresponding detrended hourly series. The functional dataset under analysis is composed of the $n = 261$ daily demand curves obtained from this detrended hourly series, measured in Megawatt-hour (MWh). The quantity of functional principal components considered was $K = 9$. These K principal components explained, at least, 98% of the variance.

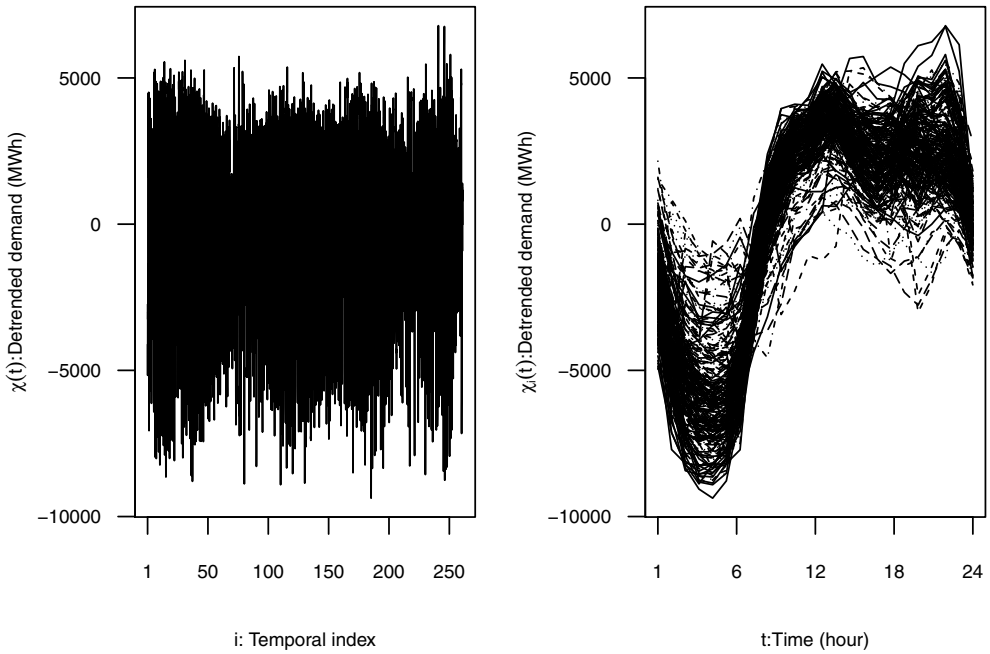


Figure 3: *Left panel: time series of electricity demand. Right panel: daily electricity demand curves.*

The functional time series and the corresponding daily curves are shown in Figure 3. Higher demands are observed in the interval 10:00h–22:00h while lower ones correspond to the interval 3:00h–5:00h.

The outliers identified from the proposed procedures PB and EB are listed in Table 5. Methods PB and EB detect 11 and 15 outliers, respectively, resulting in 20 different curves.

We can find possible causes for most of these abnormal curves. For example, April 16 and 19, November 1 and 2 and also December 24 correspond to days with zero price hours. During some hours in these days, the overproduction of wind power decreases the electricity price fixed by daily market. This have to do with the different taxations of this "green energies" because, as the wind power production increases, the electricity price decreases. As a result, if the wind power production covers and abnormally high percentage of the electricity demand, the price can drop even until zero during a period of time (this being the case of the cited days). We find also as outliers some previous or posterior days to these "zero price days", such as April 24 and 26, which are also affected by the disturbance in the price. Some of the outliers correspond to nonworking days in which the people usually behaves in a different way than the rest of the regular days (simply because most of the economical and industrial activities stop during these days), affecting the electrical consumption and, as a consequence, also the demand. This is the case of May 1 (Labour Day), August 15 (Assumption Day), October 12

Table 5: *Outliers detected in the demand data from proposed procedures PB and EB.*

	Method			Method	
Day	PB	EB	Day	PB	EB
February 14	X		November 2		X
April 16	X	X	November 7		X
April 19	X	X	November 14	X	X
April 24		X	December 3	X	
April 26	X		December 6	X	X
May 1	X	X	December 10		X
May 4	X		December 21		X
August 15	X	X	December 24	X	
October 12		X	December 25		X
November 1		X	December 28		X

(National holiday in Spain), November 1 (All Saints Day), December 6 (Constitution Day in Spain) and 25 (Christmas). December 24 (Christmas Eve) is also a special day, even if it is not officially a holiday. Friday, November 2, besides being a zero-price day, is situated also in the middle of a long weekend caused by All Saints Day, in which a lot of people take some holidays. Finally, November 14 was a strike day in Spain, which clearly affects electrical consumption as it can be considered in some sense as a holiday.

Finally, it is worth pointing out that electricity demand curves observed at days April 16, November 14 and December 6 are detected as outliers simultaneously with the two proposed methods, but no one of these curves is identified as an outlier from either the HDR or DBT procedures (remember that neither HDR nor DBT take dependence in the data into account). Actually, as can be seen in Figure 4, these three curves have features that can, to say the least, be considered suspicious: demand curve observed at April 16 takes high values throughout the first hours (possibly because the electricity price at 3:00h–6:00h was zero); demand curve corresponding to November 14 (strike day) maintains low values from 7:00h, this being the typical behaviour of demand curves corresponding to nonworking days; December 6 is a holiday.

5.2. Case study: electricity price

A similar study is conducted in this section for electricity price. Prices were available for the same period as demands, and they were obtained from the same source. The units were cents (euro) per kilowatt-hour (cents/kWh). Unlike the previous case, there was no

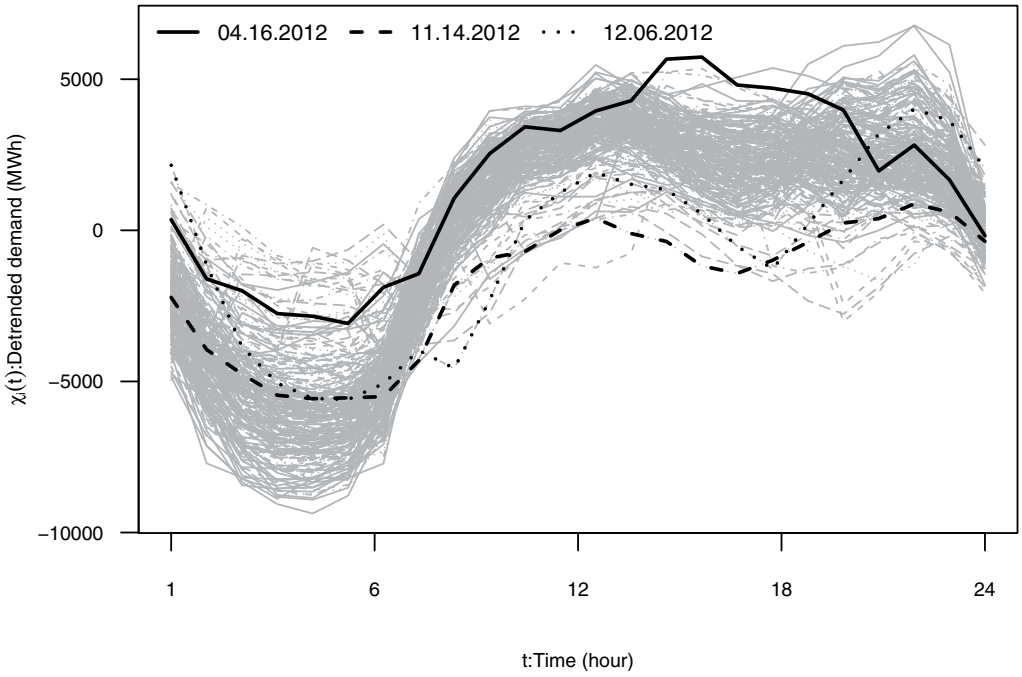


Figure 4: *Outliers simultaneously detected in the demand curves with both of the proposed procedures, but not detected when a method designed for independent data is applied.*

trend. The number of functional principal components considered was $K = 8$. These K principal components explained, at least, 98% of the variance.

Figure 5 displays the functional time series of electricity prices and the associated daily curves. Note that periods of low and high prices roughly correspond with periods of low and high demand, respectively. Greater variability is observed in the time series of prices, taking into account the different scale with respect to the electricity demand. It is easy to distinguish some of the zero-price days present in some points along the year, caused by the overproduction of wind power, and also in the daily curves (specially between 3:00h–6:00h).

The outliers identified by the proposed procedures PB and EB are listed in Table 5. Note that a total of 20 observations are detected as abnormal curves (13 from the PB method and 15 from the EB method). In addition, 7 of the 20 days corresponding to such outliers were days when demand curves were also identified as outliers (compare Tables 4 and 5). Following the classical rules of any kind of market, it is usual that demand and price are very interconnected, this being also the case of electricity markets and the reason why some of the outlying curves in demand are repeated as outliers in the electricity price. As in the previous application, one can argue causes for most of the abnormal curves of electricity price, being most of them already cited in the study of outliers in electricity demand. Some of the outliers correspond to zero-price days, as

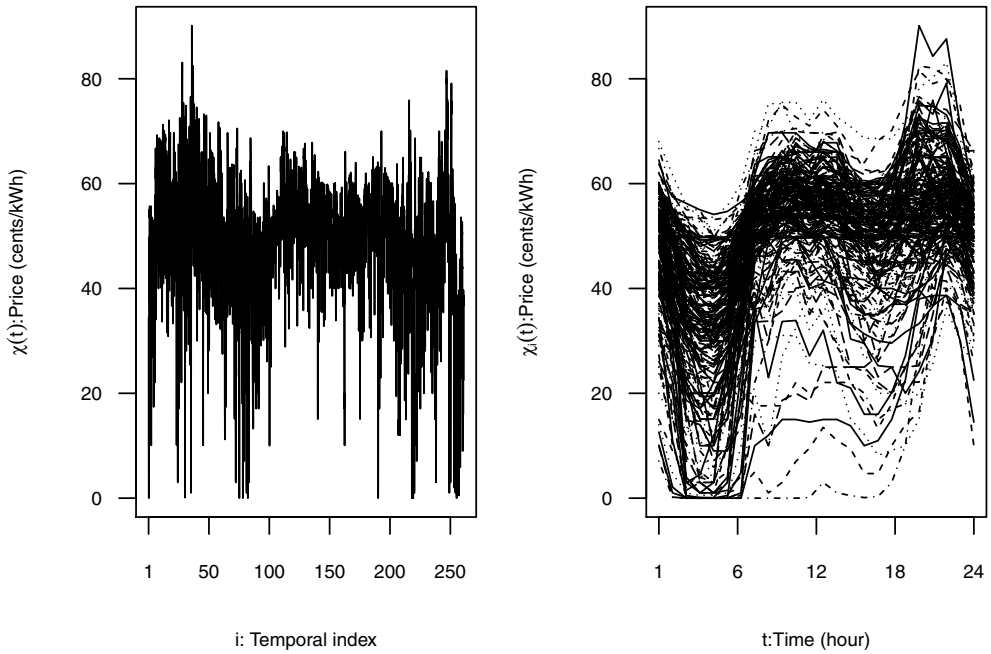


Figure 5: *Left panel: time series of electricity price. Right panel: daily electricity demand curves.*

Table 6: *Outliers detected in the price data from proposed procedures PB and EB.*

	Method			Method	
Day	PB	EB	Day	PB	EB
February 13	X		August 15	X	X
February 21		X	August 16		X
April 6	X		September 24		X
April 10	X	X	October 1		X
April 11	X	X	October 24		X
April 19	X		November 1	X	X
April 25	X	X	November 2	X	
May 1	X	X	December 14		X
May 8		X	December 24	X	X
June 11	X		December 25	X	X

April 19 and 25, September 24, November 1 and 2 or December 24 or days with a period close to zero price (February 13). Holidays have also some kind of influence over electricity prices, as April 6 (Good Friday) or May 1, August 15, November 1 and

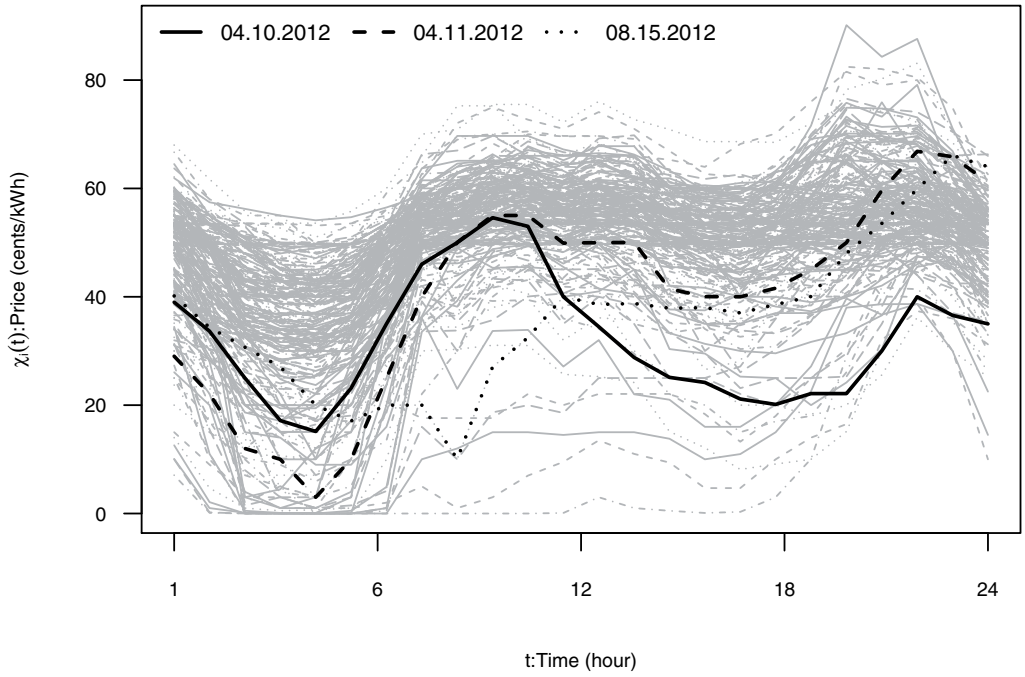


Figure 6: *Outliers in the price curves simultaneously detected using both proposed procedures, but not detected when applying a method designed for independent data.*

December 25. Finally, we find also some special days related to other holidays or linked to holidays, this being the case of February 21 (Carnival, holiday in part of Spain), August 16 (posterior to a nonworking day), November 2 (in the middle of a long weekend) or December 24 (Christmas Eve).

Finally, again as in the case of the demand, three price curves are detected as outliers simultaneously by the two proposed methods, but none of these curves is identified as an outlier with either the HDR or DBT procedure; we refer to the curves corresponding to April 10 and 11, and August 15, see Figure 6. It seems to make sense to consider them as outliers: the price was very low in the second half of the day on April 10 and the first half of the following day, April 11; August 15 is a holiday, and the pattern of the corresponding price curve of this day is different from the working days pattern.

6. Conclusions

This article proposes two methods to detect outliers in functional time series, the projections-based (PB) and the errors-based (EB) methods. These methods take dependence in the data into account and use robust functional principal component analysis (FPCA).

Our simulation studies have shown that the proposed methods present good performance when they are applied either on independent curves or dependent curves.

However, procedures designed for independent data, such as the functional HDR boxplot (Hyndman and Shang, 2010), the depth-based trimming (Febrero, Galeano, and González-Manteiga, 2008) or the integrated squared error (Hyndman and Ullah, 2007) methods, fail to detect outliers in functional time series. Thus, it has also been shown the need to take dependence in the time series into account. The PB method has very low false detection rate (p_f) while the sensitivity (p_c) of the EB approach is very high. Although in our simulation study small contamination sizes have been considered, both methods show acceptable trade-off between p_c and p_f. In fact, they improve the trade-off corresponding to the DBT for dependent data (Raña, Aneiros, and Vilar, 2015), this being (to the best of our knowledge) the only method in the statistical literature that includes the effect of dependence in the detection of outliers. Both PB and EB have also shown good performance in different situations, considering the kind of outlier (magnitude or shape outliers) and also de dependence scenario (low or high dependence). Their output is generally better than the other methods included in the comparison. Regarding sensitivity (p_c), EB seems to be the best option for both magnitude or shape outliers. Furthermore, PB is also very accurate and, although its p_c is lower than the EB, its false detection rate is the lowest of all the methods. We have also shown that both proposals improve their results as long as the sample size increases. The practical usefulness of our methodology has been illustrated on the daily curves of electricity demand and price.

Finally, it is worth pointing out that, as in all procedures based on FPCA, the proposed methods depend on the quantity of principal components considered, K. In this article, K was selected by imposing a lower bound to the cumulative percentage of variance explained from the first K principal components (cumulative percentage variance (CVA) approach). As in Hyndman and Booth (2008), we find that a general recommendation is to use a larger than necessary K (for instance, a K explaining at least 98% or, even, 99.9% of the variability). Alternatives to the CVA approach are, for instance, methods based on the cross-validation score (Yao, Müller, and Wang, 2005a) or the Akaike information criterion (Yao, Müller, and Wang, 2005b).

Acknowledgments

The authors wish to thank two anonymous referees for their helpful comments and suggestions, which greatly improved the quality of this paper. This research was partially supported by Grants MTM2014–52876–R from Spanish Ministerio de Economía y Competitividad, and CN2012/130 from Xunta de Galicia.

References

Aneiros-Pérez, G., Cardot, H, Estévez-Perez, G., and Vieu, P. (2004). Maximum ozone concentration forecasting by functional non-parametric approaches. *Environmetrics*, 15, 675–685.

Aneiros, G., Vilar, J.M., Cao, R., and Muñoz-San-Roque, A. (2013). Functional prediction for the residual demand in electricity spot markets. *IEEE Transactions on Power Systems*, 28, 4201–4208.

Aneiros, G., Vilar, J., and Raña, P. (2016). Short-term forecast of daily curves of electricity demand and price. *Electrical Power and Energy Systems*, 80, 96–108.

Antoch, J., Prchal, L., De Rosa, M.R., and Sarda, P. (2010). Electricity consumption prediction with functional linear regression using spline estimators. *Journal of Applied Statistics*, 37, 2027–2041.

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2006). A functional wavelet kernel approach for time series prediction. *Journal of the Royal Statistics Society B*, 68, 837–857.

Arribas-Gil, A., and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15, 603–619.

Baíllo, A., Cuesta-Albertos, J.A., and Cuevas, A. (2011). Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics*, 38, 480–498.

Besse, P.C., Cardot, H., and Stephenson, D. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27, 673–688.

Boente, G., and Fraiman, R. (2000). Kernel-based functional principal components. *Statistics and Probability Letters*, 48, 335–345.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, 45, 11–22.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108, 7–21.

Cryer, J.D., and Chan, K.S. (2008). *Time Series Analysis*, New York: Springer.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1–23.

Cuevas, A., Febrero, M., and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, 51, 1063–1074.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22, 481–496.

Delsol, L., Ferraty, F., and Vieu, P. (2011). Structural test in regression on functional variables. *Journal of Multivariate Analysis*, 102, 422–447.

Febrero, M., Galeano, P., and González-Manteiga, W. (2007). Functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics*, 22, 411–427.

Febrero, M., Galeano, P., and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, 19, 331–345.

Ferraty, F., and Romain,Y. (Eds.) (2011). *The Oxford Handbook of Functional Data Analysis*, Oxford: Oxford University Press.

Ferraty, F., van Keilegom, I., and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109, 10–28.

Ferraty, F., and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17, 545–564.

Ferraty, F., and Vieu, P. (2006). *Nonparametric Functional Data Analysis*, New York: Springer-Verlag.

Fraiman, R., and Svarc, M. (2013). Resistant estimates for high dimensional and functional data based on random projections. *Computational Statistics and Data Analysis*, 58, 326–338.

García-Portugués, E., González-Manteiga, W., and Febrero-Bande, M. (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23, 761–778.

Gervini, D. (2012). Outlier detection and trimmed estimation for general functional data. *Statistica Sinica*, 22, 1639–1660.

González-Manteiga, W., and Martínez-Calvo, A. (2011). Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference*, 141, 453–461.

Hall, P. (2011). Principal component analysis for functional data: methodology, theory, and discussion. *The Oxford Handbook of Functional Data Analysis*, F. Ferraty and Y. Romain, Eds., Oxford: Oxford University Press, 210–234.

Hall, P., Müller, H.G., and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34, 1493–1517.

Horváth, L., and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, New York: Springer.

Hubert, M., Rousseeuw, P.J., and Verboven, S. (2002). A fast method of robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50, 120–126.

Hyndman, R.J., and Ullah, M.S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51, 4942–4956.

