Statistics and Operations Research Transactions,

vol. 43, n. 2 (2019)

SORT 43 (2) July-December 2019, 193-222 DOI: 10.2436/20.8080.02.85

Modelling human network behaviour using simulation and optimization tools: the need for hybridization

Aljoscha Gruler[1], Jesica de Armas[2], Angel A. Juan[1] and David Goldsman[3]

Abstract

The inclusion of stakeholder behaviour in Operations Research / Industrial Engineering (OR/IE) models has gained much attention in recent years. Behavioural and cognitive traits of people and groups have been integrated in simulation models (mainly through agent-based approaches) as well as in optimization algorithms. However, especially the influence of relations between different actors in human networks is a broad and interdisciplinary topic that has not yet been fully investigated. This paper analyses, from an OR/IE point of view, the existing literature on behaviour-related factors in human networks. This review covers different application fields, including: supply chain management, public policies in emergency situations, and Internet-based human networks. The review reveals that the methodological approach of choice (either simulation or optimization) is highly dependent on the application area. However, an integrated approach combining simulation and optimization is rarely used. Thus, the paper proposes the hybridization of simulation with optimization as one of the best strategies to incorporate human behaviour in human networks and the resulting uncertainty, randomness, and dynamism in related OR/IE models.

MSC: 90B50, 91B06.

Keywords: Modelling human behaviour, human networks, simulation, optimization, simheuristics.

1 Introduction

Operations Research/Industrial Engineering (OR/IE) methods such as simulation and optimization are frequently employed in the design, development and optimization of complex networks and systems (Derigs, 2009). The realistic representation of these systems and networks through suitable models is hereby of major importance. Even

[1] IN3 – Dept. of Computer Science. Universitat Oberta de Catalunya. Castelldefels, Spain. {agruler, ajuanp}@uoc.edu

[2] Department of Economics and Business. Universitat Pompeu Fabra. Barcelona, Spain. jesica.dearmas@upf.edu

[3] Stewart School of Industrial and Systems Engineering. Georgia Institute of Technology. Atlanta, USA. sman@gatech.edu

Received: April 2018
Accepted: June 2019

though complex systems and networks from different application fields have been extensively studied by the OR/IE community, the consideration of realistic stakeholder behaviour in these models is not so usual (Crespo Pereira et al., 2011, Elkosantini, 2015, Neumann and Medbo, 2009). However, behavioural factors (either associated to isolated individuals or complete collective entities) are usually among the most important components in any real-life system. As such, simplifying behavioural assumptions neglecting the major impact of uncertainty, randomness, and dynamism that characterizes individual stakeholder behaviour often make OR/IE methods inapplicable in practice (Baines et al., 2004, Bendoly, Donohue and Schultz, 2006, Schultz, Schoenherr and Nembhard, 2010, Wang et al., 2015).

Nevertheless, the consideration of behavioural factors related to cognitive and social psychology is only one side of the coin. As individual agents are highly influenced by the contacts, ties, and connections shaping the group- and system dynamics of the human networks in which they operate, the modelling of human network interrelations is also of highest importance (Renfro, 2001, Russel and Norvig, 2003). Knoke and Yang (2008) even suggest that structural relations between different network actors follow the same patterns: (i) they are often more important in explaining behaviour than individual traits such as age, gender, etc.; (ii) they affect the perceptions, beliefs, and actions of individual network agents through structural mechanisms of human networks; and (iii) they are dynamic over time.

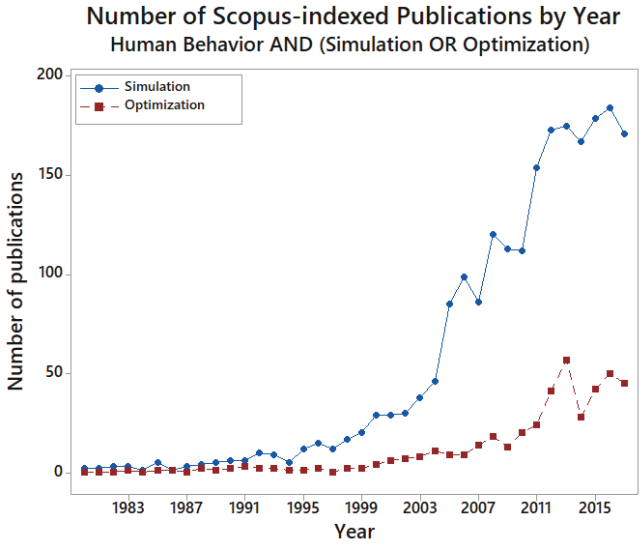


Figure 1: *Evolution of publications related to human behaviour in combination with simulation and/or optimization in Scopus indexed journals.*

The inclusion of human behaviour in simulation and optimization models has received increased attention in recent years. Figure 1 shows a clear increase in Scopus-indexed publications related to human behaviour in the context of simulation or opti-

mization. Especially in the areas of computer science, engineering, and mathematics the incorporation of complex system dynamics through behavioural traits seems to be of interest (Figure 2).

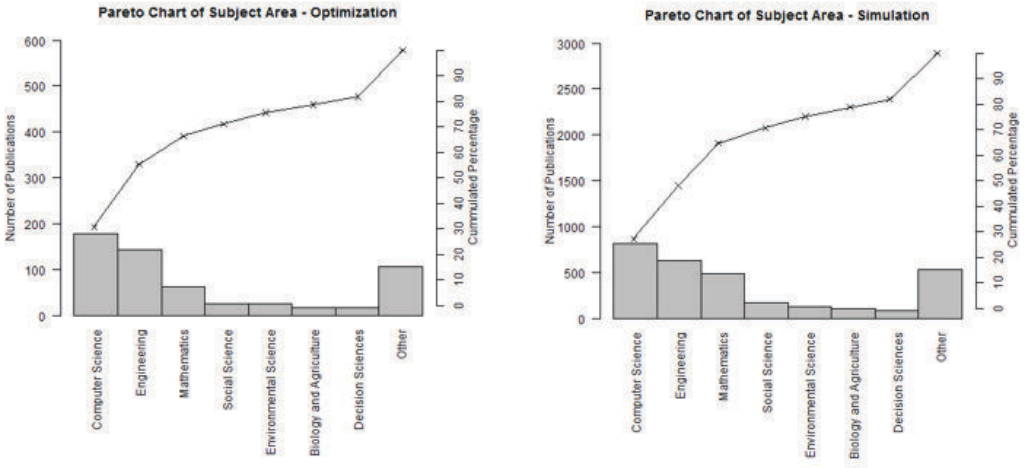


Figure 2: *Subject area of publications related to human behaviour in combination with simulation and/or optimization in Scopus indexed journals.*