Hyndman, R.J., and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24, 323–342.

Hyndman, R.J., and Shang, H.L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19, 29–45.

Künsch, H.R. (1989). The jacknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217–1241.

Li, Y., and Hsing, T. (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98, 1782–1804.

Liebl, D. (2013). Modeling and forecasting electricity spot prices: a functional data perspective. *Annals of Applied Statistics*, 7, 1562–1592.

Ocaña, F.A., Aguilera, A.M., and Escabias, M. (2007). Computational considerations in functional principal component analysis. *Computational Statistics*, 22, 449–465.

Ramsay, J.O., and Silverman, B.W. (2005). *Functional Data Analysis*. New York: Springer-Verlag.

Raña, P., Aneiros, G., and Vilar, J.M. (2015). Detection of outliers in functional time series. *Environmetrics*, 26, 178–191.

Sawant, P., Billor, N., and Shin, H. (2012). Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27, 83–102.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *Test*, 23, 725–750.

Shang, H.L. (2014). Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density. *Journal of Nonparametric Statistics*, 26, 599–615.

Sun, Y., and Genton, M.G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20, 316–334.

Tsay, R.S., Peña, D., and Pankratz, A.E. (2000). Outliers in multivariate time series. *Biometrika*, 87, 789–804.

Vilar, J.M., Cao, R., and Aneiros, G. (2012). Forecasting next-day electricity demand and price using non-parametric functional methods. *International Journal of Electrical Power and Energy Systems*, 39, 48–55.

Yao, F., Müller, H.G., and Wang, J.L. (2005a). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33, 2873–2903.

Yao, F., Müller, H.G., and Wang, J.L. (2005b). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577–590.

Yu, G., Zou, C., and Wang, Z. (2012). Outlier detection in functional observations with applications to profile monitoring. *Technometrics*, 54, 308–318.

SORT 40 (2) July-December 2016, 349-374

Log-ratio methods in mixture models for compositional data sets

M. Comas-Cufí, J.A. Martín-Fernández and G. Mateu-Figueras

Abstract

When traditional methods are applied to compositional data misleading and incoherent results could be obtained. Finite mixtures of multivariate distributions are becoming increasingly important nowadays. In this paper, traditional strategies to fit a mixture model into compositional data sets are revisited and the major difficulties are detailed. A new proposal using a mixture of distributions defined on orthonormal log-ratio coordinates is introduced. A real data set analysis is presented to illustrate and compare the different methodologies.

MSC: 62E99, 62G07, 62H30, 62H99.

Keywords: Compositional data, Finite Mixture, Log ratio, Model-based clustering, Normal distribution, Orthonormal coordinates, Simplex.

1. Introduction

A *finite mixture distribution* is a probability distribution with probability density function (pdf) given by the expression

$$\pi_1 f_1(\,\cdot\,;\boldsymbol{\theta}_1) + \cdots + \pi_k f_k(\,\cdot\,;\boldsymbol{\theta}_k), \tag{1}$$

where $f_1,\ldots,f_k$ are pdf's of distributions with parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ respectively, and $\pi_1,\ldots,\pi_k$ are positive numbers with $\sum_{i=1}^{k}\pi_i = 1$ (McLachlan and Peel, 2000). The pdfs $f_1,\ldots,f_k$ are typically called *mixtures components*. In this paper we assume the most common case where all the mixture components, f_i, in a mixture belong to a unique family (Gaussian, skew-normal, etc) with pdf, f, and parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ belonging to a unique set $\boldsymbol{\Theta}$.

According to Scott and Symons (1971) and McLachlan and Peel (2000), finite mixture models provide reasonable results in several multivariate techniques, for instance,

Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona. gloria.mateu@udg.edu

Received: February 2016
Accepted: October 2016

discriminant analysis, density estimation and model-based clustering (Banfield and Raftery, 1993), even for high-dimensional data (Bouveyron and Brunet-Saumard, 2014). The Gaussian mixture is the most common model thanks to its theoretical and computational simplicity (McLachlan and Peel, 2000). However, because of its simplicity, Gaussian mixtures have some significant limitations which triggered the proposal of alternative models. For example, Student t mixtures were introduced to fit distributions with heavier tails (Andrews and McNicholas, 2012, Lee and McLachlan, 2014, Lin, 2010); and skew-normal and skew-t (Azzalini and Capitanio, 1999, 2003) mixtures were proposed to fit asymmetrical distributions (Lee and McLachlan, 2011). Moreover, Browne and McNicholas (2013) introduced the Generalized Hyperbolic mixture, a more general mixture model which includes, either asymptotically or explicitly, different types of well-known families of mixture models. A crucial point to note is that all these mixture models were designed for data in real space. For data in a different sample space, there is a general agreement that other distributions should be used. For example, Bickel and Scheffer (2004) used multinomial mixture distributions for discrete data in text classification, and Bouguila (2011) proposed other extensions of multinomial mixture distributions for count data. Another example is circular data, whose sample space is the sphere. Banerjee et al. (2005) and Mardia et al. (2007) proposed mixtures of Von Mises probability distributions, defined for random vectors in the sphere.

Finite mixture modelling for compositional data (CoDa) also needs its own probability distributions because the CoDa sample space, the simplex $\mathscr{S}^D$, has a particular algebraic-geometric structure, different from the one in real space (Pawlowsky-Glahn and Egozcue, 2001). CoDa, also called D-part compositions, are vectors $\mathbf{x} = (x_1, ..., x_D)$ with all its parts strictly positive and carrying only relative information. A D-part composition is usually restricted to sum to a fixed constant κ, i.e.

$$\sum_{i=1}^{D} x_i = \kappa. \tag{2}$$

As a convention, it is usual to assume $\kappa = 1$ for proportions and $\kappa = 100$ for percentages. Because the value of κ is irrelevant, in this paper we will assume that $\kappa = 100$ for simplicity. Typical examples of CoDa are frequent in economics (income and expenditure distributions), medicine (body composition: fat, bone, muscle), the food industry (food composition: fat, sugar, etc), geochemistry and chemometrics (chemical composition), ecology (abundance of different species), sociology (time-use surveys), and genetics (genotype frequency). When a problem is compositional, one assumes that the absolute value of each part is irrelevant and the interest is focused on the ratios of the parts. Following this idea, Aitchison (1986) introduced the log-ratio methodology to deal with compositional data. According to this methodology, the compositions are expressed in terms of log-ratio coordinates and traditional techniques are applied to them. This log-ratio methodology is coherent with the algebraic-geometric structure of the simplex

introduced later by Pawlowsky-Glahn and Egozcue (2001). In the literature we find a large number of papers where a specific methodology for CoDa is developed following the log-ratio approach (e.g., Martín-Fernández et al., 2015, Vives-Mestres et al., 2014, Palarea-Albaladejo et al., 2012).

As in many other statistical methods, log-ratio methodology requires complete data sets. When measuring concentrations, some elements are often not present in sufficient concentrations and measuring instruments report them as values below detection limits. In the literature this issue is also known as the rounded zero problem. The data matrix is completed by using imputation strategies, replacing non-detected values with reasonable estimates, and by allowing the computation of log-ratios for applying to any multivariate data analysis. The interested reader can refer to Palarea-Albaladejo et al. (2014), whose work encompasses the recent advances in this area.

Another approach to the zero problem consists in transforming the data from the simplex into the real space using a transformation defined on the zero, for example the hyperspherical transformation (Neocleous et al., 2011, Wang et al., 2007). Scealy et al. (2015) recommend the square root transformation because it handles zero components. While these possibilities can exhibit good results, in practice they lack of geometric structure (see discussion in Aitchison, 1982). In this work we consider the log-ratio methodology, which can be seen as a transformation but it also provides a geometry to the simplex with its own operations.

It is difficult to find in the literature finite mixture models for CoDa that consider distributions restricted to the simplex. The exception are a few studies (e.g., Albert and Gupta, 1982, Bouguila et al., 2004, Calif et al., 2011) where finite mixture models using Dirichlet distributions, a traditional probability distribution in the simplex, are used. Nevertheless, it is more frequent to ignore the compositional nature of the CoDa data and to use mixtures models of distributions on real space (e.g., Papageorgiou et al., 2001). Recently, in practical works, the log-ratio methodology had been considered to fit a mixture model (e.g., Ferrer-Rosell et al., in press) without theoretical and methodological considerations. As a consequence, there is a methodological gap in the analysis of CoDa where the latest advances in log-ratio methods can contribute to mixture modelling. In the present work, we introduce a new technique to model CoDa using mixtures of distributions well-defined on the simplex using orthonormal log-ratio coordinates and consequently coherent with its algebraic-geometric structure. In particular we use the normal and the skew-normal distributions on the simplex (Mateu-Figueras and Pawlowsky-Glahn, 2007, Mateu-Figueras et al., 2013).

This paper is organized as follows: in Section 2 a brief introduction of CoDa analysis is provided. Section 3 describes the pros and cons of each of the traditional mixture models when applied to CoDa. Section 4 is devoted to introducing log-ratio mixture models and two real data sets are analysed in Sections 5 and 6 to compare the traditional and log-ratio approaches. Finally, Section 7 contains conclusions and final remarks. The programming of the data analyses discussed in this work has been conducted using the open-source R statistical environment (R Core Team, 2014). Computer rou-

tines implementing the methods can be obtained from the R packages `Mclust`, `Rmixmod`, `EMMIXuskew` and also from the website www.compositionaldata.com. As an accompaniment to this article, the data and the programs used to fit the mixtures in Sections 5 and 6 are provided as supplementary material.

2. Compositional data analysis

Aitchison (1986) stated that there are two basic operations in the simplex $\mathscr{S}^D$: *perturbation* ($\oplus$) and *powering* ($\odot$). *Perturbation* is defined between two compositions $\mathbf{x}$ and $\mathbf{y}$, and *powering* is defined between a composition $\mathbf{x}$ and a scalar value α as:

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, \ldots, x_D y_D), \qquad \alpha \odot \mathbf{x} = C({x_1}^{\alpha}, \ldots, {x_D}^{\alpha}), \tag{3}$$

where $C(\mathbf{x}) = \frac{\kappa}{\sum x_k}(x_1, \ldots, x_D)$ is the closure operation for rescaling a vector.

These operations respectively play analogous roles to translation and scalar multiplication in $\mathbb{R}^D$, and provide a vector space structure of dimension $D-1$ to the simplex. Pawlowsky-Glahn and Egozcue (2001) stated that the inner product

$$< \mathbf{x}, \mathbf{y} >_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \tag{4}$$

provides $\mathscr{S}^D$ with the structure of an Euclidean space of dimension $D-1$. Note that a norm and a distance can be derived from the inner product given by Equation 4. This Euclidean space structure allows us to establish the principle of working on coordinates (Mateu-Figueras et al., 2011). The idea is to express compositions in terms of their coordinates with respect to an orthonormal basis on $\mathscr{S}^D$ and apply traditional statistical methods to these coordinates. These coordinates are formed by log-ratios, therefore we use the log-ratio methodology mentioned above. Once an orthonormal basis $\mathscr{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ is fixed, any D-part composition $\mathbf{x}$ can be expressed as the linear combination

$$\mathbf{x} = (h_1 \odot \mathbf{v}_1) \oplus \cdots \oplus (h_{D-1} \odot \mathbf{v}_{D-1}).$$

The elements of vector $\mathbf{h}_{\mathscr{B}}(\mathbf{x}) = (h_1, \ldots, h_{D-1})$ are the orthonormal log-ratio coordinates of composition $\mathbf{x}$ with respect to the basis $\mathscr{B}$. Egozcue et al. (2003) introduced an example of these coordinates where

$$h_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \; i = 1, \ldots, D-1, \tag{5}$$

whose corresponding basis is $\mathscr{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ with

$$\mathbf{v}_i = C\Big(\underbrace{e^{1/\sqrt{i(i+1)}}, \ldots, e^{1/\sqrt{i(i+1)}}}_{i}, 1/e^{\sqrt{i/(i+1)}}, \underbrace{1, \ldots, 1}_{D-(i+1)}\Big).$$

In this paper we use the coordinates in Equation 5 but any other orthonormal basis can also be considered. Determining which basis or coordinates are the most appropriate to solve a specific problem, is not straightforward. Nevertheless, the sequential binary partition introduced by Egozcue and Pawlowsky (2005) is a very useful tool to construct a particular basis to increase the interpretability of the corresponding coordinates.

One can define a pdf on the simplex by a pdf over the vector of orthonormal log-ratio coordinates. Indeed, let $f^*(\cdot\,;\boldsymbol{\theta}) : \mathbb{R}^{D-1} \to \mathbb{R}^+$ be a pdf defined on real space with parameters $\boldsymbol{\theta}$. Then, $f_{\mathscr{B}}(\mathbf{x}\,;\boldsymbol{\theta}) = f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x})\,;\boldsymbol{\theta})$ defines a pdf on the simplex, $f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}) : \mathscr{S}^D \to \mathbb{R}^+$, with respect to the Aitchison measure on $\mathscr{S}^D$. For example, fixing an orthonormal basis $\mathscr{B}$, the log-ratio normal distribution with parameters μ and $\boldsymbol{\Sigma}$ is defined as

$$f_{\mathscr{B}}(\mathbf{x}\,;\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{(D-1)/2}|\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{h}_{\mathscr{B}}(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{h}_{\mathscr{B}}(\mathbf{x})-\boldsymbol{\mu})}. \tag{6}$$

Note that it is a density on the simplex with respect to the Aitchison measure. The Aitchison measure, $d\lambda_a$, is a natural measure on $\mathscr{S}^D$, compatible with its Euclidean vector space structure (see Mateu-Figueras et al., 2013, for an in-depth discussion). This measure is absolutely continuous with respect to the Lebesgue measure on real space, $d\lambda$, and the relationship between them is $|d\lambda_a/d\lambda| = (\sqrt{D}x_1x_2\cdots x_D)^{-1}$.

Figure 1 (left) shows the contour lines of three normal distributions in the simplex $\mathscr{S}^3$. Note that the distribution in the centre of the ternary diagram is similar to the cir-

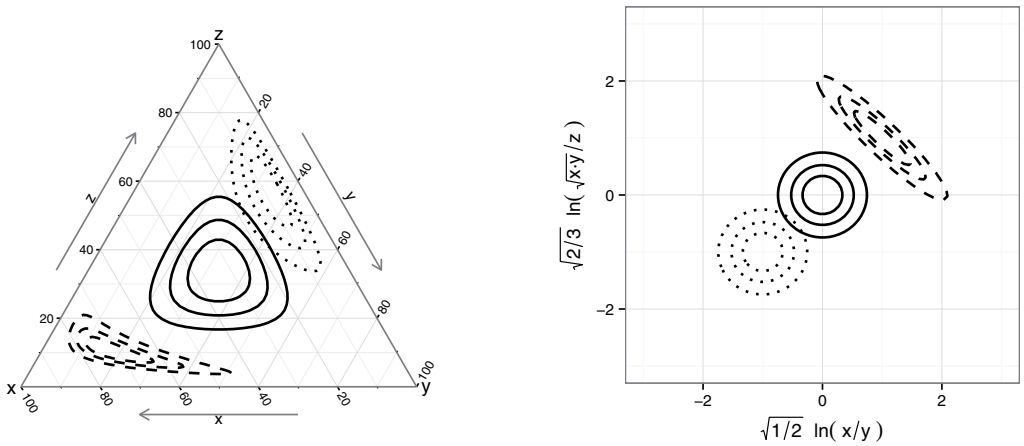


Figure 1: *Contour lines of typical log-ratio normal distribution on the simplex: (left) in the ternary diagram; (right) in log-ratio coordinates.*

cular contour lines in real space. However, note that, the farther the distribution from the centre is, the more different the contours from the traditional Gaussian shape are. These shapes are frequent in real data sets from industrial and scientific applications (Buccianti, 2011, Vives-Mestres et al., 2014). When these distributions are plotted using their orthonormal log-ratio coordinates (Figure 1 (right)) the traditional Gaussian contour lines are obtained. This idea can be applied by using other distributions on real space as, for example, the skew-normal (Mateu-Figueras and Pawlowsky-Glahn, 2007).

The well-known additive log-ratio vector (Aitchison, 1986) can be interpreted as the coordinates of a composition with respect to a non-orthogonal basis. Although the expression of the corresponding pdf is similar to Equation 6, the distances are not preserved among the additive log-ratio components and the principle of working on coordinates cannot always be applied (Mateu-Figueras et al., 2011). The equally well-known centred log-ratio vector (Aitchison, 1986) can be interpreted as the coordinates of a composition with respect to a generating system, not a basis. Despite the distances being preserved in this case, we do not recommend its use in a mixture model context because the fitted densities will be degenerate (Mateu-Figueras et al., 2011).

3. Modelling compositional data using traditional mixtures

When the goal is to fit a finite mixture model, the researcher can encounter different difficulties such as unbounded likelihood function, different local maximum, etc. The reader interested in knowing how to deal with these difficulties can consult McLachlan and Peel (2000) for an in-depth exposition. In this article we will indicate all the decisions taken in the process of fitting the finite mixtures.

3.1. Finite mixtures using traditional distributions defined on the real space

This approach assumes that $\mathscr{S}^D$ is a subset of $\mathbb{R}^D$ and its particular Euclidean space structure described in Section 2 is ignored. It is assumed that compositions are generated from a finite mixture distributions with pdf given by Equation 1 where $f(\,\cdot\,;\boldsymbol{\theta}_i) : \mathbb{R}^D \to \mathbb{R}^+$ is a pdf defined on the real space and with respect to the Lebesgue measure (e.g., a multivariate normal distribution or a t-student distribution). The main reason for using this approach is the simplicity of working without having to consider any restriction. However, this strategy exhibits some significant limitations and misleading results could be obtained.

When one uses traditional distributions defined on the real space, the mixture pdf is strictly positive in all the space, giving positive probability to impossible events. For example, the *impossible* event of having the i-th part negative has positive probability, i.e $P(\{\mathbf{x} \in \mathscr{S}^D | x_i < 0\}) > 0$. This difficulty is similar to the traditional confidence interval of a very small or very large proportion, i.e. it may provide lower or upper limit respectively beyond the restricted space.

Table 1: *CoDa set with three parts (a,b,c) from 20 compositions. (h_1, h_2) are its log-ratio coordinates. Two categorical covariates were considered: site and condition.*

a	**b**	**c**	h_1	h_2	**site**	**condition**
54.73	34.37	10.90	0.329	1.128	S1	C1
64.75	25.08	10.18	0.671	1.123	S1	C1
64.18	24.91	10.91	0.669	1.060	S1	C1
83.53	11.85	4.61	1.381	1.568	S1	C1
62.72	28.15	9.13	0.566	1.246	S1	C1
62.10	27.73	10.17	0.570	1.148	S1	C1
69.46	22.53	8.00	0.796	1.305	S1	C1
68.25	26.43	5.32	0.671	1.696	S1	C1
66.88	26.16	6.96	0.664	1.464	S1	C1
61.62	28.38	9.99	0.548	1.169	S1	C1
31.65	55.23	13.12	−0.394	0.946	S2	C1
24.32	61.47	14.21	−0.656	0.817	S2	C1
24.47	59.49	16.04	−0.628	0.708	S2	C1
18.75	68.00	13.25	−0.911	0.809	S2	C1
15.72	72.96	11.32	−1.085	0.895	S2	C1
18.83	32.85	48.32	−0.394	−0.542	S2	C2
12.11	30.61	57.27	−0.656	−0.890	S2	C2
10.75	26.14	63.10	−0.628	−1.082	S2	C2
10.31	37.38	52.31	−0.911	−0.800	S2	C2
8.15	37.81	54.05	−1.085	−0.918	S2	C2

In addition, this approach defined on the real space also ignores the constant sum constraint. Therefore, a further limitation is the collinearity that appears between parts after restricting the parts to sum a constant (Equation 2). This collinearity implies that the covariance matrix is singular, and therefore some methods can not be directly applied. Frequently, mixture models are estimated using the Expectation–Maximization (EM) algorithm (Dempster et al., 1977). In the E-step of the EM-algorithm a pdf computed from the sample is evaluated. Because most pdf depend on the inverse of the covariance matrix (e.g., multivariate normal and skew-normal), the common solution consists of removing one part of the composition for the rest of the analysis (e.g., Papageorgiou et al., 2001). However, this strategy may produce misleading results. For example, let $\mathbf{X}$ be the CoDa set recorded in Table 1. It is a simulated 3-part compositional data set representing proportions of 3 different elements, denoted a, b and c. Assume that the compositions come from two different locations, S_1 and S_2; and that they were collected under two possible weather conditions, C_1 and C_2. In addition, assume that it is well known that these weather conditions only affect part c: in condition C_1 the level of element c is lower than in condition C_2 (for example, element c is water and condition C_1 is a sunny day while condition C_2 is a rainy day). In this way, the compositions from row numbers 16 to 20 (Table 1) are the perturbed corresponding counterparts of

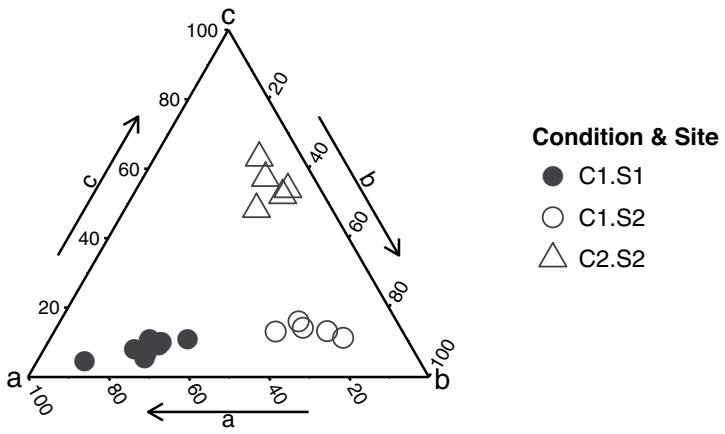


Figure 2: *CoDa set* $\mathbf{X}$ *in the ternary diagram. Filled and empty symbols are respectively used for data from location* S_1 *and* S_2*. Circles and triangles respectively correspond to condition* C_1 *and* C_2*.*

compositions from row numbers 11 to 15 after the perturbation $(1,1,r)$, where r is a random number depending on condition C_2. In this example we have modelled r as a lognormal random variable with parameters $\mu = 2$ and $\sigma = 0.25$. We have considered that condition C_1 and C_2 were an effect of the component c regardless of the magnitude of components a and b. Therefore, the effect of condition C_1 and C_2 could be modelled by means of a perturbation (Equation 3), which is a movement in the simplex with the Aitchison geometry.

The ternary diagram in Figure 2 shows that $\mathbf{X}$ is formed by three groups: the first group consists of the observations collected in site S_1 (filled circles), all of them collected under condition C_1; the second group with observations collected in site S_2 under condition C_1 (empty circles) and the third group with observations collected in site S_2 under condition C_2 (empty triangles). Suppose that an analyst, who is interested in fitting a traditional mixture model to $\mathbf{X}$, is not informed about the two different weather conditions and he or she only knows the information about the location. Because of the collinearity he/she decides to eliminate part c for the rest of the analysis. After eliminating part c, the researcher is working with the data set represented in Figure 3. This plot suggests that the analyst might conclude that $\mathbf{X}$ is formed by three mixture components as a result of the information collected in only the first two elements. This is a misleading conclusion because, by construction, we know that exclusively attending to the raw information provided by the first two elements the CoDa set $\mathbf{X}$ is formed by only two groups (one group for each location). But, when we work with proportions (a,b,c), despite part c having been eliminated, its effect (weather condition) is still present and interpretations about the nature of the groups based only on parts (a,b) may be misleading. An interested reader could find other examples about the misleadings conclusions and problems resulting from applying standard analysis to compositional data in Aitchison (1999, 2002).

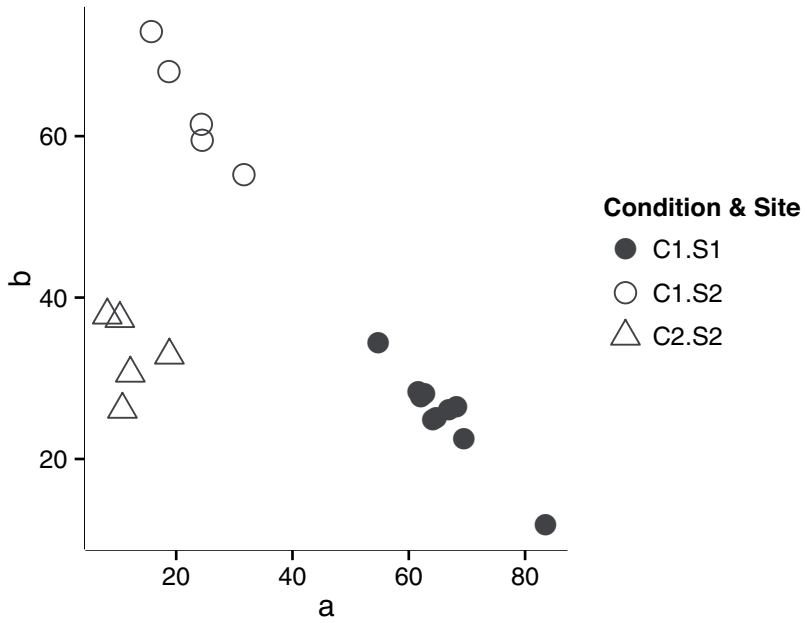


Figure 3: *Scatterplot of parts* (a,b) *of CoDa set* $\mathbf{X}$*. Filled and empty symbols are respectively used for data from location* S_1 *and* S_2*. Circles and triangles respectively correspond to condition* C_1 *and* C_2*.*

3.2. Finite mixtures using traditional distributions defined on the simplex

A finite mixture of distributions defined on the simplex is a probability distribution with pdf given by Equation 1 where $f(\,\cdot\,;\boldsymbol{\theta}) : \mathscr{S}^D \to \mathbb{R}^+$, is a pdf defined on the simplex. The Dirichlet distribution has been traditionally used as the probability distribution on $\mathscr{S}^D$. It can be obtained by the projection on the simplex of a random vector formed by independent and equally scaled gamma distributed parts. Despite its simplicity and its good mathematical properties, it has a very strong independence structure (Aitchison, 1986). In particular, any ratio x_i/x_j of two parts have to be independent from another ratio x_k/x_m formed from other two parts. In practice, such an independence structure cannot be assumed for most real data sets and consequently it heavily restricts the Dirichlet potential modelling application (Aitchison, 1986). To solve this difficulty, many generalizations of the Dirichlet distribution with less independence structure have been proposed: the Connor and Mosimann's distribution (Connor and Mosimann, 1969), the scaled Dirichlet distribution (Aitchison, 1986). In addition, Rayens and Srinivasan (1994) extend the Liouville distribution further to the generalized Liouville family. Later Smith and Rayens (2002), due to the limited applicability of the Liouville family of distributions, propose a generalization called Conditional Liouville distribution. Ongaro and Migliorati (2013) present the Flexible distribution, a generalization of the Dirichlet that exhibits greather flexibility in terms of dependence/independence structure and shape of the density. Finally, Monti et al. (2011) introduce the shifted-scaled Dirichlet distribution. This

generalized distribution is defined by adding the perturbation and powering operations (Equation 3) to the standard Dirichlet distribution. Unfortunately, all of these attempts have had limited success in fitting the general dependence structure of CoDa. Note that all these distributions are usually expressed through their density function with respect to the Lebesgue measure on $\mathscr{S}^D$ but the density with respect to the Aitchison measure could be easily obtained using the relationship between them (see Monti et al. (2011) for a detailed analysis of the implications of changing the measure).

In the literature different methods are found to estimate the parameters of a Dirichlet distribution. As it is an exponential family, the log-likelihood function is globally concave and a global optimum can be obtained. However, there is no closed form solution for the ML equations and numerical methods must be employed. According to Ng et al. (2011), the MLE via Newton-Raphson algorithm converges to the global optimum. Narayanan (1991) provides a Fortran subroutine with three different possibilities to estimate the initial parameter required. We can also obtain MLE estimates via the EM gradient methods (Ng et al., 2011). Recently the performance of different algorithms and starting value strategies to obtain the MLE of the Dirichlet parameters have been compared by Giordan and Wehrens (2015) using high-dimensional data. Nevertheless, the main problem is that final estimates can be outside the correct range for the parameters. Also, a large amount of iterations could be required to reach convergence. In practice, given a CoDa set, there is no straightforward method to fit a Dirichlet mixture or any of its generalizations. However, to obtain an approximation of the MLE estimator of a Dirichlet mixture, it is possible to apply the classification EM-algorithm (Celeux and Govaert, 1992) using any of the mentioned approaches to fit a Dirichlet model (see example in Section 5).

4. Modelling compositional data using a mixture of log-ratio distributions

To model CoDa using a finite mixture of log-ratio distributions, we consider

$$\pi_1 f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}_1) + \cdots + \pi_k f_{\mathscr{B}}(\,\cdot\,;\boldsymbol{\theta}_k) \tag{7}$$

where $f_{\mathscr{B}}(\mathbf{x};\boldsymbol{\theta}_i)$ are pdf's defined on the simplex with parameters $\boldsymbol{\theta}_i$, that is, they are densities defined considering the particular algebraic-geometric structure of the simplex defined in Section 2 and consequently are expressed with respect to the Aitchison measure. As indicated before and according to the principle of working on coordinates, we have

$$f_{\mathscr{B}}(\mathbf{x};\boldsymbol{\theta}) = f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x});\boldsymbol{\theta})$$

where $f^*(\,\cdot\,;\boldsymbol{\theta})$ are pdf on $\mathbb{R}^{D-1}$ for the orthonormal log-ratio coordinates vectors $\mathbf{h}_{\mathscr{B}}(\mathbf{x})$. Let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ be a CoDa set. Thus fitting the parameters $\pi_1,\ldots,\pi_k$ and $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k$ of Equation 7 using maximum likelihood estimators is equivalent to fitting the parameters in

$$\pi_1 f^*(\,\cdot\,;\boldsymbol{\theta}_1)+\cdots+\pi_k f^*(\,\cdot\,;\boldsymbol{\theta}_k) \tag{8}$$

using the data set $\mathbf{X}^{\mathrm{T}}=\{\mathbf{h}_{\mathscr{B}}(\mathbf{x}_1),\ldots,\mathbf{h}_{\mathscr{B}}(\mathbf{x}_n)\}$, that is, the log-ratio coordinates of the data set with respect to a selected orthonormal basis $\mathscr{B}$.

Indeed, the likelihood function evaluated for the CoDa set $\mathbf{X}$ is

$$\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_{\mathscr{B}}(\mathbf{x}_i;\boldsymbol{\theta}_j)=\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x}_i);\boldsymbol{\theta}_j). \tag{9}$$

Because the likelihood functions are the same, the maximum likelihood estimators $\hat{\pi}_1,\ldots,\hat{\pi}_k,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_k$ are also the same

$$\left(\hat{\pi}_1,\ldots,\hat{\pi}_k,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_k\right)=\underset{\pi_1,\ldots,\pi_k,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\arg\max}\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f_{\mathscr{B}}(\mathbf{x}_i;\boldsymbol{\theta}_j)= \tag{10}$$

$$=\underset{\pi_1,\ldots,\pi_k,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\arg\max}\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f^*(\mathbf{h}_{\mathscr{B}}(\mathbf{x}_i);\boldsymbol{\theta}_j). \tag{11}$$

Following this approach, we cannot obtain the misleading results shown in Section 3.1.. Taking the example from Section 3.1, we were interested in fitting a mixture to a sample $\mathbf{X}$ formed by parts a, b and c (Table 1). Instead of eliminating one part, now the analyst decides to express parts a, b and c in log-ratio coordinates. Before starting the analysis, a basis $\mathscr{B}$ of $\mathscr{S}^3$ is selected, for example

$$\mathscr{B}=\left\{C\left(e^{1/\sqrt{2}},1/e^{\sqrt{1/2}},1\right),C\left(e^{1/\sqrt{6}},e^{1/\sqrt{6}},1/e^{\sqrt{2/3}}\right)\right\}, \tag{12}$$

and the compositions of $\mathbf{X}$ are expressed in terms of their coordinates $\mathbf{X}^{\mathrm{T}}$ ($h_1=\sqrt{1/2}\ \ln(a/b)$ and $h_2=\sqrt{2/3}\ln(\sqrt{ab}/c)$) (see Table 1). Figure 4 shows the plot of these coordinates where the different effect of the location (parts a and b) and the weather conditions (part c) are highlighted. Note that the compositions from S_2 under condition C_1 take the same value in the first coordinate as their counterparts under condition C_2.

In this case the interpretations based only in terms of parts a and b will not be misleading. In fact, if the analyst also decides to remove part c, a basis $\mathscr{B}'$ of $\mathscr{S}^2$ is selected as:

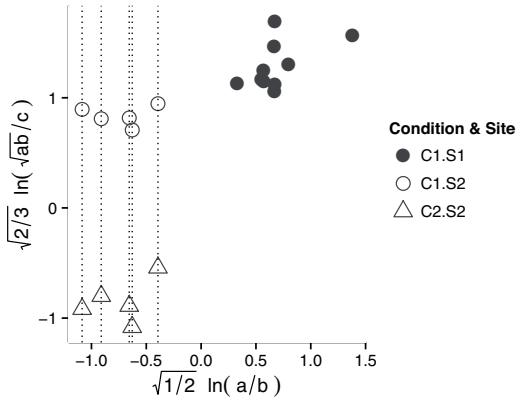


Figure 4: *Scatterplot of log-ratio coordinates for the CoDa set* **X**. *Filled and empty symbols are respectively used for data from location S_1 and S_2. Circles and triangles respectively correspond to condition C_1 and C_2.*

$$\mathscr{B}' = \left\{ C\left(e^{1/\sqrt{2}}, 1/e^{\sqrt{1/2}}\right) \right\}.$$

In this way, the corresponding coordinate h_1 is the same as before. Figure 5 shows the histograms of coordinate h_1 separated by weather conditions in two stratas. Note that, regardless of the condition, all the data collected in S_2 take the same value, forming one cluster (between -1 and 0). On the other hand, the compositions collected in S_1 are close to one.

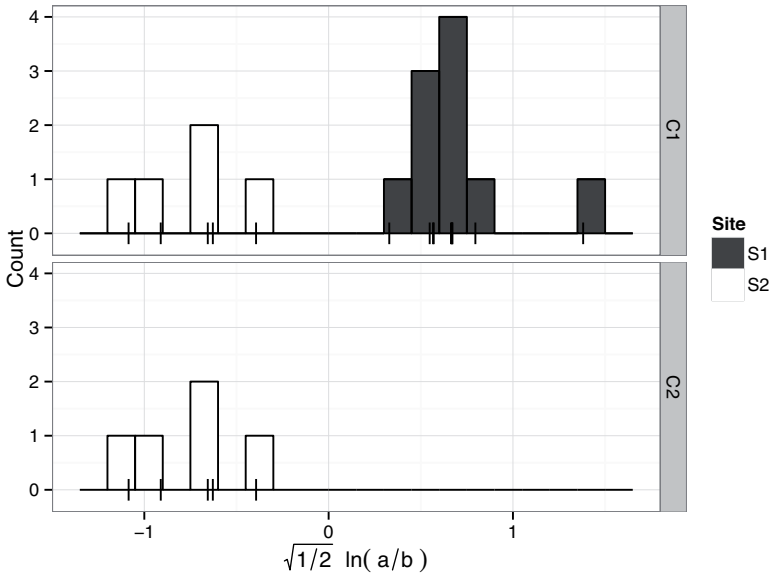


Figure 5: *Histograms of first log-ratio coordinate for CoDa set* **X**. *Two stratas correspond to weather conditions.*

In Equations 9 and 10, we fit the mixture using the coordinates $\mathbf{h}_{\mathscr{B}}(\mathbf{x})$ with respect to a specific basis $\mathscr{B}$ but any other orthonormal basis could have been chosen as well. Thus, in any compositional analysis involving coordinates, it is important to check the invariance of the results under changes of basis. When fitting a mixture of log-ratio distributions, it is enough to check that the family of distributions used to fit the mixture is basis invariant, that is, it satisfies the following definition.

Definition 1 *Let $\mathscr{B}_1$ and $\mathscr{B}_2$ be two basis on $\mathscr{S}^D$. Let $\boldsymbol{\Theta}$ be a parameter space for a probability density function $f^* : \mathbb{R}^{D-1} \to \mathbb{R}^+$. A probability density function f^* is* basis invariant *if for any two different basis $\mathscr{B}_1$, $\mathscr{B}_2$, for any parameters $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, there are parameters $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ such that*

$$f^*(\boldsymbol{h}_{\mathscr{B}_1}(\boldsymbol{x});\boldsymbol{\theta}_1) = f^*(\boldsymbol{h}_{\mathscr{B}_2}(\boldsymbol{x});\boldsymbol{\theta}_2).$$

Most common distributions are basis invariant when we do not restrict the parameters. For example, the log-ratio normal distribution (Equation 6) is formulated in terms of Mahalanobis distance and of covariance matrix determinant, that are both invariant elements under change of basis (Barceló-Vidal et al., 1999). Moreover, using the linear transformation property (Azzalini and Capitanio, 1999), it can easily be proved that the multivariate log-ratio skew-normal distribution is also invariant under change of basis.

5. A real data set: Forensic Glass

To illustrate and compare the different described approaches, we analysed the USA Forensic Science Service data set, also known as the Forensic Glass data set. This data is available from the UCI Machine Learning Repository (Bache and Lichman, 2013). The data set is composed of 214 fragments of glass samples where the percentages of eight chemical elements were measured. The fragments of glass were originally come from seven types of glass. In order to easily display the results using ternary diagrams and bivariant plots, we only consider three chemical elements: Calcium (Ca), Silica (Si) and Aluminium (Al). For simplicity, we only consider three types of glass (containers, vehicle headlamps and vehicle windows) but all types of glass could be considered and lead to similar conclusions. We call this data set the Reduced Forensic Glass data set (Table 2). Figure 6 shows this data set formed by 59 glass samples in the ternary diagram. We can see that the types of glass do not form well-separated groups and consequently there will be a weak relation between the components of the mixture and the types of glass. This was already observed by Venables and Ripley (2002) in a discriminant context.

We fit a mixture model using the normal distribution on real space, the Dirichlet distribution and the log-ratio normal and skew-normal distributions on the simplex. For all cases the index BIC indicates that $k = 3$ are the optimal number of components

Table 2: *Reduced Forensic Glass data set: parts (Ca, Si, Al) and its log-ratio coordinates. The categorical covariate (type) shows the provenance of glass.*

Ca	**Si**	**Al**	h_1	h_2	**type**
10.43	88.23	1.35	−1.510	2.541	Veh
10.12	88.26	1.63	−1.531	2.375	Veh
10.23	88.10	1.67	−1.523	2.359	Veh
10.31	88.06	1.63	−1.517	2.382	Veh
10.14	87.73	2.13	−1.526	2.155	Veh
11.60	87.39	1.01	−1.428	2.818	Veh
10.81	88.40	0.79	−1.486	2.994	Veh
10.12	88.40	1.48	−1.533	2.455	Veh
10.63	87.79	1.58	−1.493	2.418	Veh
10.36	88.12	1.52	−1.514	2.441	Veh
10.48	87.97	1.55	−1.504	2.429	Veh
11.77	87.53	0.71	−1.419	3.112	Veh
10.67	87.48	1.85	−1.488	2.290	Veh
10.69	87.33	1.98	−1.485	2.234	Veh
10.87	87.26	1.86	−1.473	2.292	Veh
10.80	88.29	0.91	−1.486	2.878	Veh
11.23	87.66	1.12	−1.453	2.721	Veh
7.41	88.18	4.42	−1.751	1.433	Con
11.92	85.88	2.20	−1.396	2.186	Con
13.29	84.89	1.82	−1.311	2.380	Con
13.41	84.78	1.80	−1.304	2.393	Con
13.26	84.84	1.90	−1.312	2.344	Con
11.84	86.03	2.13	−1.402	2.210	Con
13.15	84.81	2.04	−1.318	2.282	Con
14.23	83.94	1.84	−1.255	2.395	Con
8.65	87.57	3.78	−1.637	1.621	Con
8.59	87.66	3.74	−1.643	1.627	Con
14.51	83.87	1.63	−1.241	2.501	Con
11.54	85.88	2.58	−1.419	2.043	Con
13.08	85.17	1.75	−1.325	2.407	Con
6.78	90.96	2.26	−1.836	1.957	Head
7.31	89.89	2.80	−1.774	1.808	Head
10.71	87.80	1.49	−1.488	2.469	Head
11.89	85.60	2.51	−1.396	2.076	Head
10.72	87.65	1.63	−1.486	2.396	Head
10.38	87.48	2.14	−1.507	2.160	Head
10.38	86.80	2.82	−1.502	1.931	Head
10.60	86.13	3.27	−1.481	1.816	Head
10.21	87.40	2.39	−1.518	2.062	Head
10.17	87.47	2.36	−1.522	2.071	Head
10.65	86.20	3.15	−1.479	1.848	Head

Table 2 (cont.)