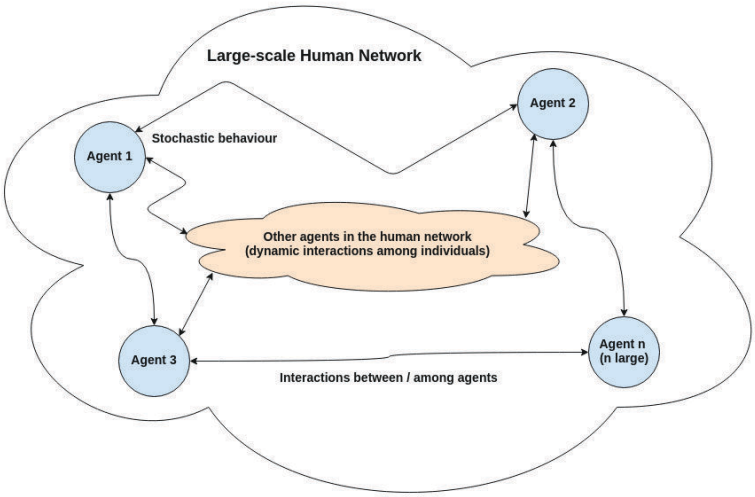


Figure 3: *Representation of a multi-agent human network.*

Simulation is mostly used to evaluate complex systems in which multiple actors interact in specific multi-agent human networks, similar to the one outlined in Figure 3. In this context, multi-agent systems (MAS) and agent-based modelling (ABM) arose with the desire to study complex and adaptive systems and their behaviours (Heath and Hill, 2010). Individual agents are thus modelled with unique attributes and behaviours, reacting to the actions, perceptions, and interrelations with other stakeholders in the mod-

Table 1: *Overview over reviewed human network environments.*

Human Network Environment	Main Focus
Supply Chain Management (Section 2)	**Manufacturing & Production** x Cooperation among workers x Workload balance and corporate social responsibility x Workers initiative and autonomy x Ergonomic conditions at work x Robustness of human network with increasing size **Logistics & Transportation** x Personal attitudes towards uncertainty in demands x Collaborative transportation management x Horizontal cooperation among carriers x Collaboration of urban freight stakeholders x Coordination in the use of shared parking spots
Public Policies in Emergency Situations (Section 3)	**Disease & Epidemics Dynamics** x Disease propagation and dynamics x Policies to limit the impact of fatal disease spreading x Infection control policies **Healthcare Emergencies** x Policies for an efficient patient care x Human resources allocation for managing patient overflow **Evacuations** x Perception of hazards in emergency situations x Efficient and real-time communication during evacuations x Navigation within social groups (crowd flow patterns) x Evacuation policies and procedures x Human interaction during evacuation of buildings x Movements of vehicles and pedestrians under emergencies x Detouring and avoiding congestion vs. greedy behaviour
Internet Social Networks (Section 4)	**Influential User Definition** x Viral marketing campaigns x Optimal display of advertising x Asymmetric influence relationships x Pricing policies **Community Establishment** x Discovery of network communities x True and false friend links x Degree of separation among users x Identification of market target groups **Other** x Trust-based relationships x Propagation of Internet-based human network viruses x Evolution of the network due to individual behaviour x Multi-agent rumor spread x Information propagation

elled system environment (Bandini, Manzoni and Vizzari, 2009, Kennedy, 2010, Macal and North, 2010, Siebers, Aickelin and Menachof, 2008). While ABM is the most common simulation approach to model behavioural traits in a human network environment, discrete event simulation (DES) has also been applied in to model resulting system dynamics (Robinson, 2014). However, this more process-oriented approach is less commonly applied to model agent behaviour and their impact on human networks (Siebers et al., 2010). DES frameworks for modelling social behaviour in networks are presented for example by Alt and Lieberman (2010) and Hou et al. (2013).

While simulation seems to be the natural way to incorporate human behaviour dynamics and the resulting randomness in the evaluation of human network structures, optimization is usually required to increase the efficiency of related processes. Resulting optimization problems are either addressed by exact solution methods for smaller instances, or approximate methods such as metaheuristics for larger problem settings (Talbi, 2006, Vazirani, 2012). Dynamic and uncertainty conditions are usually modelled using random variables in objective functions and constraints, e.g. through Monte Carlo simulation (MCS) or fuzzy logic.

This paper reviews existing work from the simulation and optimization fields in which human behaviour in human networks has been successfully modelled. Accordingly, the focus is put on three human network environments in which agent behaviour and stakeholder interrelations play a decisive role, namely supply chain management (SCM), the evaluation of public policies in emergency situations, and the structural analysis human network users in the Internet. Table 1 provides a more detailed overview over the discussed application fields. From this critical analysis of existing simulation and optimization models, a second contribution emerges: we provide arguments supporting the need for hybridizing simulation with optimization methods as a natural way to include human network behaviour in OR/IE models.

Accordingly, this work is structured as follows: Section 2 analyses the literature focusing on human behaviour in SCM. Section 3 discusses works on human behaviour in public policies in emergency situations. Section 4 is devoted to examine literature on human behaviour in Internet human networks. Hybrid simulation-optimization, as a natural way to include human network behaviour in optimization models, is closer discussed in Section 5. To conclude, Section 6 highlights the main contributions of this work.