Ca	Si	Al	h_1	h_2	type
11.05	85.97	2.98	−1.451	1.908	Head
10.58	86.65	2.77	−1.487	1.953	Head
10.70	86.16	3.14	−1.475	1.853	Head
10.46	86.56	2.97	−1.494	1.891	Head
9.92	87.41	2.68	−1.539	1.957	Head
10.47	88.14	1.40	−1.506	2.513	Head
9.93	87.21	2.86	−1.536	1.903	Head
9.93	87.68	2.39	−1.540	2.052	Head
10.33	86.97	2.69	−1.506	1.968	Head
10.32	87.52	2.16	−1.512	2.150	Head
10.36	87.40	2.24	−1.508	2.121	Head
7.97	89.78	2.24	−1.712	2.025	Head
11.11	85.67	3.22	−1.444	1.845	Head
10.84	85.76	3.40	−1.463	1.791	Head
10.07	87.55	2.38	−1.529	2.061	Head
10.06	87.53	2.41	−1.530	2.050	Head
10.09	87.60	2.31	−1.528	2.086	Head
10.25	87.27	2.47	−1.514	2.036	Head

except for the Dirichlet distribution whose optimal value is for $k = 5$. For illustration purposes and in order to easily compare all described approaches, we will use $k = 3$ for all different cases. For each mixture approach, we fit the mixture 100 times using different starting points to avoid local maximums.

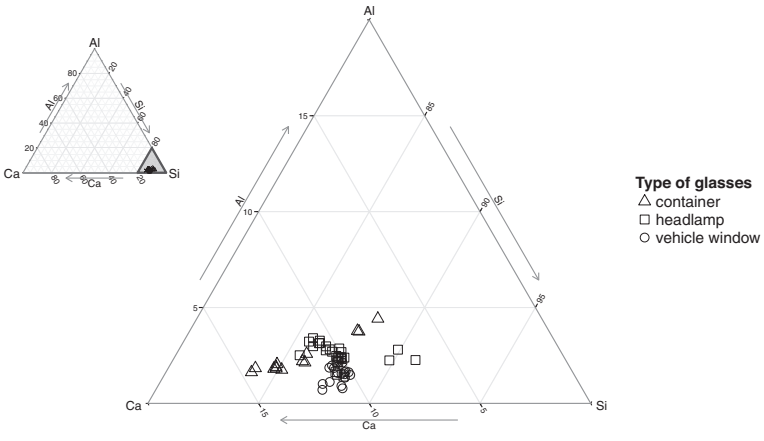


Figure 6: *Reduced Forensic Glass data set in ternary diagram: Calcium (Ca), Silica (Si) and Aluminium (Al) chemical elements. Three groups of glass: containers (circles), headlamps (triangles) and vehicle windows (squares). The large ternary diagram is a zoom of the shadow area seen in the smaller initial ternary diagram.*

Using the traditional approach introduced in Section 3.1 we fit a mixture of distributions on real space with three mixture components. In particular we choose a traditional Gaussian mixture. As mentioned, we need to eliminate one part to avoid the constant sum constraint. For example, when we removed the Calcium (Ca) part, the corresponding mixture model (BIC = -763.4) obtained is $\pi_1 f(\cdot\,;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+\pi_2 f(\cdot\,;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+\pi_3 f(\cdot\,;\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ with estimates

$$\hat{\pi}_1 = 0.12,\ \ \hat{\boldsymbol{\mu}}_1 = (88.76, 1.65),\ \ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.66 & 0.81 \\ 0.81 & 0.52 \end{pmatrix},$$

$$\hat{\pi}_2 = 0.38,\ \ \hat{\boldsymbol{\mu}}_2 = (85.85, 2.68),\ \ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.17 & 0.72 \\ 0.72 & 0.58 \end{pmatrix},$$

$$\hat{\pi}_3 = 0.5,\ \ \hat{\boldsymbol{\mu}}_3 = (87.67, 1.97) \text{ and } \ \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.16 & -0.18 \\ -0.18 & 0.27 \end{pmatrix}.$$

Figure 7 (top-left) shows the isodensity curves for the fitted mixture of Gaussian distributions. Figure 7 (top-right and bottom-left) also shows the isodensity curves of the finite mixture when the parts removed were Aluminium (Al) and Silica (Si), respectively. The dashed lines represent the limit of the simplex, i.e. the region were restrictions given by Equation 2 are held. In Figure 7 (bottom-right) the isodensity curves have been completed to be represented in the ternary diagram. Note that the distribution is giving positive probability to impossible regions.

Despite the fact that in Gaussian mixtures the maximum likelihood function is invariant whatever part is removed, we stated that in practice the numerical algorithm gets stuck in a local optimum. That is, the invariance of the results is not guaranteed, and different mixtures may be obtained depending on the part removed.

A Dirichlet probability distribution is specified by the parameters $\boldsymbol{\alpha} = \left(\alpha^1,\ldots,\alpha^D\right)$. Therefore, to fit a mixture of K Dirichlet distributions the parameters π_1, ..., π_K and $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K$ need to be estimated. To make this estimation we approximated the MLE estimator of a Dirichlet mixture using the EM-algorithm proposed by Celeux and Govaert (1992). The mixture of Dirichlet distributions obtained (BIC =-732.9) was $\pi_1 f(\cdot\,;\boldsymbol{\alpha}_1)+\pi_2 f(\cdot\,\boldsymbol{\alpha}_2)+\pi_3 f(\cdot\,;\boldsymbol{\alpha}_3)$ with estimates

$$\hat{\pi}_1 = 0.37, \qquad \hat{\boldsymbol{\alpha}}_1 = (281.2, 2343.1, 71.6),$$

$$\hat{\pi}_2 = 0.15, \qquad \hat{\boldsymbol{\alpha}}_2 = (272.9, 1777.2, 41.2),$$

$$\hat{\pi}_3 = 0.48 \text{ and } \quad \hat{\boldsymbol{\alpha}}_3 = (34.6, 304.3, 6.3).$$

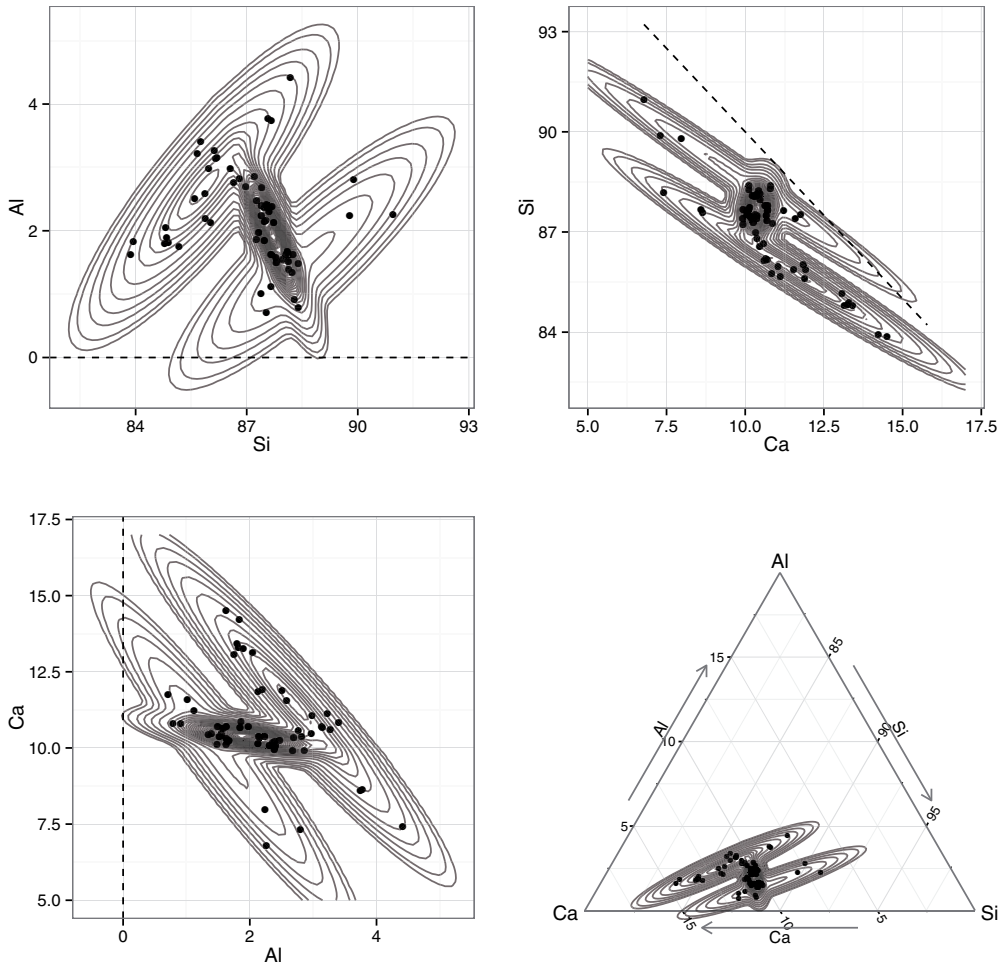


Figure 7: *Reduced Forensic Glass data set. On the top-left, top-right and bottom-left isodensity curves for mixtures of Gaussian distributions in R^2 after removing the Ca, the Al and the Si part respectively. On bottom-right the isodensity curves transformed into the simplex.*

Note that for $k = 3$ the Dirichlet BIC value is worse than the value for the normal distribution. Using the Dirichlet parameter estimates we can, respectively, obtain the centre of each mixture component in the simplex: $(10.43, 86.91, 2.66)$, $(13.05, 84.98, 1.97)$ and $(10.02, 88.15, 1.83)$, expressed in percentages.

Figure 8 shows how the Dirichlet mixture fits the data set. Due to the strong independence structure of the Dirichlet model (noted above in Section 3.2), the density can only take nearly elliptical shapes. Consequently, the mixture obtained cannot capture non-elliptical forms of variability.

Finally, we use the log-ratio approach introduced in Section 4. To fit a mixture of log-ratio distributions it is necessary first to express each composition with respect to a

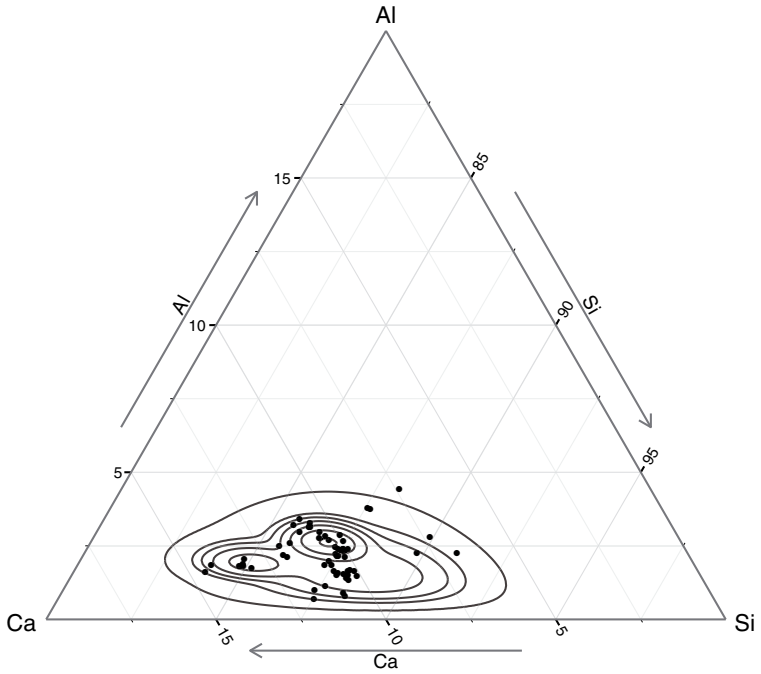


Figure 8: *Reduced Forensic Glass data set: classification given by a standard Dirichlet mixture model.*

basis of $\mathscr{S}^3$. Consider the same basis $\mathscr{B}$ defined in Equation 12. Table 2 contains the data set expressed in log-ratio coordinates with respect to basis $\mathscr{B}$, resulting in coordinates $h_1 = \sqrt{1/2}\ln(\mathrm{Ca}/\mathrm{Si})$ and $h_2 = \sqrt{2/3}\ln(\sqrt{\mathrm{Ca}\cdot\mathrm{Si}}/\mathrm{Al})$.

Fitting a Gaussian mixture to the log-ratio coordinates (BIC = -84.3) results in mixture model $\pi_1 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 f_{\mathscr{B}}(\cdot\,;\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ with estimates

$$\hat{\pi}_1 = 0.59,\ \ \hat{\boldsymbol{\mu}}_1 = (-1.5, 2.31),\ \ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 8e-04 & 0.0059 \\ 0.0059 & 0.0949 \end{pmatrix},$$

$$\hat{\pi}_2 = 0.1,\ \ \hat{\boldsymbol{\mu}}_2 = (-1.73, 1.75),\ \ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.005 & -0.0059 \\ -0.0059 & 0.0422 \end{pmatrix},$$

$$\hat{\pi}_3 = 0.31,\ \ \hat{\boldsymbol{\mu}}_3 = (-1.39, 2.12) \text{ and } \ \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.0065 & 0.0186 \\ 0.0186 & 0.0581 \end{pmatrix}.$$

Note that the difference between the BIC value for the log-ratio normal distribution and the previous distributions seems to be unusually large. However, these values can not be directly comparable because the latter is calculated using log-ratio coordinates. In Figure 9 the isodensity curves of the log-ratio normal distribution are represented in

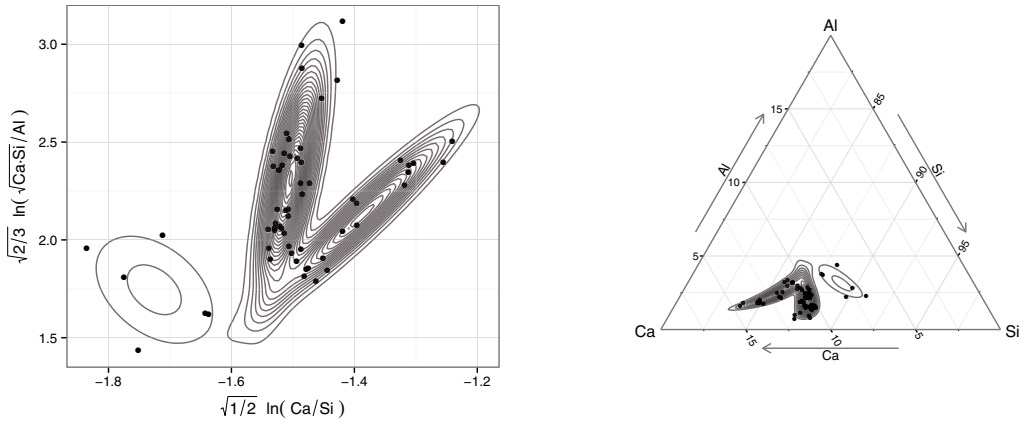


Figure 9: *Log-ratio Gaussian mixtures for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.*

the space of coordinates (left) and in the ternary diagram (right). Looking at the coordinate space, we see that this mixture can model elliptical forms of variability and consequently, on the simplex the estimated mixture is able to model those typical arc shaped forms (Figure 9 (right)). Because multivariate log-ratio normal is basis invariant (Section 4), working with another orthonormal log-ratio basis results in the same mixture as that represented in the ternary diagram (Figure 9 (right)). As noted above, there is low similarity between mixture components and types of glass. In this case the adjusted Rand index (Hubert and Arabie, 1985) is equal to 0.219.

Note that the parameters of the mixture are expressed with respect to coordinates h_1 and h_2. To better interpret the parameters of the mixture, we back-transformed the parameters $\boldsymbol{\mu}_i$ into the simplex: $(10.46, 87.75, 1.79)$, $(7.77, 89.13, 3.10)$ and $(12.02, 85.59, 2.39)$, into percentages. Note that only the centre of the first log-ratio normal mixture component is similar to the centre of the first Dirichlet mixture component. To better interpret the covariance parameter $\boldsymbol{\Sigma}_i$, Aitchison (1986) proposes using the variation matrix, that is, the variance of each log-ratio. In this case, the corresponding log-ratio variances are shown in Table 3.

The first mixture component is characterised by the highest relative variability of the ratio between the Calcium and Aluminium parts and lowest between the Calcium and

Table 3: *Forensic Glass data set: log-ratio variances for each mixture component fitted by a log-ratio Gaussian mixture.*

Mixture component	var(ln(Ca/Si))	var(ln(Ca/Al))	var(ln(Si/Al))
1	0.0016	0.1530	0.1324
2	0.0101	0.0556	0.0760
3	0.0131	0.1226	0.0582

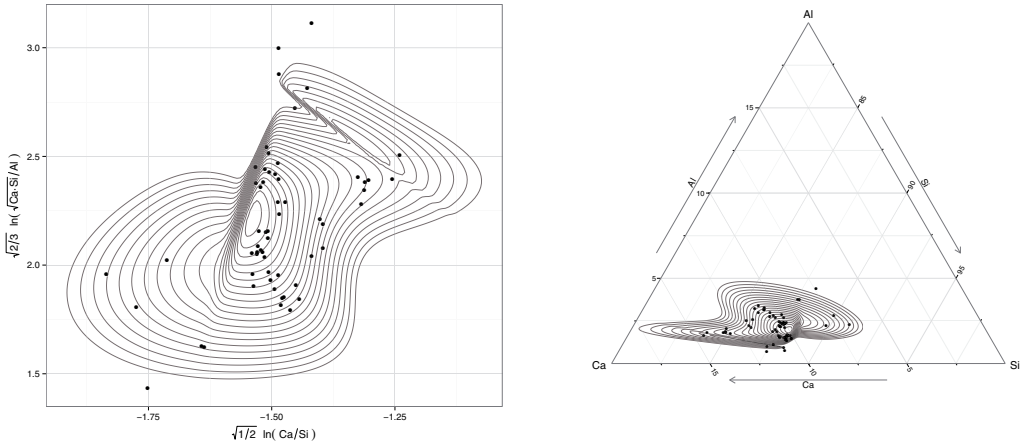


Figure 10: *Log-ratio skew normal mixture adjusted for Forensic Glass data set: (left) in log-ratio coordinates; (right) in the ternary diagram.*

Silica elements. Due to var(ln(Ca/Si)) being close to zero, the concentration of these elements are nearly proportional (Martín-Fernández et al., 2015). Note that this behaviour is common across the three mixture components. All the variances take small values for the second mixture component, while the third mixture component differs from the first due to the small value in the variance of ln(Si/Al).

Following an analogous approach, it is possible to fit other non-Gaussian models. For example, in Figure 10 the data set is modelled with a mixture of multivariate log-ratio skew-normal distributions using the package provided by Prates et al. (2013) (BIC = -62.3). The log-ratio skew-normal model extends the modelling possibilities because it contains the log-ratio normal model as a particular case. Nevertheless, the final model is more complex because a skew parameter is added for each density in the mixture. This complexity also contributes to the BIC value which is worse than the value for the log-ratio normal distribution. For the sake of brevity, we prefer not to give the estimated parameters here. The multivariate log-ratio skew-normal model is also basis invariant, thus working with another orthonormal log-ratio basis results in the same mixture as that represented in the ternary diagram (Figure 10 (right)). Although the adjusted Rand index increased slightly to 0.348, there is low similarity between mixture components and types of glass.

6. A second real data set: C-horizon of the Kola data set

To illustrate how to proceed when the number of parts is greater than three, we analysed a reduced data set of the C-horizon of the Kola data set (Reimann, Filzmoser). We selected a subsample formed by 69 observations belonging to three groups: Alkaline (7), Sediments (39) and Granite (23). For these samples we created the subcomposition

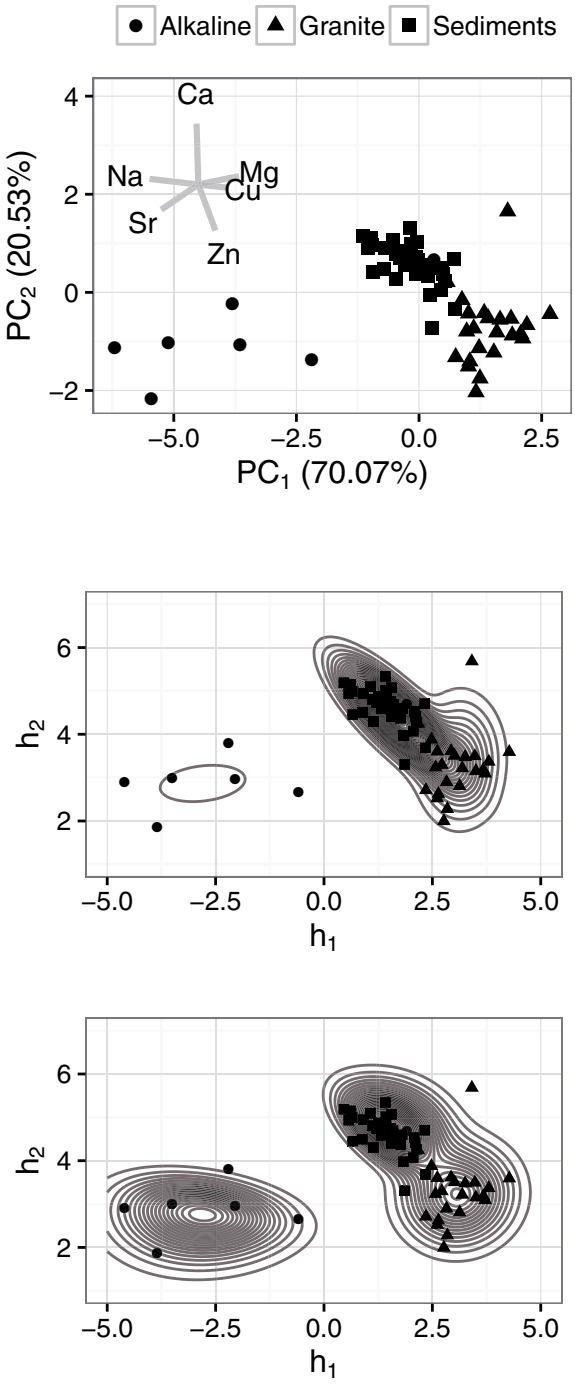


Figure 11: *Mixtures adjusted to the reduced C-horizon of Kola data set: (top) compositional biplot; (middle) marginal of the log-ratio Gaussian mixture for the two first coordinates: h_1 and h_2; (bottom) marginal of the log-ratio skew normal mixture for the two first coordinates.*

formed by the chemical elements: Calcium (Ca), Copper (Cu), Magnesium (Mg), Sodium (Na), Strontium (Sr) and Zinc (Zn).

Figure 11 (top) shows the compositional biplot, which consists of a principal component plot applied to the centred log-ratio coordinates. The two principal components explain a 90.6% variance, which is a high percentage of the total variance of the sample. The first principal axis (PC_1) is associated to the relative variation in parts Na and Sr as opposed to Mg and Cu. On the other hand, the axis of the PC_2 is associated to the relative variation of element Ca versus Zn. The group of Alkaline observations has a high concentration of elements Na and Sr with respect to the proportion in the groups Granite and Sediments that have a high concentration of Mg and Cu elements. The main differences between the groups Granite and Sediments is that the former has a higher proportion of the element Ca, whilst the latter has high concentration in the Zn part.

We fit a mixture model using the normal and the skew-normal distributions on log-ratio coordinates. For the sake of brevity, the estimated parameters are not provided. In both cases the BIC index indicates that $k = 3$ is the optimal number of components. To avoid local maximums we recalculated the parameters for each mixture until no improvement was obtained in the likelihood function during 100 simulations. To calculate the orthonormal log-ratio coordinates in this example we considered the orthonormal basis $\mathscr{B}$ formed by the directions of the principal components.

Figure 11 (middle) shows the marginal of the adjusted log-ratio normal mixture with respect to the first (h_1) and second (h_2) orthonormal log-ratio coordinates. For the log-ratio normal distributions the Rand index was 0.580, with 29 observations misclassified. In Figure 11 (bottom) the marginal (h_1, h_2) of the adjusted log-ratio skew normal mixture is shown. In this case the Rand index is better (0.760) and the misclassification rate is also improved because only 5 observations were misclassified.

7. Final remarks

Traditional distributions in finite mixtures for compositional data sets show significant difficulties. If densities for real data are used, probabilities of impossible events are obtained. Additionally, as a part of a composition is often removed to estimate the model, the results depend on that part. Dirichlet density and some generalizations on the simplex can not capture the variability of many compositional data sets due to their strong independence structure. The proposed log-ratio models are defined on the simplex using its particular algebraic-geometric structure. Consequently probabilities for impossible events are not obtained and there is no need to eliminate any part. The log-ratio normal model is a flexible model that can describe different forms of variability and dependence structures. It is a simple model and provides a rich enough parametric class of distributions on the appropriate sample space. Certainly, the model has the equivalent limitations as the traditional Gaussian mixtures in real space. Nevertheless, the proposed methodology allows different and alternative models. Indeed, any mixture model

defined on the real space can be considered to model data on the simplex space using the principle of working on coordinates. In this paper we have proposed a mixture of normal and skew-normal distributions to the log-ratio coordinates of a compositional sample. These two options extend the range of possibilities we have had up to now with the Dirichlet model or its generalizations. Interestingly, both proposed log-ratio models are invariant with respect to the orthonormal basis chosen to compute the log-ratios. The proposed log-ratio methodology could be extended by studying the possibilities of other known distributions on real space, like Student-t and skewed-t mixtures. Furthermore, in a non-parametric context, an analogy of these models with the P-spline methodology for CoDa should be explored Eilers et al. (2015).

Acknowledgments

This research was supported by the Ministerio de Economía y Competividad through the projects "METRICS" and "CoDa-RETOS" (MTM2012-33236; MTM2015-65016-C2-1-R: MINECO/FEDER,UE) and the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR: 2014SGR551). The authors gratefully acknowledge the constructive comments of the anonymous referees which have undoubtedly helped to signficanlty improve the quality of the paper.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London (UK). Reprinted in 2003 by Blackburn Press.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology*, 31, 563–580.

Aitchison, J. (2002). Simplicial inference. In *Algebraic Methods in Statistics and Probability* (ed. Viana MA and Richards DS), vol 287. Contemporary Mathematics Series: American Mathematical Society, Providence, RI (USA), 1–22.

Albert, J. H. and Gupta, A. K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables. *The Annals of Statistics*, 10, 1261–1268.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Statistics and Computing*, 22, 1021–1029.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 579–602.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 367–389.

Bache, K. and Lichman, M. (2013). UCI machine learning repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6, 1345–1382.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49, 803–821.

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (1999). Comment on "Singularity and nonnormality in the classification of compositional data" by Bohling, G. C., Davis, J. C., Olea, R. A. and Harff, J. *Mathematical Geology*, 31, 581–585.

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In Rastogi, R., Morik, K., Bramer, M., and Wu, X., editors, *ICDM 2004, fourth IEEE International Conference on Data Mining*, 19–26, Brighton. IEEE Computer Society.

Bouguila, N. (2011). Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22, 186–198.

Bouguila, N., Ziou, D. and Vaillancourt, J. (2004). Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13, 1533–1543.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71, 52–78.

Browne, R. P. and McNicholas, P. D. (2013). A mixture of generalized hyperbolic distributions. ArXiv e-prints arXiv:1305.1036

Buccianti, A. (2011). *Natural Laws Governing the Distribution of the Elements in Geochemistry: The Role of the Log-Ratio Approach*, 255–266. John Wiley and Sons, Ltd.

Calif, R., Emiliol, R. and Soubdhan, T. (2011). Classification of wind speed distributions using a mixture of Dirichlet distributions. *Renewable Energy*, 36, 3091–3097.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14, 315–332.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64, 194–206.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37, 795–828.

Eilers, P.H.C., Marx, B.D. and Durban, M. (2015). Twenty years of P-splines. *SORT*, 39, 149–186.

Ferrer-Rosell, B., Coenders, G., and Martínez-García, E. (in press). Segmentation by tourist expenditure composition. An approach with compositional data analysis and latentclasses. *Tourism Analysis*.

Giordan, M. and Wehrens, R. (2015). A comparison of computational approaches for maximum likelihood estimation of the Dirichlet parameters on high-dimensional data. *SORT*, 39, 109–126.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.

Lee, S. X. and McLachlan, G. J. (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. ArXiv e-prints arXiv:1109.4706

Lee, S. X. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24, 181–202.

Lin, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, 20, 343–356.

Mardia, K. V., Taylor, C. C. and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63.

Martín-Fernández, J. A., Daunis-i-Estadella, J. and Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *SORT*, 39, 231–252.

Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2007). The skew-normal distribution on the simplex. *Communications in Statistics-Theory and Methods*, 36, 1787–1802.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. In *Compositional Data Analysis*, 29–42. John Wiley and Sons, Ltd.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *SORT*, 37, 29–56.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models, Willey Series in Probability and Statistics*. John Wiley and Sons, New York.

Monti, G. S., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2011). Notes on the scaled Dirichlet distribution. In *Compositional Data Analysis*, 128–138. John Wiley and Sons, Ltd.

Monti, G. S., Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The shifted-scaled Dirichlet distribution in the simplex. In Egozcue, J. J., Tolosana-Delgado, R. and Ortego, M. I., editors, *CoDaWork 2011, the 4th International Workshop on Compositional Data Analysis*, Sant Feliu de Guíxols. CIMNE.

Narayanan, A. (1991). Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 40, 365–374.

Ng, K. W., Tian, G.-L. and Tang, M.-L. (2011). *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley and Sons.

Neocleous, T., Aitken, C. and Zadora, G. (2011). Transformations for compositional data with zeros with an application to forensic evidence evaluation. *Chemometrics and Intelligent Laboratory Systems*, 109, 77–85.

Ongaro, A. and Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114, 412–426.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Buccianti, A. (2014). Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *Journal of Geochemical Exploration*, 141, 71–77.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Soto, J. A. (2012). Dealing with distances and transformations for fuzzy C-means clustering of compositional data. *Journal of Classification*, 29, 144–169.

Papageorgiou, I., Baxter, M. J. and Cau, M. A. (2001). Model-based cluster analysis of artefact compositional data. *Archaeometry*, 43, 571–588.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15, 384–398.

Prates, M. O., Lachos, V. H. and Cabral, C. R. B. (2013). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna, Austria.

Rayens, W. S. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the Simplex. *Journal of the American Statistical Association*, 89, 1465–1470.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R. (2011). *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. John Wiley and Sons Ltd, Chichester (UK).

Scealy, J. L., Patrice de Caritat, Grunsky, E. C., Tsagris, M. T and Welsh, A. H. (2015). Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, 136–148, DOI: 10.1080/01621459.2014.990563.

Scott, A. and Symons, M. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 387–397.

Smith, B. and Rayens, W. (2002). Conditional generalized Liouville distributions on the simplex. *Statistics*, 36, 185–194.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York.

Vives-Mestres, M., Daunis-i Estadella, J. and Mártin-Fernández, J. A. (2014). Individual T-2 control chart for compositional data. *Journal of Quality Technology*, 46, 127–139.

Wang, H., Liu, Q., Mok, H. M. K., Fu, L. and Tse, W. M. (2007). A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179, 459–468.

SORT 40 (2) July-December 2016, 375-398

Smoothed landmark estimators of the transition probabilities

Luís Meira-Machado

Abstract

One important goal in clinical applications of multi-state models is the estimation of transition probabilities. Recently, landmark estimators were proposed to estimate these quantities, and their superiority with respect to the competing estimators has been proved in situations in which the Markov condition is violated. As a weakness, it provides large standard errors in estimation in some circumstances. In this article, we propose two approaches that can be used to reduce the variability of the proposed estimator. Simulations show that the proposed estimators may be much more efficient than the unsmoothed estimator. A real data illustration is included.

MSC: 62N02, 62G05, 62P10.

Keywords: Kaplan-Meier, Multi-state model, Nonparametric estimation, Presmoothing, Survival Analysis.

1. Introduction

The analysis of survival data may be described by the Markov process considering the transition from an initial 'alive' state to a single ultimate state or endpoint 'dead'. However, in most longitudinal medical studies more than one endpoint can be defined. In breast cancer trials, for instance, several endpoints, such as disease-free survival, local recurrence, distant metastasis or death are possible. Multi-state models are a useful way of describing such a process in which an individual moves through a number of finite states in continuous time. A wide range of medical situations have been modeled using multi-state methods, for example, HIV infection and AIDS (Gentleman et al., 1994), liver cirrhosis (Andersen et al., 2002), breast cancer (Pérez-Ocón et al., 2001; Putter et al., 2007) and problems following heart transplantation (Meira-Machado et al., 2009). A commonly-used model is the illness-death model, with three states representing health, illness and death (Figure 1). Individuals start in the healthy state and subsequently move

Centre of Mathematics and Department of Mathematics and Applications, University of Minho, Guimarães, Portugal. lmachado@math.uminho.pt

Received: March 2016
Accepted: November 2016

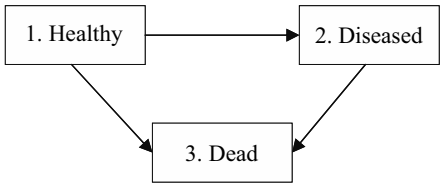


Figure 1: *Illness-death model.*

either to the diseased state or to the dead absorbing state. In the irreversible version of this model, individuals in the diseased state will eventually move to the dead state without any possibility of recovery. Methods developed for the progressive illness-death model have a wide range of applications in longitudinal medical studies.

One important feature of multi-state models is their ability to obtain predictions of the clinical prognosis of a patient at a certain point in his/her recovery or illness process. Various aspects of the model dynamics can be captured through the transition probabilities. Traditionally, the transition probabilities are estimated using Aalen-Johansen estimator (Aalen and Johansen, 1978) which assumes the process to be Markovian. Substitute estimators for the Aalen-Johansen estimator for a general non-Markov illness-death process without recovery were introduced by Meira-Machado et al. (2006). These authors showed that the new estimators may behave much more efficiently than the Aalen-Johansen when the Markov assumption does not hold. This work has been revisited by Allignol et al. (2014), who proposed a closely related non-Markov estimator too. However, both of the proposed non-Markov estimators have the drawback of requiring that the support of the censoring distribution contains the support of the lifetime distribution, which is not often the case. To avoid this problem, de Uña-Álvarez and Meira-Machado (2015) propose new estimation methods which are consistent regardless the Markov condition and the referred assumption about the censoring support. The idea behind the proposed methods is to use specific subsamples or portions of data at hand (namely, those observed to be in a given state at a pre-specified time point). Such an approach is known in the literature as the landmark methodology (van Houwelingen, 2007). Simulations reported in the paper by de Uña-Álvarez and Meira-Machado (2015) reveal significant improvements on the behaviour of the new method. For small sample sizes and/or large proportion of censored data the landmark approach may result in a wiggly estimator with fewer jump points. This will be more prominent is some transition probabilities. To avoid this problem, we propose two approaches that can be used to reduce the variability of the landmark estimator. A simple approach is based on spline smoothing. Another valid approach is to consider a modification of the landmark estimator based on presmoothing (Dikta, 1998). Simulation studies reported in Section 3 show that the proposed estimators may be much more efficient than the completely nonparametric estimator. In addition, we introduce nonparametric estimators based on the landmark approach that account for the influence of covariates in the transition probabilities.

The organization of the paper is as follows. In Section 2, we introduce the notation and revisit the estimator proposed by de Uña-Álvarez and Meira-Machado (2015). New smoothed estimators are also introduced. The performance of the three sets of estimators is investigated through simulations in Section 3, while in Section 4 the methods are compared through the analysis of medical data from a clinical trial on breast cancer from Germany. In Section 5 we give a brief overview of the R package developed by the authors. Main conclusions are reported in Section 6.

2. Transition probabilities

2.1. Notation and preliminaries

A multi-state model is a stochastic process $(Y(t), t \in \mathscr{T})$ with a finite state space in continuous time. These models are a useful way of describing a process in which an individual moves through a series of states. In this paper, we consider the progressive illness-death model depicted in Figure 1 and we assume that all the subjects are in State 1 at time $t = 0$. Extensions to progressive processes beyond the three-state illness-death model can also be considered following the ideas given in the paper by de Uña-Álvarez and Meira-Machado (2015) (Section 5).

The progressive illness-death model is characterized by the three random variables T_{ij}, $1 \leq i < j \leq 3$, that represent the potential transition times from State i to State j. According to this notation, subjects not visiting State 2 will reach State 3 at time T_{13}. This time will be $T_{12} + T_{23}$ if he/she passes through State 2 before, where the variables T_{12} and T_{23} are recorded successively, rather than simultaneously. In this model we have two competing transitions leaving State 1. Therefore, we denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting State 2 at some time, $Z = \min(T_{12}, T_{13})$ the sojourn time in State 1, and $T = Z + \rho T_{23}$ the total survival time of the process. This means that $\rho = I(Z < T)$.

As usual, assume that these event times are subject to univariate right-censoring denoted by C, which we assume to be independent of (Z,T). Define $\widetilde{Z} = \min(Z,C)$ and $\widetilde{T} = \min(T,C)$ for the censored versions of Z and T. Then, put $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$ for the respective censoring indicators. Finally, the available data is $(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i)$, $1 \leq i \leq n$, iid copies of $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta)$.

In the illness-death model, the target is each of the five transition probabilities $p_{ij}(s,t) = P(Y(t) = j | Y(s) = i)$, where $1 \leq i \leq j \leq 3$ and $s \leq t$ are two pre-specified time points. However, since we have two obvious relations, $p_{12}(s,t) = 1 - p_{11}(s,t) - p_{13}(s,t)$ and $p_{22}(s,t) = 1 - p_{23}(s,t)$, in practice one only need to estimate three of these quantities. According to our notations, the transition probabilities are written as

$$p_{11}(s,t) = P(Z > t \mid Z > s), \quad p_{12}(s,t) = P(Z \leq t, T > t \mid Z > s),$$

$$p_{13}(s,t) = P(T \leq t \mid Z > s), \quad p_{22}(s,t) = P(Z \leq s, T > t \mid Z \leq s, T > s),$$
$$p_{23}(s,t) = P(T \leq t \mid Z \leq s, T > s). \tag{1}$$

2.2. Landmark estimators

According to the landmark approach (van Houwelingen, 2007) nonparametric estimators for the transition probabilities can be introduced by considering specific subsamples or portions of the data. For example, given the time point s, to estimate $p_{1j}(s,t)$ for $j = 1,2,3$ the analysis can be restricted to the individuals observed in State 1 at time s. This set is just $\mathscr{S}_1 = \left\{ i : \widetilde{Z}_i > s \right\}$. As explained in de Uña-Álvarez and Meira-Machado (2015) as long as C is independent of Z, a subject in $\mathscr{S}_1$ is representative of those individuals for which Z exceeds s. On the other hand, for the subpopulation $\widetilde{Z} > s$, the censoring time C is still independent of the pair (Z,T) and, therefore, Kaplan-Meier-based estimation will be consistent. The same applies to the analysis restricted to the individuals observed in State 2 at time s, say $\mathscr{S}_2 = \left\{ i : \widetilde{Z}_i \leq s < \widetilde{T}_i \right\}$, which serves to introduce landmark estimators for $p_{2j}(s,t)$, $j = 2,3$.

The transition probability $p_{11}(s,t)$ is defined as the survival function at time t, among the individuals observed in State 1 at time s, which can be estimated by the ordinary Kaplan-Meier estimator (Kaplan and Meier, 1958) of the sojourn time distribution in State 1, based on the pairs $(\widetilde{Z}_i, \Delta_{1i})$'s in the subsample $\mathscr{S}_1$. Similarly, the transition probability $p_{13}(s,t)$ is defined as one minus the survival function (of the total time) at time t in the same subset $\mathscr{S}_1$. The transition probability $p_{23}(s,t)$ is defined as one minus the survival function (of the total time) at time t in the subset $\mathscr{S}_2$. The landmark estimators given in the paper by de Uña-Álvarez and Meira-Machado (2015) are defined in terms of multivariate 'Kaplan-Meier integrals' with respect to the marginal distribution of the first time, for the transition probability $p_{11}(s,t)$, and with respect to the marginal distribution of the total time T in the remaining transitions.