2 Considering behavioural traits in supply chain networks

The efficient organization of material and information flows in SCM requires effective interaction and cooperation between different supply chain agents. To realistically model the resulting human network dynamics, different behavioural issues have to be addressed. In the following, Section 2.1 analyses existing literature concerning manufacturing & production processes in which human network dynamics (mainly through

the interaction between individual employees) are considered. Later, behavioural aspects in the design and evaluation of logistics & transportation concepts such as collaborative transportation management or city logistics are reviewed in Section 2.2. An overview of the analysed papers and their OR/IE modelling approach to address human network behaviour is given in Table 2. Notice that many of these papers do not consider optimization.

2.1 Applications in Manufacturing & Production

The importance of considering human behaviour in manufacturing systems by integrating psychological and emotional aspects is stressed by Elkosantini and Gien (2009). In their work, they propose an agent-based simulation model to represent a production line including workers, whereby special attention is paid to individual behaviour and social relationships between workers. The authors highlight the influence of social interaction among employees on individual performance levels. Other OR/IE models in the context of manufacturing & production address individual human factors such as fatigue, motivation, education, or personalities (Digiesi et al., 2009, Elkosantini, 2015, Huerta, Fernandez and Koutanoglu, 2007, Khan, Jaber and Guirida, 2012, Riedel et al., 2009, Silva et al., 2013). Also in this context, Grosse et al. (2015) developed a framework that allows the integration of human-related factors into models associated with the planning of tasks.

Spier and Kempf (1995) were among the first authors in proposing the inclusion of human interrelations in simulation models. The authors use object-oriented simulation to test learning effects among workers in a small manufacturing line, showing that proactive and cooperative agents provide the best company performance. More recent work on similar issues apply ABM or DES to analyse, simulate, and evaluate production lines, workforce allocation in manufacturing cells, or the impact of engineers in the product design process. Okuda et al. (1999) stress that cooperation can be a key attribute in the planning of efficient production lines. The authors test different production process designs (e.g. U-shaped production lines and manufacturing cells) in terms of workload balance and total throughput in small-lot manufacturing, characterized by a high need for production flexibility. By using ABM, the impact of cooperation through human-oriented production lines is assessed. The paper concludes that production processes taking into account human behaviours (inter-worker learning effects) achieve the most balanced working times and the highest company output.

Various simulation models focusing on workforce allocation in production lines have been developed in the past. However, most of them do not consider the impact of human behaviour and collaboration in their models. Zhang et al. (2015) overcome this drawback by integrating different models of human agents in the context of a dynamic systems with a discrete-event behaviour, which they use to evaluate changeover processes in manufacturing processes. By modelling and simulating the dynamics of work process together with the dynamics of human behaviour, the authors show that the in-

corporation of different cooperation styles and skill levels can has a noticeable effect on the expected throughput of the system. From the reported simulation experiments, it can also be concluded that changeover assignments based on collaborative strategies lead to the best system performance.

The effects of collaborative product design processes (PDPs) are studied by Yang, Song and Zhang (2007). They argue that many simulation methods applied in the planning of efficient PDPs are too task oriented and do not represent the central role of designers in the process, which is deeply impacted by human initiative and autonomy, but also by collaboration within project teams. They elaborate an ABM to predict, manage, evaluate, and improve manufacturing design processes. Therefore, the design evaluation depends on the degree of cooperative behaviour among the product design team-members. Furthermore, human factors such as efficiency of designer and organization, human workload, error, and collaboration levels are taken into account. Another ABM to represent the dominant role of product designers in PDPs is proposed by Li, Zhang and Zhang (2011). The designer agents in their model have distinctive characteristics such as initiative, autonomy, and collaboration skills. They construct a simulation model of a motorcycle design project, allowing to analyse PDP traits such as organizational structure, scheduling strategies, and partner selection while considering individual and social behavioural traits.

ABM seems to be the predominant method of choice for modelling and simulating social interactions in manufacturing systems. However, there are some works that address the issue by using DES approaches. Crespo Pereira et al. (2011) propose a manufacturing DES environment that allows them to conduct training and research on how human operations take place. Their experimental system allows the consideration of human factors such as inter-group differences, worker experience, buffer capacities, work-sharing, and process state perception. Experimental results based on a real-life case show that inter-group variations, experience, and ergonomic conditions have a significant impact on the process outcome. Also using DES, Putnik et al. (2015) test the robustness of large production networks in environments with demand uncertainty. By modelling the behaviour of socially connected individuals, their work shows that system robustness and production rates depend on system sizes and human networks. According to their simulation experiments, large human networks with lots of business relations positively impact network robustness, while the production rate exhibits a nontrivial relation to the number of connections.

2.2 Applications in Logistics & Transportation

In the face of increasing market complexity driven by rapidly changing customer preferences, globalization, and fierce competition, the need for effective supply chain management (SCM) among suppliers, manufacturers, distributors, and retailers lead to complex dynamic systems. In response to this, many innovative planning models of logistics & transportation systems such as collaborative transportation management (CTM) or city

logistics (CL) are based on the idea of stakeholder collaboration (Crainic, Ricciardi and Storchi, 2009, Taniguchi, Thompson and Yamada, 2012, Benjelloun and Crainic, 2009). Consequently, behavioural factors on an individual and network level have to be considered in the design of sustainable and integrated transportation & logistics structures (Geary, Disney and Towill, 2006, Sarimveis et al., 2008).

In the context of CTM, different simulation approaches have been used to include cognitive behaviour (e.g., individual thinking, deciding, and reasoning processes) and social factors (e.g., relationships and inter-organizational influences). As such, Yuan and Shon (2008) propose a CTM model based on the collaboration in transportation management among different supply chain partners. Their simulation tool is developed as realistic representation of a beer supply chain with four levels. The authors show that transportation costs and vehicle utilization levels can be significantly improved by collaboration and coordination of transportation activities. Chan and Zhang (2011) use Monte Carlo simulation (MCS) to evaluate benefits of CTM in long term relationships between retailers and carriers. The authors illustrate the concept of carrier flexibility to optimize delivery lead times. Their results show that collaboration between both parties can reduce retail costs while improving service levels. A conceptual framework for a behavioural multi-agent model considering the impact of cognitive and social behaviour is presented by Okdinawati, Simatupang and Sunitiyoso (2014). They propose the inclusion of Drama Theory (see Bryant (2003, 2004)) in their model. This allows the consideration of stakeholder behaviour in conflict and collaboration scenarios during the hierarchical decision making process of CTM strategies on an operational, tactical, and strategic level.

Using ABM, Li and Chan (2012) describe the impact of CTM on SCM with stochastic demands. They simulate a three-level supply chain while taking into account factors such as company characteristics, their types of action, and changes in company behaviour. By comparing the efficiency level reached in non-cooperative scenario with the cooperative case, the authors show that CTM can reduce global costs while increasing supply chain flexibility. Their work concludes that CTM is an efficient approach to tackle demand disruptions. Yu, Ting and Chen (2010) use DES to test different information sharing scenarios between supply chain members. More specifically, they consider information sharing about demand, inventories, capacities, and their different combinations. Their results suggest that especially information-sharing concerning customer demands is critical for supply chain success. Furthermore, they show that a full cooperative scenario based on shared information and assets is ideal for obtaining higher levels of efficiency in most supply chains. There are some metaheuristic approaches that address similar concepts (e.g. Horizontal Cooperation) in which interactions between network actors are highly important (Pérez-Bernabeu et al., 2015, Quintero-Araujo et al., 2019), but these optimization methods have not yet reached the same integration level of behavioural issues as simulation approaches.

Related to CL, Tamagawa, Taniguchi and Yamada (2010) develop a multi-agent methodology to evaluate different CL measures (road pricing, truck bans, motorway

tolls) taking into account the behaviour of partners in urban freight transportation. More specifically, the modelled agents represent motorway operators, administrators, residents, shippers, and freight carriers. The authors develop an acceptable network environment for all stakeholders by considering conflicting objectives, transportation cost, profits, and environmental effects. To evaluate different road networks, they apply a genetic algorithm to calculate different routing options from the resulting vehicle routing problem (VRP), which has to consider time windows. Furthermore, a learning prototype affecting the behaviour of different agents is implemented. This paper extends a similar work of Taniguchi, Yamada and Okamoto (2007), in which the authors show that the implementation of road pricing can reduce pollution emissions but may increase freight shipment costs. To avoid such effects, cooperative freight transportation systems are proposed.

Teo, Taniguchi and Qureshi (2012) test government measures affecting urban road networks (e.g., road pricing for trucks) in an e-commerce delivery system. Their ABM considers the behaviour of major stakeholders in the transportation environment. In particular, they propose a reinforcement learning strategy for administrators to represent realistic agent behaviour. Furthermore, the resulting routing problem is optimized with an insertion heuristic. According to their outcomes, when the government administrator considers freight vehicle road pricing, truck emissions can be significantly reduced. A multi-agent approach to evaluate the financial and environmental impact of implementing urban distribution centres in urban areas is presented by Duin et al. (2011). The authors consider and test the dynamic behaviour among different CL stakeholders. Moreover, the impact of stakeholders' behaviour and actions towards city measures like tolls, operational subsidies, or time windows, and entry restrictions within city centres is evaluated. Experimental results suggest that the development of a positive business environment for urban freight consolidation centres depends not only on physical factors such as traffic congestion, but also on the actions and behaviour of each system agent. Their ABM also incorporates a genetic algorithm for routing optimization.

Joint delivery systems, urban distribution centres, and car parking management within city centres are the CL measures analysed by Wangapisit et al. (2014). The focus of this study lies on the interaction and cooperation among urban freight stakeholders when CL measures are implemented. The authors use ABM combined with reinforcement learning and an insertion heuristic to solve extended VRP versions with pick-up-and-delivery and time windows. Their results suggest that urban distribution centres and joint delivery systems can improve the environmental impact of urban freight transportation. They fine-tune the distribution centre implementation by applying urban parking management and subsidies for shopping street associations. Car parking management in an ABM is also the matter of research in the paper by Boussier et al. (2009). Their simulation models considers the behaviour of different agents, focusing of shared parking spaces between private and commercial vehicles. Furthermore, the use of electric vehicle fleets in the development of 'greener' transportation systems is taken into account in some works (Juan et al., 2016, Eskandarpour et al., 2019).

Table 2: *Summary of reviewed papers related to Supply Chain Management.*

Area	Paper	Simulation			Optimization
		ABM	DES	Other	(Meta-) Heuristics
Manufacturing & Production	Spier and Kempf (1995)			x	
	Okuda et al. (1999)	x			
	Yang et al. (2007)	x			
	Li et al. (2011)	x			
	Crespo Pereira et al. (2011)		x		
	Zhang et al. (2015)	x			
	Putnik et al. (2015)		x		
Logistics & Transportation	Taniguchi et al. (2007)	x			x
	Yuan and Shon (2008)			x	
	Boussier et al. (2009)	x			
	Yu et al. (2010)		x		
	Tamagawa et al. (2010)	x			x
	Chan and Zhang (2011)			x	
	Li and Chan (2012)	x			
	Teo et al. (2012)	x			x
	van Duin et al. (2012)	x			x
	Wangapisit et al. (2014)	x			x
	Okdinawati et al. (2014)	x			

3 The impact of public policies on human networks behaviour in emergency situations

This section reviews different approaches in which human network behaviour in emergency situations is addressed. In this field, especially the reaction of complete population groups to public policies in the face of disease and epidemic dynamics (Section 3.1), other healthcare emergencies (Section 3.2), and evacuation situations (Section 3.3), has recently been a topic of interest. Table 3 summarizes the reviewed works.