To formally present the estimators, we need to introduce the expressions for the Kaplan-Meier weights: $w_i^{(s)}$ - the Kaplan-Meier weights attached to $\widetilde{Z}_i$ when estimating the marginal distribution of Z from the $(\widetilde{Z}_i, \Delta_{1i})$'s in subset $\mathscr{S}_1$, $W_i^{(s)}$ - the Kaplan-Meier weights attached to $\widetilde{T}_i$ when estimating the marginal distribution of T from the $(\widetilde{T}_i, \Delta_i)$'s in subset $\mathscr{S}_1$, and $W_i^{[s]}$ - the Kaplan-Meier weights attached to $\widetilde{T}_i$ when estimating the marginal distribution of T from the $(\widetilde{T}_i, \Delta_i)$'s in subset $\mathscr{S}_2$. Let $\left(\widetilde{Z}_{(i)}^{(s)}, \Delta_{1[i]}^{(s)}\right)$, $i = 1,...,n_{1s}$, be the $\left(\widetilde{Z}, \Delta_1\right)$-sample in $\mathscr{S}_1$ ordered with respect to $\widetilde{Z}$, and $\left(\widetilde{T}_{(i)}^{(s)}, \Delta_{[i]}^{(s)}\right)$, $i = 1,...,n_{1s}$, be the $\left(\widetilde{T}, \Delta\right)$-sample in $\mathscr{S}_1$ ordered with respect to $\widetilde{T}$. Then,

$$\widehat{p}_{11}^{ldm}(s,t) = 1 - \sum_{i=1}^{n_{1s}} w_i^{(s)} I(\widetilde{Z}_{(i)}^{(s)} \leq t), \tag{2}$$

$$\widehat{p}_{13}^{ldm}(s,t) = \sum_{i=1}^{n_{1s}} W_i^{(s)} I(\widetilde{T}_{(i)}^{(s)} \leq t), \tag{3}$$

where

$$w_i^{(s)} = \frac{\Delta_{1[i]}^{(s)}}{n_{1s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{1[j]}^{(s)}}{n_{1s}-j+1} \right], \quad 1 \leq i \leq n_{1s};$$

and

$$W_i^{(s)} = \frac{\Delta_{[i]}^{(s)}}{n_{1s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[j]}^{(s)}}{n_{1s}-j+1} \right], \quad 1 \leq i \leq n_{1s}.$$

Similarly, one can introduce the corresponding estimator for $p_{23}(s,t)$. Let $\left(\widetilde{T}_{(i)}^{[s]}, \Delta_{[i]}^{[s]}\right)$, $i = 1, ..., n_{2s}$, is the $\left(\widetilde{T}, \Delta\right)$-sample in $\mathscr{S}_2$ ordered with respect to $\widetilde{T}$. Then,

$$\widehat{p}_{23}^{ldm}(s,t) = \sum_{i=1}^{n_{2s}} W_i^{[s]} I(\widetilde{T}_{(i)}^{[s]} \leq t), \tag{4}$$

where

$$W_i^{[s]} = \frac{\Delta_{[i]}^{[s]}}{n_{2s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[j]}^{[s]}}{n_{2s}-j+1} \right], \quad 1 \leq i \leq n_{2s}.$$

The estimators $\widehat{p}_{ij}^{ldm}(s,t)$ have the simple form of a Kaplan-Meier estimator, based on a certain subsample which is determined by the time point s. Thus, they can also be expressed in the form of inverse of probability censoring weighted average (IPCW) (Satten and Datta, 2001),

$$\widehat{p}_{11}^{ldm}(s,t) = 1 - \frac{1}{n_{1s}} \sum_{i=1}^{n_{1s}} \frac{I(\widetilde{Z}_{(i)}^{(s)} \leq t)\Delta_{1[i]}^{(s)}}{\widehat{G}(\widetilde{Z}_{(i)}^{(s)})},$$

$$\widehat{p}_{13}^{ldm}(s,t) = \frac{1}{n_{1s}} \sum_{i=1}^{n_{1s}} \frac{I(\widetilde{T}_{(i)}^{(s)} \leq t)\Delta_{[i]}^{(s)}}{\widehat{K}_1(\widetilde{T}_{(i)}^{(s)})},$$

and

$$\widehat{p}_{23}^{ldm}(s,t) = \frac{1}{n_{2s}} \sum_{i=1}^{n_{2s}} \frac{I(\widetilde{T}_{(i)}^{[s]} \leq t)\Delta_{[i]}^{[s]}}{\widehat{K}_2(\widetilde{T}_{(i)}^{[s]})},$$

where G, K_1 and K_2 are the survival functions of the censoring variable C, which can be consistently estimated by the Kaplan-Meier approach considering events as 'censored' observations and censored observations as 'events'. Here, $\widehat{G}$ stands for the Kaplan-Meier estimator (of the censoring survival function) based on the $(\widetilde{Z}_i, 1-\Delta_{1i})$'s in subset $\mathscr{S}_1$; whereas, $\widehat{K}_1$ and $\widehat{K}_2$ stand for the Kaplan-Meier estimator (of the censoring survival function) based on the $(\widetilde{T}_i, 1-\Delta_i)$'s in subset $\mathscr{S}_1$ and $\mathscr{S}_2$, respectively.

It is important to mention that $\widehat{p}_{11}^{ldm}(s,t)$ is equivalent to the estimator given by Meira-Machado et al. (2006) and the so-called Aalen-Johansen estimator (Aalen and Johansen, 1978) of $p_{11}(s,t)$, which is consistent regardless of the Markov assumption. In addition, for $s=0$, the landmark estimators are known as the occupation probabilities and they are equivalent to those provided by Meira-Machado et al. (2006).

2.3. Smooth landmark estimators

The standard error of the landmark estimators introduced in the previous subsection may be large when the censoring is heavy, particularly with a small sample size. This problem may be more obvious when estimating the transition probabilities $p_{ij}(s,t)$ for large values of s. In this section, we propose two smoothed versions of the nonparametric landmark estimators given in the previous subsection. One simple approach is based on the use of constrained penalized regression splines (Meyer, 2008, 2012; Wood, 2006). We also introduce a semiparametric estimator which uses a presmoothed version of the Kaplan-Meier estimator (Dikta, 1998; Jácome and Iglesias, 2008; López-de-Ullibarri and Jácome, 2013) pertaining to the distribution of the survival times to weight the data.

2.3.1. Constrained penalized splines

Constrained penalized regression splines can be used as a simple approach which provides smooth estimation of the transition probabilities. These methods can be used under some constraints of shape, such as monotonicity (required for the transition probabilities $p_{11}(s,t)$, $p_{13}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$) and to force a fit curve to go through a particular point. The later constraint is also important since for $s=t$ obvious conditions are required ($p_{11}(s,s)=p_{22}(s,s)=1$ and $p_{12}(s,s)=p_{13}(s,s)=p_{23}(s,s)=0$). To obtain spline-based landmark estimators, $\widehat{p}_{ij}^{crs}(s,t)$, we propose the use of the cubic regression splines.

The key assumption underlying regression spline smoothing is that, for a fixed value of s, the unknown functions $p_{ij}(s,t)$ can be approximated by polynomial splines, defined

on a set of knots (join points) within the domain of $A = [s, \tau_T]$ where τ_T is the upper bound of the support of T. For a fixed value of s, we first obtain the (landmark) estimates of the transition probabilities over all possible time values t with $s < t$, and then, define a cubic spline basis defined by a modest sized set of knots spread evenly through the interval $A = [s, \tau_T]$. For each transition probability $p_{ij}(s,t)$ the use of regression splines provide one approach that allows flexible relationships between a covariate X (time values in A) and the average response (i.e., the landmark estimates of the transition probabilities) as a function of the variable X.

Cubic spline functions are piecewise continuous curves defined by polynomial functions of degree 3. These functions are built joining the piecewise functions on equally spaced join points (also known as knots) so that they are continuous in value, as well as its first two derivatives. This is done by choosing a cubic regression spline basis for which many alternatives can be found (see for example Durrleman and Simon (1989) or Wood (2006)). One approach is to parameterize the spline terms of its values at the knots (Wood, 2006). Specifically, we can write the spline function as a function of $u \in A$

$$f(u) = \sum_{i=1}^{q} \delta_i b_i(u)$$

where q is the number of knots, the b_i are the basis functions of at least second order for representing smooth functions over a given interval, and the δ_i are the spline coefficients. Constraints forcing the curve to pass through a specific point can be imposed. This can be done by creating a regression spline basis, making sure there is a knot at the constraint point. Monotonicity constraints can also be imposed. Penalization is achieved by the conventional integrated square second derivative cubic spline penalty. In practice, we use the gam function in the R package mgcv (Wood, 2006) to obtain the transition probability curves as well for obtaining the predicted values of the smooth curves at the new values. The mgcv implementation of gam, by default uses basis functions for these splines that are designed to be optimal, given the number basis functions used. For details about these methods, see for example Wood (2006) or Pya and Wood (2015).

2.3.2. Presmoothed estimators

The variance of the landmark estimators may also be reduced by presmoothing. Successful applications of presmoothed estimators include estimation of the survival function (Dikta, 1998; Meira-Machado et al., 2016), nonparametric curve estimation (Cao and Jácome, 2004), regression analysis (de Uña-Álvarez and Rodríguez-Campos, 2004; Jácome and Iglesias, 2010), estimation of the bivariate distribution of censored gap times (de Uña-Álvarez and Amorim, 2011), and the estimation of the transition probabilities (Amorim et al., 2011; Moreira et al., 2013). All these references concluded that the presmoothed estimators have improved variance when compared to purely nonparamet-

ric estimators. In this paper, we show that presmoothing is also useful to improve efficiency of the landmark estimators introduced in a previous section. This 'presmoothing' is obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights by a smooth fit. This preliminary smoothing may be based on a certain parametric family such as the logistic, or on a nonparametric estimator of the binary regression curve. When the parametric family is the right one, parametric presmoothing (Dikta, 1998) leads to more efficient estimation than that associated to the unsmoothed estimator. Nonparametric presmoothing (Cao et al., 2005) is useful when there is a clear risk of a miss-specification of the parametric model. The validity of a given parametric model for presmoothing can be checked graphically or formally, by applying a goodness-of-fit test. In this paper we consider estimators obtained using standard logistic regression. The corresponding (semiparametric) presmoothed landmark estimators of the transition probabilities are given by

$$\widehat{p}_{11}^{prs}(s,t) = 1 - \sum_{i=1}^{n_{1s}} pw_i^{(s)} I(\widetilde{Z}_{(i)}^{(s)} \leq t), \tag{5}$$

$$\widehat{p}_{13}^{prs}(s,t) = \sum_{i=1}^{n_{1s}} PW_i^{(s)} I(\widetilde{T}_{(i)}^{(s)} \leq t), \tag{6}$$

and

$$\widehat{p}_{23}^{prs}(s,t) = \sum_{i=1}^{n_{2s}} PW_i^{[s]} I(\widetilde{T}_{(i)}^{[s]} \leq t) \tag{7}$$

where the presmoothed Kaplan-Meier weights are defined as follow:

$$pw_i^{(s)} = \frac{m_{0n}(\widetilde{Z}_{(i)}^{(s)})}{n_{1s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m_{0n}(\widetilde{Z}_{(i)}^{(s)})}{n_{1s}-j+1} \right], \quad 1 \leq i \leq n_{1s},$$

$$PW_i^{(s)} = \frac{m_n(\widetilde{Z}_{[i]}^{(s)}, \widetilde{T}_{(i)}^{(s)})}{n_{1s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m_n(\widetilde{Z}_{[i]}^{(s)}, \widetilde{T}_{(i)}^{(s)})}{n_{1s}-j+1} \right], \quad 1 \leq i \leq n_{1s},$$

and

$$PW_i^{[s]} = \frac{m_n(\widetilde{Z}_{[i]}^{[s]}, \widetilde{T}_{(i)}^{[s]})}{n_{2s}-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m_n(\widetilde{Z}_{[i]}^{[s]}, \widetilde{T}_{(i)}^{[s]})}{n_{2s}-j+1} \right], \quad 1 \leq i \leq n_{2s},$$

where $\left(\widetilde{Z}_{[i]}^{(s)}, \widetilde{T}_{(i)}^{(s)}, \Delta_{[i]}^{(s)}\right)$, $i = 1, ..., n_{1s}$, is the $\left(\widetilde{Z}, \widetilde{T}, \Delta\right)$-sample in $\mathscr{S}_1$ ordered with respect to $\widetilde{T}$, and $\left(\widetilde{Z}_{[i]}^{[s]}, \widetilde{T}_{(i)}^{[s]}, \Delta_{[i]}^{[s]}\right)$, $i = 1, ..., n_{2s}$, is the $\left(\widetilde{Z}, \widetilde{T}, \Delta\right)$-sample in $\mathscr{S}_2$ ordered with respect to $\widetilde{T}$. Here, $m_{0n}(u)$ and $m_n(u,v)$ stand for estimators of the binary regression functions $m_0(u) = P(\Delta_1^{(s)} = 1 \mid \widetilde{Z}^{(s)} = u)$ and $m(u,v) = P(\Delta^{(s)} = 1 \mid \widetilde{Z}^{(s)} = u, \widetilde{T}^{(s)} = v)$, respectively. In this work we assume that these functions belong to a parametric (smooth) family of binary logistic regression curves. For example for $m_{0n}(u)$, we assume that $m_{0n}(u) = m(u;\beta)$ where β is a vector of parameters which typically will be computed by maximizing the conditional likelihood of the $\Delta_1^{(s)}$'s given $\widetilde{Z}^{(s)}$.

As discussed in Amorim et al. (2011) the function $m(u,v)$ will typically be discontinuous along the line $v = u$, that is, for those covariate values $(\widetilde{Z}, \widetilde{T})$ corresponding to individuals who are censored while being in state 1 or who suffer a direct transition to the absorbing state. In order to construct $m_n(u,v)$ we use the ideas proposed by Amorim et al. (2011).

Note that, unlike the unsmoothed landmark estimators, the presmoothed estimators can attach positive mass to pair of event times with censored total time. The presmoothed estimators $p_{ij}^{prs}(s,t)$ are step functions, with jumps at the observed (censored or uncensored) times. In this aspect they differ from landmark estimators ($p_{ij}^{ldm}(s,t)$) whose jumps are restricted to the uncensored times. In the limit case of no presmoothing, the Presmoothed Landmark estimator reduces to the landmark estimator.

In practice, estimation of the variance is needed for inference purposes. To this end, resampling techniques such as the bootstrap can be used. These methods can be used to construct confidence limits based on the bootstrap (e.g., using the basic or the percentile method) and thus to confirm if the proposed methods lead to a reduction in the variability of the estimators proposed in this section. These resampling techniques can be easily implemented using the R package described in Section 5.

Simulations reported in Section 3 reveal that the proposed estimators are virtually unbiased and that they may achieve good efficiency levels when compared to the unsmoothed landmark estimators.

2.4. Including covariates

In this section, we will explain how to introduce covariate information in the unsmoothed landmark estimators, $\widehat{p}_{ij}^{ldm}(s,t)$. In particular, we are interested in estimating the conditional transition probabilities $p_{ij}(s,t|X = x)$ that can be computed for any times s and t, $s < t$, but conditional to a given continuous covariate X that could either be a baseline covariate or a current covariate that is observed for an individual before the individual makes a particular transition of interest. Discrete covariates can be also included by splitting the sample for each level of the covariate and repeating the described procedures for each subsample.

To account for the covariate effect, one standard method is to consider estimators based on a Cox's model (Cox, 1972), with the corresponding baseline hazard function estimated by the Breslow's method (Breslow, 1972). Flexible effects of the covariates on the transition probabilities as those depicted in Figure 5 can be obtained using an alternative approach which introduces local smoothing by means of kernel weights based on local constant (Nadaraya-Watson) regression (Nadaraya 1965; Watson 1964).

Nonparametric estimators of the conditional transition probabilities have been recently proposed by Meira-Machado et al. (2015). These authors propose to estimate $p_{ij}(s,t \mid X=x)$ via estimation of the conditional expectations such as $E[\psi(Z,T) \mid X=x]$, where ψ is a general function defined over Z and T. Following the ideas described in Meira-Machado et al. (2015), the conditional transition probabilities are defined as follows:

$$\begin{aligned} p_{11}(s,t \mid X=x) &= \frac{1-P(Z \leq t \mid X=x)}{1-P(Z \leq s \mid X=x)}, \\ p_{13}(s,t \mid X=x) &= \frac{P(Z>s, T \leq t \mid X=x)}{1-P(Z \leq s \mid X=x)} \\ p_{23}(s,t \mid X=x) &= \frac{P(Z \leq s, s<T \leq t \mid X=x)}{P(Z \leq s \mid X=x)-P(T \leq s \mid X=x)}. \end{aligned} \tag{8}$$

The conditional transition probability $p_{11}(s,t \mid X=x)$, the denominator of $p_{13}(s,t \mid X=x)$ and the denominator of $p_{23}(s,t \mid X=x)$ involve the estimation of the conditional distribution/survival function of the response, given the covariate under random right censoring. This topic was introduced by Beran (1981) and was further studied by several authors (see e.g. papers by Akritas, 1994; van Keilegom et al., 2001; Akritas and van Keilegom, 2003). Their proposals can be used to estimate for instances the conditional distribution function of $Z \mid X=x$, that is, $F_{X=x}(u)=P(Z \leq u \mid X=x)$ which we denote by $\widehat{F}_{X=x}$ or simply by $\widehat{F}_x$. This can be done using the estimator introduced by Beran (1981),

$$\widehat{F}_{X=x}(u) = 1 - \prod_{\widetilde{Z}_i \leq u, \Delta_{1i}=1} \left[1 - \frac{NW(x,X_i,h)}{\sum_{j=1}^{n} I(\widetilde{Z}_j \geq \widetilde{Z}_i) NW(x,X_j,h)} \right], \tag{9}$$

where $NW(x,X_i,h)$ are the Nadaraya-Watson (NW) weights (Nadaraya, 1965; Watson, 1964)

$$NW(x,X_i,h) = \frac{D((x-X_i)/h)}{\sum_{j=1}^{n} D((x-X_j)/h)}$$

where D is a known probability density function (the kernel function) and h is a bandwidth.

The remaining quantities in the computation of the conditional transition probabilities involve conditional expectations of particular transformations of the pair (Z,T) given X, $E[\psi(Z,T) \mid X = x]$ which can not be estimated so simply.

In the absence of censoring, to estimate the conditional expectations $E[\psi(Z,T) \mid X = x]$ we may use kernel smoothing techniques by calculating a local average of the $\psi(Z,T)$, that is, as follows:

$$\widehat{E}[\psi(Z,T) \mid X = x] = \sum_{i=1}^{n} NW(x,X_i,h)\psi(\widetilde{Z}_i,\widetilde{T}_i),$$

where $NW(x,X_i,h)$ are the Nadaraya-Watson (NW) weights.

To handle right censoring Meira-Machado et al. (2015) propose the use of inverse of probability censoring weighting. Assuming that $\forall x$, $P(\widetilde{Z} > s, \widetilde{T} \leq t \mid X = x) > 0$ and $P(\widetilde{Z} \leq s, s < \widetilde{T} \leq t \mid X = x) > 0$, we have the following:

$$E[I(Z > s, T \leq t) \mid X = x] = E[I(\widetilde{Z} > s, s < \widetilde{T} \leq t)\Delta / K_X(\widetilde{T}) \mid X = x],$$

$$E[I(Z \leq s, T > s) \mid X = x] = E[I(\widetilde{Z} \leq s, \widetilde{T} \leq t)\Delta / K_X(\widetilde{T}) \mid X = x]$$

where K_X denotes the conditional survival function of the censoring variable C given the covariate X, that is $K_{X=x}(u) = P(C > u \mid X = x)$. Let $\widehat{K}_{X=x}$ denote Beran's estimator of K_X. Based on this, the following nonparametric estimators of the conditional transition probabilities can be introduced:

$$\widehat{p}_{11}(s,t \mid X = x) = \frac{1 - \widehat{F}_x(t)}{1 - \widehat{F}_x(s)}, \tag{10}$$

$$\widehat{p}_{13}(s,t \mid X = x) = \frac{1}{1 - \widehat{F}_x(s)} \sum_{i=1}^{n} \frac{NW(x,X_i,h_1) I(\widetilde{Z}_i > s, \widetilde{T}_i \leq t)\Delta_i}{\widehat{K}_{X_i}(\widetilde{T}_i)}, \tag{11}$$

and

$$\widehat{p}_{23}(s,t \mid X = x) = \frac{\sum_{i=1}^{n} NW(x,X_i,h_1) I(\widetilde{Z}_i \leq s, s < \widetilde{T}_i \leq t)\Delta_i / \widehat{K}_{X_i}(\widetilde{T}_i)}{\widehat{F}_x(s) - \widehat{H}_x(s)}, \tag{12}$$

where $\widehat{H}_x$ denote Beran's estimator of the conditional distribution of $T \mid X = x$.