3.1 Applications in Disease & Epidemics dynamics

Human behaviour in human networks has a strong influence on how civil infrastructures are used. Thus, whenever public polices have to be designed these human factor should be considered. Likewise, social interactions provide an ideal environment in which diseases can easily be spread out. For those reasons, social interactions need to be considered when designing public policies, since the behaviour of citizens in response to these policies as well as their reaction to situations of crisis can modify the usual social patterns.

The most remarkable papers in the literature regarding simulation of these issues use ABM simulation. Thus, related to policy making in large-scale networks, Kasaie and Kelton (2013) consider the problem of resource allocation in the control of epidemics. They assume a fixed budget to be allocated among competing healthcare interventions, with the goal of achieving the best health benefits. Interventions thus include vaccination, prevention, or treatment programs. Their constructed ABM is combined with a response surface methodology as sequential optimization technique for the resulting resource allocation problem, depending on different investment strategies.

Considering realistic and large-scale human networks, Bisset et al. (2009) analyse the evolution of human behaviour and disease dynamics. These authors use a highly detailed interaction-based simulator to simulate sixteen scenarios, established by combining different types of intervention policies during the spreading of fatal diseases: closure and reopening of schools, quarantine policies, and vaccinations policies. Published results indicate that quarantine and other isolation polices seem to have a limited impact on the overall rate of infection. Also, individual isolation policies are typically employed at late stages of the epidemic outspread, which in practice limits their effectiveness. Other isolation policies (e.g, quarantine of some individuals) tend to affect only a small portion of the total population, which also limits their efficiency. However, a combination of vaccinations and quarantine policies seems to be effective since the number of key citizens infected is reduced. Concerning school closures, the results suggest that even very low levels ($<0.1\%$) of residual infection rates among pupils can cause new infection waves after disease epidemics.

Focusing on a smaller and enclosed area, Laskowski et al. (2011) proposed a model to study, by means of simulation, the spread of influenza virus infections in the emergency department of a Canadian hospital. Their simulation used a set of patients and healthcare workers, modelling their individual properties as well as their social interactions. According to their results, those policies oriented to controlling the infection in patients (e.g., masking symptomatic patients or alternate treatment streams) are usually more efficient than those other policies focused just on healthcare workers.

3.2 Applications in Healthcare Emergencies

Effective management of Emergency Departments (ED) is an important problem in healthcare systems. The frequency of arrival of patients, the waiting time of patients, the treatment given, the emotions of the doctor, the nurse management of patients, etc. are factors that affects the quality of ED. Overcrowding and high flow in ED will have higher probability of conflict occurrence. Conflict happens in every ED, so a good policy is needed to confront the crowded situation in order to maintain the quality of ED. The analysis of this particular human network behaviour play a key role in developing policies and decision tools for overall performance improvement of the system. The ability to accurately represent, simulate and predict performance of ED is invaluable for decision makers.

Brailsford (2016) develop a review on simulation models for healthcare applications in which the simulated objects (entities) are human beings. This study focuses specifically on whether it is desirable (and possible) to incorporate human behaviour within the conceptual design of a simulation model. However, the reality is that, although there are many approaches in the literature dealing with healthcare emergencies through simulation (Almagooshi, 2015), only a few of them consider human network behaviour and optimization. Thus, Rico, Salari and Centeno (2007) study the best nurse allocation policy to manage patient overflow during a pandemic influenza outbreak. Their approach combines DES with OptQuest - an optimization software that includes metaheuristics, exact methods and neural networks - in order to analyse different configurations regarding the number of nurses needed for healthcare delivery. Some other works in the literature apply the same combination of a DES model and OptQuest including human network behaviour in healthcare emergencies. Thus, Silva and Pinto (2010) evaluate the performance of a medical emergency system creating a simulation model and using optimization to analyse different scenarios and find the best parameters for it. Similarly, Weng et al. (2011) use this combination to optimize the allocation of human resources in a hospital emergency department. With a different methodology, Liu (2017) analyse a complex Spanish ED and provide an ABM simulation considering patient arrivals based on historical data. The interaction between doctors, nurses, technicians, receptionists, and patients is studied and modelled. Additionally, some optimization methodology is used for calibrating model parameters under data scarcity.

3.3 Applications in Evacuation Situations

The perception of risk during emergency evacuations can generate stress on the population, which can derive in selfish and unorganized behaviours driven by the survival instinct (for example, by blocking narrow evacuation exits). This seriously effects survival rates and the evacuation efficiency levels. In this context, the analysis of human behaviour during emergency situations contributes to build efficient emergency management plans. As such, Parikh et al. (2013) note the importance of communication in such events. Their ABM considers population behaviour and its interaction with various interdependent infrastructures, in order to develop efficient evacuation plans considering a nuclear detonation. Their results stress the key role of agent communication, as it can beneficially alter human behaviour in the evacuation phase by reducing crowd panic and increase mobility levels.

Chu et al. (2015a) propose an agent-based simulation tool that is able to consider both human and social behaviours previously analysed in scientific works related to management of disaster and safety situations. They use several approaches to model the behaviour of each agent: (i) the user follows exits that are familiar to her; (ii) the user follows cues from building features; (iii) the user will navigate within a group of related people; (iv) and the user will follow the crowd. As expected, their simulation results show that the flow patterns might be greatly influenced by the specific arrange-

ment of exit signs, knowledge of the environment, and even social settings. Thus, these authors are able to pinpoint and evaluate the effect of social features on flow patterns. Their analysis provides insights concerning architecture, building layouts, and facility management in the design of user-centric facilities, emergency procedures, and related training programs. Later, the authors applied the same simulation tool to examine egress performance of a museum (Chu et al., 2015b). Their simulation considers different scenarios of people and group behaviour in emergency situations. Their approach allows a closer analysis of museum visitors in emergency situations to improve the design of safe egress systems and procedures.

Similarly, to represent an evacuation situation Chu and Law (2013) propose an agent-based simulation study in which social behaviour is considered. Agents are represented considering the incorporation of behavioural rules on an individual, group, and crowd level. Results reveal that social behaviour during evacuation processes can affect the overall egress time and pattern. In a model combining human behaviour with buildings, Liu et al. (2016) study the dynamic effect of damaged structures on the evacuation of buildings. Their agent-based model hybridizes probabilistic components with finite-element theory to analyse how people interact during the evacuation process. The reported simulation results show that the evacuation time can suffer a noticeable increment when considering the grouping behaviour.

Table 3: *Summary reviewed papers related to public policies in emergency situations.*

Area	Paper	Simulation			Optimization		
		ABM	DES	Other	Exact & Approximation Methods	(Meta-) Heuristics	Other
Dynamics of Diseases & Epidemics	Bisset et al. (2009)			x			
	Laskowski et al. (2011)	x					
	Kasaie and Kelton (2013)	x			x		
Healthcare Emergencies	Rico et al. (2007)		x				x
	Silva and Pinto (2010)		x				x
	Weng et al. (2011)		x				x
	Liu (2017)	x					x
Evacuations	Kagaya et al. (2005)	x					
	Zhang et al. (2009)	x					
	Song et al. (2010)	x					
	Luh et al. (2012)	x				x	
	Chu and Law (2013)	x					
	Parikh et al. (2013)	x					
	Chu et al. (2015a)	x					
	Chu et al. (2015b)	x					
	Fu et al. (2015)			x			
	Liu et al. (2016)	x					

Unlike the studies mentioned above, Luh et al. (2012) make an effort to integrate optimization techniques into their model. In order to do so, these authors use a macroscopic network-flow model. Their model takes into account different factors related to

the desire of escaping, such as smoke, fire, and even psychological ones. Thus, they employ stochastic dynamic programming to optimize escape routes for both groups and individuals. To reduce the impact of limited passage capacities on the evacuation flow, these routes are also coordinated in their model. According to the reported results, their approach is able to reduce evacuation time by diminishing bottlenecks through the path.

Next to the evacuation of people from buildings and facilities, human behaviour in human networks concerning traffic management play an important role when people are forced to leave whole areas, such as villages or cities. In particular, human behaviour might be influenced by different circumstances: *(i)* people tend to focus mainly on the prevailing situation instead of long-term interests; *(ii)* the availability and quality of traffic information might have a strong impact on human behaviour, including the choice of the escaping route; and *(iii)* instructions on evacuation paths might also affect the selection of the evacuation route. The route choice behaviour during evacuation processes is formulated by Fu et al. (2015) as a combination of the instructed route and the own user's perception on the time it might take to complete it. Thus, to model the user's behaviour a logit model and fuzzy set theory is used. According to the results provided by a simulation, there is a nonlinear impact of traffic data on the efficiency of the evacuation flow. Also, whenever real-time traffic data is available online, users are able to adapt their chosen paths, thus reducing the associated egressing times. Furthermore, a strong compliance enforcement concerning policy instructions contributes to higher evacuation efficiency.

Another urban emergency transportation simulation system is presented by Song, Yang and Du (2010). Their system is based on the Beijing metropolitan area with the focus of simulating vehicle and pedestrian movements in emergency situations. Results demonstrate the effectiveness of this system for producing evacuation routing strategies, optimizing emergency resources, identifying total evacuation times, and evaluating the performance of the whole operation. Similarly, Zhang, Chan and Ukkusuri (2009) address human interaction during evacuation processes. They use greedy agents that use a probabilistic rule and take into account the dynamic conditions of the network to select between the shortest path and the least congested one. By detouring and avoiding congested roads, some agents might be able to diminish their individual egressing times. However, this greedy behaviour also tends to increase the total time employed by all the agents to evacuate the system. Considering hazards caused by earthquakes, Kagaya et al. (2005) build the reproduction of human traffic behaviour and considering agent interaction. They classify evacuation behaviour into various patterns, which they then use to establish different rules concerning evacuation behaviour.

4 User behaviour in Internet-based human networks

The growing use of Internet-based human networks increasingly influences individual and collective conduct of people (Jin et al., 2013). Especially in the context of defining influential users and communities – for example, related to internet security and online marketing – structural network analysis using network and graph theories has received much attention. Social factors are implicit in this application area. For this reason, OR/IE problems related to the analysis of Internet-based human networks are a direct consequence of human networks and their dynamic behaviours. In contrast to previous application areas, where social factors should to be taken into account because of their influence on particular situations, problems discussed in this Section can be seen as direct consequence of human network behaviour among users. In more detail, Section 4.1 deals with research on individual network users. Then, Section 4.2 reviews papers in which network community structures are defined and analysed using OR/IE approaches. Furthermore, other related works are discussed in Section 4.3. Table 4 summarizes the works discussed in this section. Notice again that column DES is not included in this table, since none of the approaches use it.

4.1 Identifying influential network users

Internet-based human networks are becoming more important for companies in the context of efficient and productive viral marketing campaigns. The influence maximization problem in Internet-based human networks was proposed by Domingos & Richardson (Domingos and Richardson, 2001, Richardson and Domingos, 2002). When modelling the Internet-based human network on a graph, the goal is to find a subset of nodes with the highest influence on the rest of the network. As shown in Figure 4, some individuals (nodes) might be more 'influential' than others, meaning that they have a larger number of connections (i.e., their opinions or actions might reach a large number of individuals). Also, not all connections are symmetrical: while some agents might be very influential over their contacts, the opposite is not always true. Heuristics are proposed as a tool to decide upon the most influential customers in the network. The idea is to focus marketing activities on customers with a high network value, instead of only considering the related expected intrinsic (direct) marketing value of each network member.