Similar ideas as those explained above can be used to introduce nonparametric estimators for the conditional transition probabilities based on landmark. For example, given the time point s, the estimation of the conditional transition probabilities $p_{11}(s,t \mid X = x)$ and $p_{13}(s,t \mid X = x)$ are restricted to the individuals in State 1 at time s. Thus, the landmark estimators for these quantities are given as follows:

$$\widetilde{p}_{11}(s,t \mid X=x) = 1 - \sum_{i=1}^{n_{1s}} NW(x, X_{(i)}^{(s)}, h_1) \frac{I(\widetilde{Z}_{(i)}^{(s)} \leq t)\Delta_{1[i]}^{(s)}}{\widehat{G}_{X_i}(\widetilde{Z}_{(i)}^{(s)})}, \tag{13}$$

and

$$\widetilde{p}_{13}(s,t \mid X=x) = \sum_{i=1}^{n_{1s}} NW(x, X_{(i)}^{(s)}, h_1) \frac{I(\widetilde{T}_{(i)}^{(s)} \leq t)\Delta_{[i]}^{(s)}}{\widehat{K}_{1,X_i}(\widetilde{T}_{(i)}^{(s)})}, \tag{14}$$

where $\widehat{G}_X$ and $\widehat{K}_{1,X}$ are Beran's estimators for the conditional survival function of the censoring variable of the sojourn time in State 1 (respectively, total time) given X in subset $\mathscr{S}_1$.

Similarly, the built of the landmark estimator of the conditional transition probability $p_{23}(s,t \mid X=x)$ is restricted to the individuals in State 2 at time s:

$$\widetilde{p}_{23}(s,t \mid X=x) = \sum_{i=1}^{n_{2s}} NW(x, X_{(i)}^{[s]}, h_2) \frac{I(\widetilde{T}_{(i)}^{[s]} \leq t)\Delta_{[i]}^{[s]}}{\widehat{K}_{2,X_i}(\widetilde{T}_{(i)}^{[s]})}, \tag{15}$$

where $\widehat{K}_{2,X}$ is Beran's estimator of the conditional survival function of the censoring variable of the total time given X in subset $\mathscr{S}_2$.

Simulation results (not reported here) reveal that the landmark based estimators $\widetilde{p}_{12}(s,t \mid X=x)$ and $\widetilde{p}_{13}(s,t \mid X=x)$ perform favourably when compared to $\widehat{p}_{12}(s,t \mid X=x)$ and $\widehat{p}_{13}(s,t \mid X=x)$, respectively. In contrast, the landmark estimator $\widetilde{p}_{23}(s,t \mid X=x)$ have a worst performance when compared to $\widehat{p}_{23}(s,t \mid X=x)$ particularly when computed at time points s for which few individuals are observed in State 2.

3. Simulation study

In this section, we report the results of a simulation study carried out to investigate the empirical behaviour of the estimators, introduced in Section 2, for finite sample sizes. More specifically, the landmark unsmoothed estimators, $\widetilde{p}_{ij}^{ldm}(s,t)$, with the smoothed estimators, $\widetilde{p}_{ij}^{crs}(s,t)$, based on cubic regression splines and the semiparametric presmoothed estimators, $\widetilde{p}_{ij}^{prs}(s,t)$.

To simulate the data in the illness-death model, we use the same scenario as that described in Amorim et al. (2011) and de Uña-Álvarez and Meira-Machado (2015). We separately consider the subjects passing through State 2 at some time, and those who directly go to the absorbing State 3. For the first subgroup of individuals ($\rho = 1$), the successive gap times $(Z, T-Z)$ are simulated according to the bivariate exponential distribution

$$F_{12}(u,v) = F_1(u)F_2(v)\left[1+\gamma\left\{1-F_1(u)\right\}\left\{1-F_2(v)\right\}\right]$$

with exponential marginal distribution functions with rate parameter 1. The single parameter γ controls the amount of dependency between the gap times. The parameter γ was set to 0 for simulating independent gap times, and also to 1, corresponding to 0.25 correlation between Z and $T-Z$. The simulation procedure is as follows:

Step 1. Draw $\rho \sim Ber(p)$ where p is the proportion of subjects passing through State 2.

Step 2. If $\rho = 1$ then:

(2.1) $V_1 \sim U(0,1), V_2 \sim U(0,1)$ are independently generated;

(2.2) $U_1 = V_1, A = \gamma(2U_1-1)-1, B = (1-\gamma(2U_1-1))^2 + 4\gamma V_2(2U_1-1)$

(2.3) $U_2 = 2V_2/\left(\sqrt{B}-A\right)$

(2.4) $Z = \ln(1/(1-U_1)), T = \ln(1/(1-U_2)) + Z$

If $\rho = 0$ then $Z = \ln(1/(1-U(0,1)))$.

Situations with $p = 1$ corresponds to the three-state progressive model, in which a direct transition to State 3 is not allowed. In our simulation we consider $p = 0.7$. An independent uniform censoring time C is generated, according to models $U[0,4]$ and $U[0,3]$. The first model results in 24% of censoring on the first gap time Z, and in 47% of censoring on the second gap time $T-Z$, for those individuals with $\rho = 1$. The second model increases these censoring levels to 32% and about 57%, respectively.

For each simulated setting we derived the analytic expression of $p_{ij}(s,t)$ for six different points (s,t) $(s<t)$, corresponding to combinations of the percentiles 20%, 40%, 60% and 80% of the marginal distributions of the gap times. Sample sizes of 100, 250 and 500 were considered. In each simulation, 1000 samples were generated and for each of the three estimators we obtain the mean bias, the standard deviation (SD), and the mean square error (MSE) based on the 1000 Monte Carlo replicates. Table 1 reports the results for the transition probabilities $p_{12}(s,t)$ and $p_{23}(s,t)$ for the case with dependent gap times; the results for independent gap times (not shown) are similar.

As would be expected, results reported in Table 1 reveal that the performance of all methods is poorer at the right tail (i.e., larger values of s and t) where the censoring effects are stronger. At these points the SD is in most cases higher. The SD decreases with an increase in the sample size and with a decrease of the censoring percentage. All methods proposed in this article obtain in all settings a small bias.

Results reported in Table 1 reveal that the SD clearly dominates the performance of the proposed estimators in most cases. This is particularly clear when comparing the semiparametric estimators with the unsmoothed landmark estimators. The semiparamet-

Table 1: *Bias and standard deviation (SD) for the three estimators of $p_{ij}(s,t)$. The MSE of $\widehat{p}_{ij}^{prs}(s,t)$ and $\widehat{p}_{ij}^{crs}(s,t)$ relative to $\widehat{p}_{ij}^{ldm}(s,t)$ are also given. Scenario of correlated exponential gap times with three sample sizes and two censoring levels.*

		$\widehat{p}_{12}^{ldm}(s,t)$		$\widehat{p}_{12}^{prs}(s,t)$		$\widehat{p}_{12}^{crs}(s,t)$			
		bias	SD	bias	SD	bias	SD	MSE^{ldm}/MSE^{prs}	MSE^{ldm}/MSE^{crs}
$(s,t)=$	(.2231,.5108)								
$n=100$	$C\sim U[0,4]$	−0.0017	0.0408	−0.0013	0.0368	0.0046	0.0377	1.2086	1.1532
	$C\sim U[0,3]$	−0.0017	0.0415	−0.0011	0.0375	0.0026	0.0385	1.2279	1.1586
$n=250$	$C\sim U[0,4]$	0.0006	0.0263	0.0015	0.0235	0.0040	0.0249	1.2468	1.0882
	$C\sim U[0,3]$	0.0006	0.0268	0.0017	0.0240	0.0022	0.0250	1.2388	1.1410
$n=500$	$C\sim U[0,4]$	0.0017	0.0188	0.0024	0.0170	0.0039	0.0170	1.2075	1.1718
	$C\sim U[0,3]$	0.0016	0.0189	0.0027	0.0169	0.0022	0.0171	1.2325	1.2100
$(s,t)=$	(.2231,.9163)								
$n=100$	$C\sim U[0,4]$	−0.0028	0.0518	−0.0025	0.0473	0.0017	0.0497	1.2034	1.0881
	$C\sim U[0,3]$	−0.0030	0.0530	−0.0035	0.0487	0.0003	0.0525	1.1845	1.0221
$n=250$	$C\sim U[0,4]$	<0.0001	0.0338	0.0027	0.0311	0.0006	0.0320	1.1855	1.1153
	$C\sim U[0,3]$	−0.0001	0.0352	−0.0007	0.0320	0.0007	0.0328	1.2078	1.1512
$n=500$	$C\sim U[0,4]$	0.0010	0.0235	0.0008	0.0215	0.0002	0.0221	1.1970	1.1326
	$C\sim U[0,3]$	0.0007	0.0243	−0.0003	0.0219	0.0017	0.0239	1.2333	1.0295
$(s,t)=$	(.5108,1.6094)								
$n=100$	$C\sim U[0,4]$	0.0051	0.0704	0.0025	0.0642	0.0024	0.0691	1.2084	1.0422
	$C\sim U[0,3]$	0.0047	0.0780	−0.0004	0.0695	0.0029	0.0774	1.2647	1.0175
$n=250$	$C\sim U[0,4]$	0.0022	0.0438	−0.0007	0.0397	0.0011	0.0444	1.2176	0.9763
	$C\sim U[0,3]$	0.0019	0.0489	−0.0028	0.0435	0.0007	0.0487	1.2607	1.0097
$n=500$	$C\sim U[0,4]$	0.0005	0.0301	−0.0015	0.0273	0.0015	0.0301	1.2112	0.9979
	$C\sim U[0,3]$	0.0008	0.0337	−0.0036	0.0296	0.0015	0.0333	1.2745	1.0227
$(s,t)=$	(.9163,1.6094)								
$n=100$	$C\sim U[0,4]$	0.0055	0.0848	0.0022	0.0775	0.0007	0.0883	1.2015	0.9261
	$C\sim U[0,3]$	0.0053	0.0956	0.0008	0.0873	−0.0070	0.0978	1.2019	0.9537
$n=250$	$C\sim U[0,4]$	0.0029	0.0547	−0.0005	0.0492	0.0027	0.0543	1.2362	1.0152
	$C\sim U[0,3]$	0.0026	0.0610	−0.0020	0.0539	0.0021	0.0589	1.2820	1.0733
$n=500$	$C\sim U[0,4]$	<0.0001	0.0383	−0.0032	0.0346	0.0022	0.0367	1.2123	1.0862
	$C\sim U[0,3]$	0.0008	0.0417	−0.0040	0.0371	0.0014	0.0408	1.2491	1.0440
		$\widehat{p}_{23}^{ldm}(s,t)$		$\widehat{p}_{23}^{prs}(s,t)$		$\widehat{p}_{23}^{crs}(s,t)$			
$(s,t)=$	(.2231,.5108)								
$n=100$	$C\sim U[0,4]$	0.0016	0.1687	−0.0028	0.1693	0.0042	0.1475	0.9937	1.3054
	$C\sim U[0,3]$	0.0024	0.1722	−0.0006	0.1706	0.0040	0.1535	1.0203	1.2570
$n=250$	$C\sim U[0,4]$	−0.0068	0.0967	−0.0087	0.0946	0.0028	0.0965	1.0422	1.0088
	$C\sim U[0,3]$	−0.0068	0.0971	−0.0063	0.0957	0.0020	0.0974	1.0294	0.9983
$n=500$	$C\sim U[0,4]$	−0.0017	0.0692	−0.0014	0.0677	0.0033	0.0661	1.0463	1.0932
	$C\sim U[0,3]$	−0.0015	0.0704	<0.0001	0.0685	0.0025	0.0658	1.0559	1.1409
$(s,t)=$	(.2231,.9163)								
$n=100$	$C\sim U[0,4]$	0.0015	0.1615	−0.0003	0.1566	<0.0001	0.1456	1.0633	1.2259
	$C\sim U[0,3]$	0.0009	0.1671	−0.0005	0.1579	−0.0034	0.1550	1.1205	1.1612
$n=250$	$C\sim U[0,4]$	0.0021	0.0939	0.0018	0.0921	0.0003	0.0910	1.0405	1.0653
	$C\sim U[0,3]$	0.0024	0.0972	0.0008	0.0943	0.0023	0.0962	1.0632	1.0209
$n=500$	$C\sim U[0,4]$	0.0031	0.0657	0.0021	0.0647	<0.0001	0.0626	1.0323	1.1039
	$C\sim U[0,3]$	0.0033	0.0691	0.0021	0.0666	0.0001	0.0651	1.0794	1.1282
$(s,t)=$	(.5108,1.6094)								
$n=100$	$C\sim U[0,4]$	−0.0006	0.1247	−0.0053	0.1169	0.0019	0.1229	1.1352	1.0293
	$C\sim U[0,3]$	−0.0058	0.1358	−0.0137	0.1268	0.0020	0.1329	1.1344	1.0456
$n=250$	$C\sim U[0,4]$	0.0005	0.0768	−0.0013	0.0731	0.0012	0.0738	1.1035	1.0828
	$C\sim U[0,3]$	0.0020	0.0835	−0.0068	0.0779	0.0012	0.0807	1.1388	1.0702
$n=500$	$C\sim U[0,4]$	0.0019	0.0540	0.0006	0.0517	0.0007	0.0522	1.0902	1.0711
	$C\sim U[0,3]$	0.0006	0.0604	−0.0038	0.0563	0.0001	0.0573	1.1445	1.1110
$(s,t)=$	(.9163,1.6094)								
$n=100$	$C\sim U[0,4]$	−0.0085	0.1391	−0.0111	0.1335	0.0023	0.1388	1.0816	1.0078
	$C\sim U[0,3]$	−0.0086	0.1525	−0.0122	0.1422	0.0070	0.1460	1.1448	1.0920
$n=250$	$C\sim U[0,4]$	−0.0031	0.0870	−0.0040	0.0828	0.0024	0.0836	1.1041	1.0841
	$C\sim U[0,3]$	−0.0042	0.0979	−0.0057	0.0904	0.0019	0.0922	1.1690	1.1290
$n=500$	$C\sim U[0,4]$	−0.0009	0.0593	−0.0004	0.0564	−0.0006	0.0590	1.1056	1.0121
	$C\sim U[0,3]$	−0.0022	0.0665	−0.0034	0.0616	−0.0012	0.0647	1.1635	1.0572

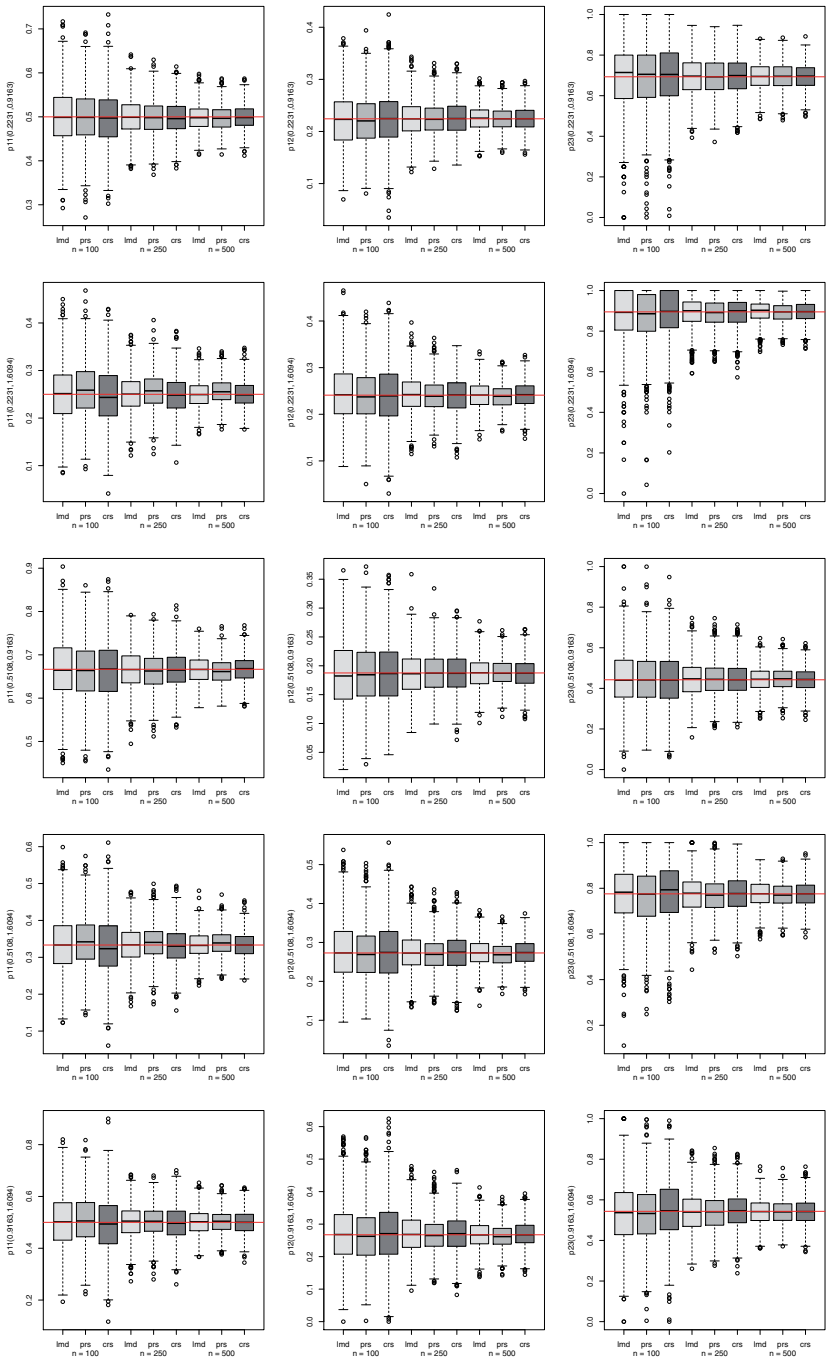


Figure 2: *Boxplots of the M = 1000 estimates of the transition probabilities of the three estimators, with three different sample sizes and correlated exponential gap times. Censoring times were generated from an uniform distribution on* $[0,3]$. *Horizontal solid red line corresponds to the true value of the transition probability.*

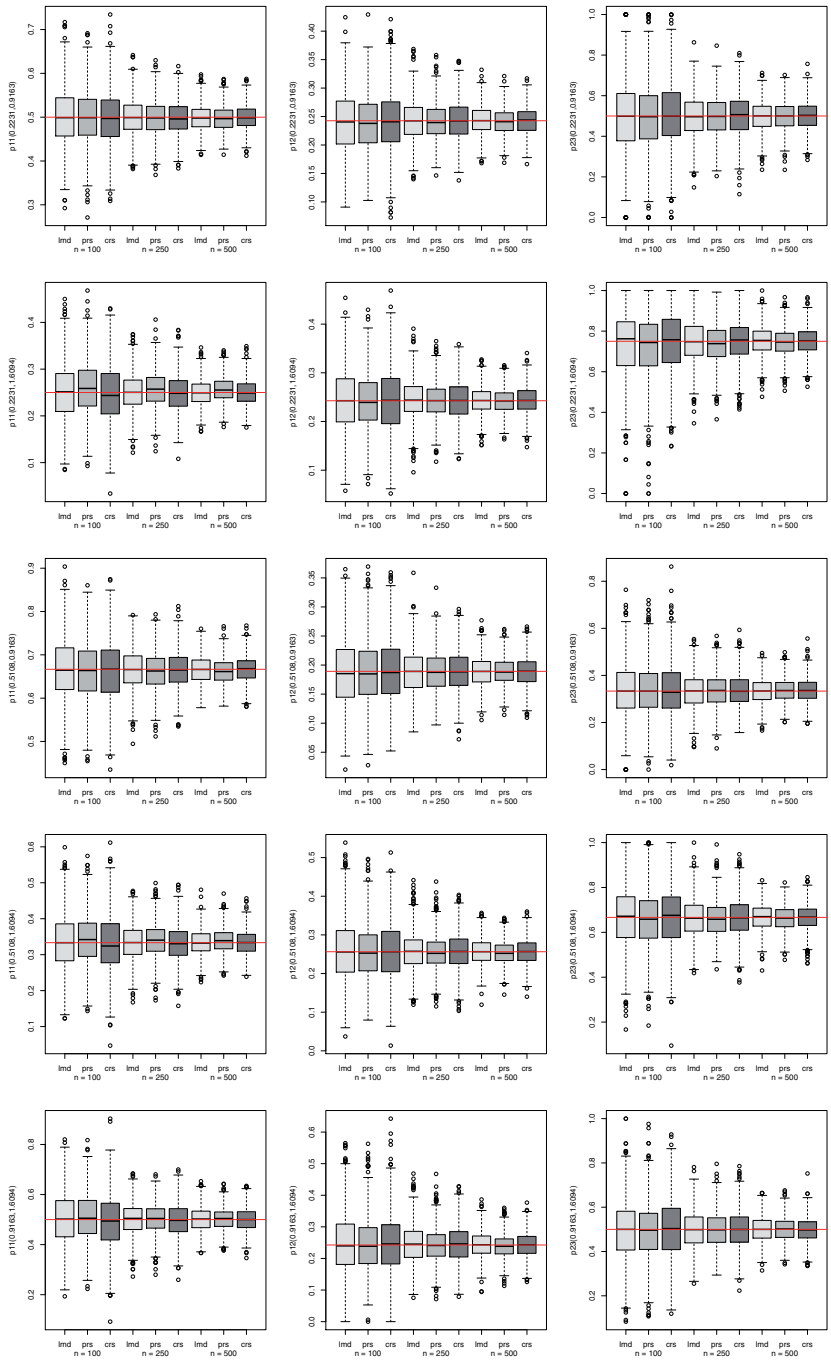


Figure 3: *Boxplots of the $M = 1000$ estimates of the transition probabilities of the three estimators, with three different sample sizes and uncorrelated exponential gap times. Censoring times were generated from an uniform distribution on $[0,3]$. Horizontal solid red line corresponds to the true value of the transition probability.*

ric estimator achieve better results with less SD and less MSE. This can be seen by the relative efficiency between the semiparametric estimator and the unsmoothed landmark estimator that was measured by the ratio between their corresponding MSEs. The semiparametric estimators reported a smaller MSE in most cases. It can also be observed that the relative advantages of the semiparametric estimator is greater for higher censoring percentages. This advantage is also apparently greater when estimating the transition probability $p_{12}(s,t)$.