Kempe, Kleinberg and Tardos (2003, 2005) develop probabilistic rules based on findings from sociology and economics, which they embed into a decreasing cascade- and linear threshold model. They use greedy approximation algorithms to achieve influence maximization. The proposed greedy algorithms were later improved by Chen, Wang and Yang (2009). These authors also discuss an efficient degree discount heuristic, which is able to reach similar influence spreading results in substantially decreased calculation times. Considering a probabilistic voter model, Even-Dar and Shapira (2007) analyse the spread maximization problem. For that, they elaborate simple and efficient algo-

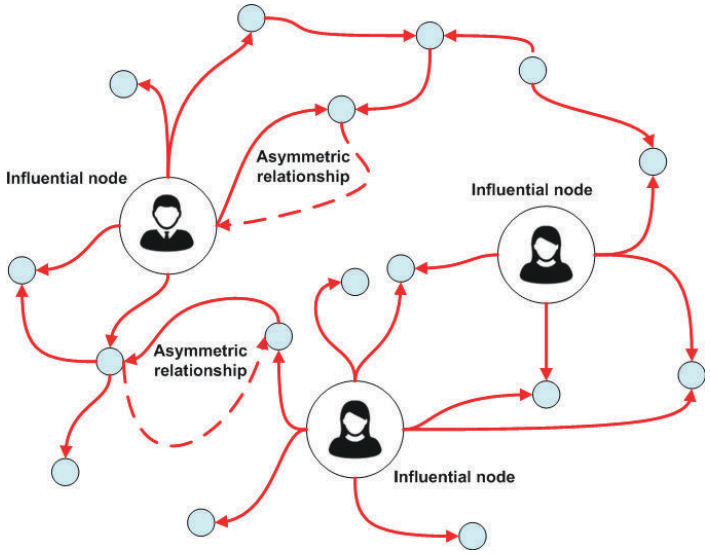


Figure 4: *Internet-based human network with influential nodes and asymmetric relationships.*

rithms. Kimura et al. (2010) later introduced an approach based on graph theory and bond percolation to reduce algorithmic computation times.

Some studies deal with the problem by considering competitive diffusion. Carnes et al. (2007) employ viral marketing to introduce a new product when a competing one already exists in the market. They assume that if an influential user chooses one product over another, then the members of his/her Internet-based human network will tend to do the same. These authors propose an approximation algorithm that is able to reach 63% of the optimal value. Later, Borodin, Filmus and Oren (2010) discussed a similar competitive environment, introducing a different approach to the original greedy one. Also, taking into account display advertising, the influence maximization problem has been explored by Abbassi, Bhaskara and Misra (2015). Here, online advertisement is shown to a pre-defined number of users. In order to find the optimal display strategy, these authors introduce alternative optimization heuristics. After completing a MCS study, their results show that especially a two-stage algorithm inspired by influence-and-exploit strategies yields promising results.

Most of the reviewed papers consider that social relationships can be modelled using undirected graphs (i.e., they are symmetric). However, trust and other social relationships might need to be modelled using directed graphs (i.e., they might be asymmetric or even unilateral). Xu et al. (2012) model the Internet-based human network on a directed graph including asymmetric influence relationships. In order to find a subset of users that have the highest influence in the network, they propose a mathematical programming approach. This is empirically evaluated using real-life data from Internet-based human networking sites. Ahmed and Ezeife (2013) develop a diffusion model that considers positive and negative trust influences in Internet-based human networks.

Influential nodes are identified with a local search based algorithm, which outperforms greedy approximation methods by as much as 35%.

The network influence effect is also considered in combination with the optimal pricing problem, in which different pricing policies are considered in the diffusion of a product. Considering a monopolistic market setting, Candogan, Bimpikis and Ozdaglar (2010) consider different scenarios concerning pricing policies (uniform, two-fold, and individual prices for the customers in the network) for a divisible good. Considering these scenarios, the authors propose an approximation algorithm for finding the optimal set of agents. Chen et al. (2010) propose an analogous concept. By taking into account incomplete information, these authors are able to extend the original model. Also related to this, a multi-stage pricing model is introduced by Hartline, Mirrokni and Sundararajan (2008). In this model, different price levels are set, at each different stage, by the manager. This work was later improved by Akhlaghpour et al. (2010) to include imperfect information of the considered agents.

4.2 Community discovery and structural analysis

Apart from identifying key network members and their influence on the behaviour of related nodes, another major research field concerning internet human networks is related to the detection of clusters and communities to structurally analyse large networks. In their comparison of network detection methods (i.e., approximation and heuristic algorithms), the concept of 'network community' is defined by Leskovec, Lang and Mahoney (2010) as "a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network". The goal is to define such communities to study their behaviour over time. The authors name different approaches to identify network clusters. Principal component analysis are used in spectral algorithms to find communities (Kannan, Vempala and Vetta, 2004). Likewise, algorithms based on network flow represent edges by means of pipes with unitary capacity, and then are able to find communities by employing algorithms such as the max flow-min cut one (Flake, Tarjan and Tsioutsiouliklis, 2003). Other authors count the number of edges pointing inside and outside a giving community (Flake, Lawrence and Giles, 2000, Radicchi et al., 2004), which allows them to identify clusters in the network. Other works are concerned with maximizing the modularity of the identified communities (Girvan and Newman, 2002, Newman and Girvan, 2004).

Addressing the problem of maximizing modularity, Nascimento and Pitsoulis (2013) propose the use of a GRASP metaheuristic combined with path relinking. In other related work especially the use of memetic- and bio-inspired algorithms seem to be a major trend in recent years. Chen and Qiu (2013) introduce a novel particle swarm optimization algorithm, showing through synthetic and real-world networks that it can effectively extract the intrinsic community structures. Other particle swarm optimization algorithms have been applied in the same context by several works (Cai et al., 2014, 2015, Biswas et al., 2015). Ant colony optimization (ACO) has been applied by Sercan,

Sima and Sule (2009) in mapping network-cliques to graph nodes. The resulting graph is then analysed using clustering based algorithms. Additional publications applying ACO algorithms to discover network communities include (Javadi et al., 2014, Jin et al., 2011, 2012, Mandala et al., 2013, Xu, Chen and Zou, 2013, Zhou et al., 2015). Further bio-inspired metaheuristics, such as artificial bee colony optimization algorithms, have also been used to deal with the problem (Abu Naser and Alshattnawi, 2014). Finally, also memetic or hybrid algorithms have been proposed in this context, e.g.: crossover operators combined with local search procedures (Gach and Hao, 2012), or a particle swarm optimization-based global search operator and tabu local search operator (Zhang et al., 2016).