For completeness purposes we show in Figures 2 and 3 the boxplots of the estimates of the transition probabilities based on the 1000 Monte Carlo replicates for the three estimators, with different sample sizes, for correlated exponential gap times (Figure 2) and independent gap times (Figure 3). Plots shown in these figures were obtained for the higher censoring levels ($C \sim U[0,3]$). In addition to the transition probabilities reported in Table 1 this figure also show the performance of the transition probability $p_{11}(s,t)$ for all methods. The boxplots shown in Figures 2 and 3 reveal some results which are in agree with our findings reported in Table 1. From these plots it can be seen that all methods have small bias and confirm the less variability of the semiparametric estimators.

Despite of offering a small bias, the bias associated to presmoothed estimators is in general larger than that of the unsmoothed landmark estimator. This bias component may be larger when there is some miss-specification in the chosen parametric model. Importantly, the validity of a given model for presmoothing can be checked graphically or formally, by applying a goodness-of-fit tests (e.g. Hosmer and Lemeshow (1989)). This implies that the risk of introducing a large bias through a miss-specified model can be controlled in practice.

4. German breast cancer study

Breast cancer is one of the most commonly diagnosed cancers in women. Prognosis of this carcinoma is related to a large variety of clinical and pathological factors such as age, tumor size, histological grade, lymph node involvement, and hormone receptor status. Another significant prognostic factor for these patients in overall survival is the presence of a recurrence. Traditionally, the effect of these time-dependent covariates is studied using extensions of the Cox proportional hazards model (Cox, 1972; Genser and Wernecke, 2005). The analysis of such studies can also be successfully performed using a multi-state model (Pérez-Ocón et al., 2001; Putter et al., 2007; Meira-Machado et al., 2009).

Several studies have been developed over the last decades regarding breast cancer. Between 1983 and 1989, four clinical trials were conducted by the German Breast Cancer Study Group (GBSG) including 2746 patients with primary node positive breast cancer. Details about these studies can be found in the paper by Schumacher et al. (1994).

Among other papers, these data were used by Schmoor et al. (2000) and Meier-Hirmer and Schumacher (2013). In both cases the main goal was to evaluate the effect on future prognosis of an isolated locoregional recurrence (ILRR). While Schmoor et al. (2000) used a Cox proportional hazards model, Meier-Hirmer and Schumacher (2013) used an illness-death model to investigate the influence of the time-dependent covariate 'recurrence'. Both studies conclude, among other things, that the increased risk after ILRR decreased significantly with increasing time since ILRR. In this paper we use data from the second trial in which a total of 720 women with primary node positive breast cancer is recruited in the period between July 1984 and December 1989. The data is available at the University of Massachusetts website for statistical software information as well as part of the R packages mfp, TH.data and flexsurv. The data which was also used by Sauerbrei and Royston (1999) considers 686 patients who had complete data for the two event times (time to recurrence and time to death). In this study, patients were followed from the date of breast cancer diagnosis until censoring or dying from breast cancer. From the total of 686 women, 299 developed a recurrence and 171 died. Besides the two event times and the corresponding indicator statuses a vector of covariates including age at acceptance tumor size, number of positive lymph nodes, progesterone and estrogen receptor status, menopausal status and tumor grade are also available. The covariate 'recurrence' is the only time-dependent covariate, while the other covariates included are fixed. This covariate can be considered as an intermediate transient state and modeled using an illness-death model with states 'Alive and disease-free', 'Alive with Recurrence' and 'Dead'. In this section, we present plots for the three different methods to estimate the transition probabilities described in Section 2. Figure 4 reports estimated transition probabilities for $p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{23}(s,t)$, for fixed values $s = 365$ and $s = 730$ (days), along time t (corresponding to 1 and 2 years after surgery). Plots shown in these figure also show the pointwise bootstrap confidence bands of the unsmoothed method. Estimators for all three methods shown in these plots report roughly the same estimates. Minor differences are appreciated when comparing the nonparametric unsmoothed method with their counterparts (the semiparametric presmoothed approach and the method based on cubic regression splines) which is in agree with our findings in the simulation study.

Plots shown at the top of Figure 4 provide the probabilities of being alive and without recurrence for the individuals who are disease free 1 year (Figure 4, top left) and 2 years (Figure 4, top right) after surgery (i.e. $p_{11}(s,t)$). These are monotonous decreasing curves. The curve do not decrease to zero due to a (disease free) censoring rate of about 56.4% (387 woman remain alive and disease free until the end of study). In addition, one can observe that these probabilities increase with an increase of the value of s. Similar conclusions can be obtained from the plots shown at the bottom of Figure 4, in which the transition probability $p_{23}(s,t)$ is estimated through the three methods. These plots report one minus the survival fraction along time, among the individuals in the recurrence state 1 year (Figure 4, bottom left) and 2 years (Figure 4, bottom right) after surgery. It can be observed from these plots that the survivorship is smaller for the first

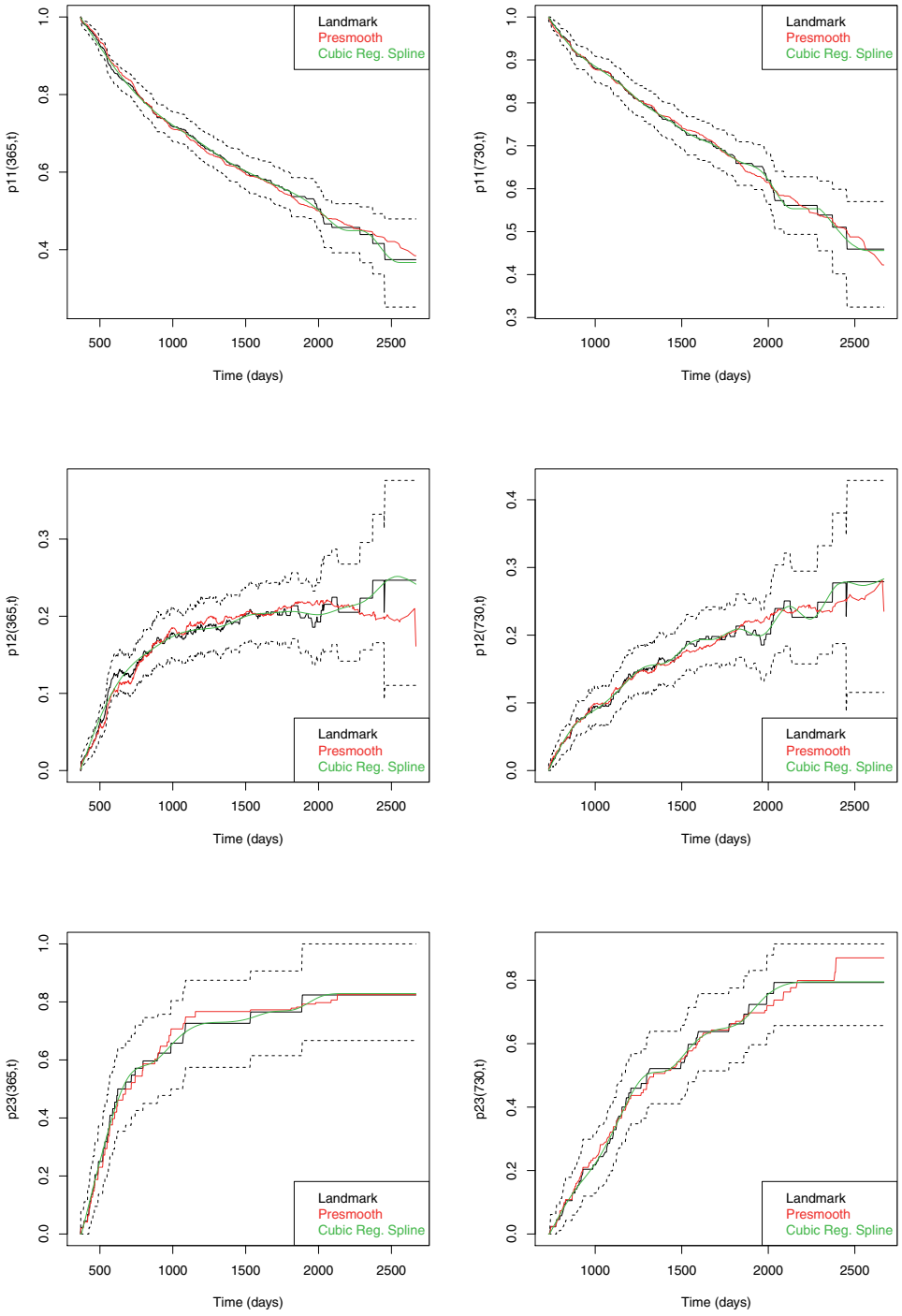


Figure 4: *Estimates of the transition probabilities $p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{23}(s,t)$ for $s = 365$ (left) and $s = 730$ (right) using the three methods (landmark, presmoothing and cubic regression splines). Pointwise confidence intervals of the landmark method is also shown. Breast cancer data.*

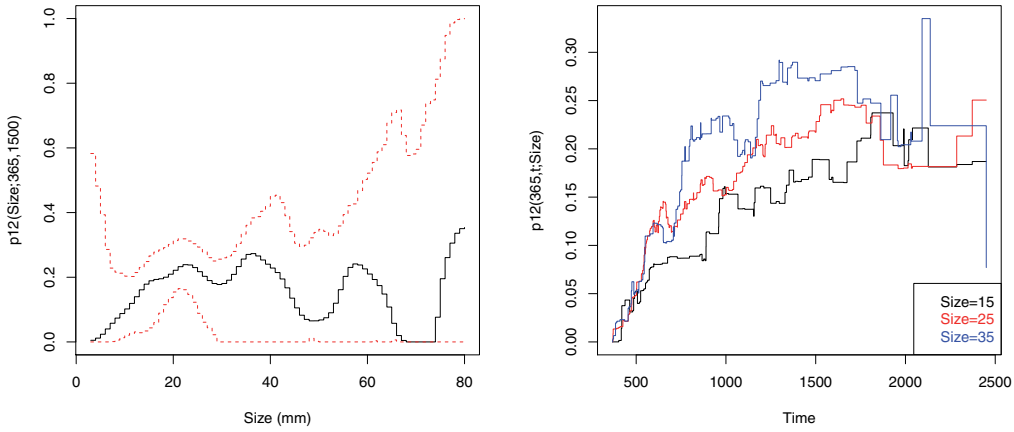


Figure 5: *Landmark estimates of the conditional transition probabilities* $p_{12}(s=365,t=1500 \mid size)$ *(left) and* $p_{12}(365,t \mid size)$ *(right). Breast cancer data.*

group for lower values of t, suggesting a negative impact of an earlier recurrence time. In contrast to these plots, curves for the transition probability $p_{12}(s,t)$ are not necessarily monotonous. Plots shown for this transition probability allows for an inspection along time of the probability of being alive with recurrence for the individuals who are disease free at 1 and 2 years after surgery. Since the recurrence state is transient, in general, this curve is first increasing and then decreasing. However, in this case, for $s = 365$, the curve has a rapid increase at lower times and afterwards remain roughly constant. The increase is more gradual for $s = 730$. The reason why the curve does not decreases can be explained by the percentage of about 46% of woman that remain in the recurrence state at the end of study. Departures between estimated curves can be more appreciated for larger time values where the censoring effects are stronger.

Figure 5 depict the landmark estimates of the conditional transition probability on the recurrence transition. Plot shown at the left depicts the estimates of the transition probability $p_{12}(\text{size};365,1500)$ as functions of the covariate tumor size together with a 95% pointwise confidence bands based on simple bootstrap which resamples each datum with probability $1/n$. Plot at the right depicts the estimates of the transition probability $p_{12}(365,t)$ conditional on the covariate tumor size. The effects of tumor size according to three groups depicted in these plots, which are purely nonparametric, indicate the real influence of this covariate in the recurrence transition. Both plots are in agreement and indicate that patients with higher tumor sizes have a larger probability of recurrence. To compute the conditional transition probabilities shown in this figure we have used a common bandwidth selector and Gaussian kernels. To this end we have used the dpik function which is available from the R KernSmooth package.

5. Software development

To provide the biomedical researchers with an easy-to-use tool for obtaining estimates of the transition probabilities we develop an R package called tprob. This package can be used to implement all nonparametric and semiparametric estimators for the transition probabilities discussed in Section 2. In addition, estimators are also implemented that account for the influence of covariates. Bootstrap confidence bands are provided for all methods. This package is composed by several functions that allow users to obtain estimates and plots of the transition probabilities. Details on the usage of these functions can be obtained with the corresponding help pages. The CPU time needed for running some of the proposed methodologies varies according to whether bootstrap confidence bands are requested or not, the sample size, and the type of processor in the computer. To minimize these problems the most computationally demanding parts of the code were developed and implemented in the C programming language. This software is available at the author web site `http://w3.math.uminho.pt/~lmachado/R/tprob`.

6. Discussion

There have been several recent contributions for the estimation of the transition probabilities in the context of non-Markov multi-state models. Meira-Machado et al. (2006) introduced a substitute for the Aalen-Johansen estimator in the case of a non-Markov illness-death model. They showed that the new estimator may behave much more efficiently than the Aalen-Johansen when the Markov assumption does not hold. However, the proposal of Meira-Machado et al. (2006) has the drawback of requiring that the support of the censoring distribution contains the support of the lifetime distribution, otherwise they only report valid estimators for truncated transition probabilities. Recently, the problem of estimating the transition probabilities in a non-Markov illness-death model has been reviewed, and new estimators have been proposed which are consistent regardless the Markov condition and the referred assumption on the censoring support. These estimators are built by considering specific subsets of individuals (namely, those observed to be in a given state at a prespecified time point s for which the ordinary Kaplan-Meier survival function leads to a consistent estimator of the target. As a weakness, it provides large standard errors for large values of s and higher censoring percentages.

In this article we propose two approaches that can be used to reduce the variability of the proposed estimator. A simple approach is based on spline smoothing (cubic regression splines). Another valid approach is to consider a semiparametric estimator based on a presmoothed version of the Kaplan-Meier estimator. The provided simulations suggest that both approaches are preferable to the original nonparametric estimator, since they often have less variance while providing more reliable curves. Between the two new methods, the one based on presmoothing is recommended.

Acknowledgements

This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade-COMPETE" and by Portuguese Funds through FCT-"Fundação para a Ciência e a Tecnologia", in the form of grant UID/MAT/00013/2013. We thank the two anonymous referees for comments and suggestions which have improved the presentation of the paper.

References

Aalen, O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5, 141–150.

Akritas, M. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, 22, 1299–1327.

Akritas, M.G. and van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data, *Journal of Royal Statistical Society, B*, 65, 457–471.

Allignol, A., Beyersmann, J., Gerds, T. and Latouche, A. (2014). A competing risks approach for nonparametric estimation of transition probabilities in an non-Markov illness-death model. *Lifetime Data Analysis*, 20, 495–513.

Amorim, A., de Uña-Álvarez, J. and Meira-Machado, L. (2011). Presmoothing the transition probabilities in the illness-death model. *Statistics & Probability Letters*, 81, 797–806.

Andersen, P.K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods Medical Research*, 11, 91–115.

Beran, R. (1981). *Nonparametric Regression with Randomly Censored Survival Data*. Technical report, Univ. California, Berkeley.

Breslow, N.E. (1972). Discussion of the paper by D. R. Cox. *Journal of Royal Statistical Society, B*, 34, 216–217.

Cao, R. and Jácome, M.A. (2004). Presmoothed kernel density estimator for censored data, *Journal of Nonparametric Statistics*, 16, 289–309.

Cao, R., López-de-Ullibarri, I., Janssen, P. and Veraverbeke N (2005). Presmoothed Kaplan-Meier and Nelson-Aalen estimators, *Journal of Nonparametric Statistics*, 17, 31–56.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–200.

de Uña-Álvarez, J. and Rodríguez-Campos, C. (2004). Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present. *Statistics*, 38, 483–496.

de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal*, 53, 113–127.

de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. *Biometrics*, 71, 364–375.

Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference*, 66, 253–279.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8, 551–561.

Genser, B. and Wernecke, K.D. (2005). Joint modelling of repeated transitions in follow-up data: a case study on breast cancer data. *Biometrical Journal*, 47, 388–401.

Gentleman, R.C., Lawless, F.F., Lindsey, J.C. and Yan, P. (1994) Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13, 805–821.

Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley.

Jácome, M.A. and Iglesias, M.C. (2008). Presmoothed estimation with left truncated and right censored data. *Communications in Statistics/Theory and Methods*, 37, 2964–2983.

Jácome, M.A. and Iglesias, M.C. (2010). Presmoothed estimation of the density function with truncated and censored data. *Statistics*, 44, 217–234.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. em Journal of the American Statistical Association, 53, 457–481.

López-de-Ullibarri, I. and Jácome, M.A. (2013) survPresmooth: An R Package for Presmoothed Estimation in Survival Analysis. *Journal of Statistical Software*, 54, 1–26.

Meyer, M.C. (2008) Inference using shape-restricted regression splines, *The Annals of Applied Statistics*, 2, 1013–1033.

Meyer, M.C. (2012) Constrained penalized splines, *The Canadian Journal of Statistics*, 40, 190–206.

Meira-Machado, L., de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 12, 325–344.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen, P.K. (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, 18, 195–222.

Meira-Machado, L., de Uña-Álvarez, J. and Datta, S. (2015). Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. *Computational Statistics*, 30, 377–397.

Meira-Machado, L., Sestelo, M. and Gonlçalves, A. (2016). Nonparametric estimation of the survival function for ordered multivariate failure time data: A comparative study. *Biometrical Journal*, 58, 623–634.

Meier-Hirmer, C. and Schumacher, M. (2013). Multi-state model for studying an intermediate event using time-dependent covariates: Application to breast cancer. *BMC Medical Research Methodology*, 13, 80.

Moreira, A.C., de Uña-Álvarez, J. and Meira-Machado, L. (2013). Presmoothing the Aalen-Johansen estimator in the illness-death model. *Electronic Journal of Statistics*, 7, 1491–1516.

Nadaraya, E. (1965). On nonparametric estimates of density functions and regression curves. Theory of Applied Probability, 10, 186–190.

Pérez-Ocón, R., Ruiz-Castro, J.E. and Gámiz-Pérez, M.L. (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society: Series C*, 50, 111–124.

Putter, H, Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26, 2389–2430.

Pya, N. and Wood, S.N. (2015) Shape constrained additive models. *Statistical Computing*, 25, 543–559.

Satten, G.A. and Datta, S. (2001). The Kaplan-Meier Estimator as an inverse-probability-of-censoring weighted average. *American Statistician*, 55, 207–210.

Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, A*, 161, 71–94.

Schmoor, C., Sauerbrei, W. Bastert, G., and Schumacher, M. (2000). Role of isolated locoregional recurrence of breast cancer: Results of four prospective studies. *Journal of Clinical Oncology*, 18, 1696–1708.

Schumacher, M., Bastert, G., Bojar, H., Hiibner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R.L.A., and Rauschecker, H.F. for the German Breast Cancer Study Group (GBSG) (1994). A randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12, 2086–2093.

van Houwelingen, H.C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34, 70–85.

van Keilegom, I., Akritas, M. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: A comparative study. *Computational Statistics and Data Analysis*, 35, 487–500.

Watson, G.S. (1964). Smooth regression analysis *Sankhya*, 26, 359–372.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press.