Related as well to network community research and their structural analysis, in human networks over the Internet Huang, Lin and Wu (2011) propose an exact- and heuristic algorithm to define links that are either true- or false-friend ones. Other works are interested in identifying the degree of separation between two users. For instance, Bakhshandeh et al. (2011) present new heuristic search techniques to provide optimal or near-optimal solutions. Finally, Rivero et al. (2011) elaborated a metaheuristic algorithm based on ACO to perform the path search between two nodes in a graph. This algorithm outperforms other ACO algorithms when considering large-scale networks.

4.3 Other Internet-based human network analysis

Next to the definition of influential users and analysis of community structures in Internet-based human networks, other related topics can be defined in the discussed context. Zhang et al. (2008) and Ben-Zwi et al. (2009) determine marketing target groups – the set of users with the highest influence on their network acquaintances – by studying the trust relationships between customers in virtual communities. On the one hand, this problem is not exactly the same as the influence maximization one, since the objective is not to arrive to a higher number of nodes, but to identify node clusters with high trust levels. On the other hand, it is also different from the community discovery problem, since it is not based on network connectivity, but rather on trust-based relationships, making traditional clustering algorithms inapplicable in this kind of scenario.

Wen et al. (2013) use numerical simulation for their susceptible-infectious-immunized model, which allows them to analyse worm propagation in Internet-based human networks. In a similar approach, Singh and Singh (2012) study the inoculation of a certain fraction of nodes against rumors. For the modelling of specific agent behaviour in particular situations (e.g., when studying the evolution of the network as a result of personal member attributes and behaviours), numerical approaches are unsuitable, usually making ABM the preferred method of choice. Blanco-Moreno, Fuentes-Fernández and Pavón (2011) make use of agent-based simulation to analyse Internet-based human networks. Their framework allows the study of scenarios in which network members are modelled by characterized agents. These agents are customized taking into account other individuals, environmental conditions, groups, and the status of the entire net-

Table 4: *Summary of reviewed papers related to Internet human networks.*

Area	Paper	Simulation		Optimization	
		ABM	**Other**	**Exact & Approximation Methods**	**(Meta-) Heuristics**
Influence of Individual Network Users	Domingos and Richardson (2001)				x
	Richardson and Domingos (2002)				x
	Kempe et al. (2003)			x	
	Kempe et al. (2005)			x	
	Even-Dar and Shapira (2007)			x	
	Carnes et al. (2007)			x	
	Hartline et al. (2008)			x	
	Chen et al. (2009)			x	x
	Kimura et al. (2010)			x	
	Borodin et al. (2010)			x	
	Candogan et al. (2010)			x	
	Chen et al. (2010)			x	
	Akhlaghpour et al. (2010)			x	
	Xu et al. (2012)			x	
	Ahmed and Ezeife (2013)				x
	Abbassi et al. (2015)		x		x
Analysis of Network Community Structures	Shi et al. (2009)				x
	Sercan et al. (2009)				x
	Jin et al. (2011)				x
	Huang et al. (2011)			x	x
	Bakhshandeh et al. (2011)				x
	Rivero et al. (2011)				x
	Jin et al. (2012)				x
	Jin et al. (2011)				x
	Gach and Hao (2012)				x
	Nascimento and Pitsoulis (2013)				x
	Chen and Qiu (2013)				x
	Mandala et al. (2013)				x
	Chang et al. (2013)				x
	Xu et al. (2013)				x
	Qu (2014)				x
	Cai et al. (2014)				x
	Javadi et al. (2014)				x
	Abu Naser and Alshattnawi (2014)				x
	Biswas et al. (2015)				x
	Cai et al. (2015)				x
	Zhou et al. (2015)				x
	Zhang et al. (2016)				x
Other related Papers	Zhang et al. (2008)				x
	Ben-Zwi et al. (2009)			x	
	Blanco-Moreno et al. (2011)	x			
	Xiao and Yu (2011)	x			
	Singh and Singh (2012)		x		
	Sabater and Sierra (2002)	x			
	Kannabe et al. (2012)	x			x
	Wen et al. (2013)		x		

work. Notice that the use of agent-based simulation helps to develop more realistic models as well as to understand how the networks perform from the interactions among their nodes. Xiao and Yu (2011) develop a multi-agent rumor spread model in virtual communities. Different simulation tests are conducted to show the impact of network structures, rumor tolerance frequency, and the user's believing rate. Also using ABM, Sabater and Sierra (2002) propose a model based on the user's reputation, which contributes to enhance its level of representativeness as regards as certain networks.

Finally, Kannabe et al. (2012) constitute an excellent example of a hybrid approach combining metaheuristics with simulation (ABMS). These authors develop a propagation model to analyse how information spread in a Internet-based human network, thus affecting human behaviour. According to their outcomes, the effects of this propagation varies from homogeneous networks (those in which agents share similar characteristics) to heterogeneous ones.

5 Need for an integrated simulation-optimization approach

The literature review completed in the previous sections shows that the choice of the appropriate OR/IE methodology is highly context dependent. It seems that in some application areas (especially Manufacturing & Production and public policies in emergency situations), individual and network behaviour is mainly considered within the simulation community, whereas optimization tools are often applied in the design and evaluation of Internet human networks. However, in all discussed application areas it is necessary to account for the uncertainty associated with individual behaviour And the system dynamics that characterize complex network interactions when modelling behavioural traits.

Simulation techniques seem to offer a natural and efficient way to model both uncertainty and system dynamics over time. In particular, ABM has been successfully applied in a myriad of different application fields. Simulation itself, however, is not an optimization tool. Thus, whenever the problem at hand requires maximization or minimization of a given objective function (or several ones in the case of multi-objective optimization), simulation alone is not enough. A logical way to proceed in those cases is to combine simulation with optimization techniques.

As pointed out by Figueira and Almada-Lobo (2014), 'sim-opt' methods are designed to combine the best of both worlds in order to face: (i) optimization problems with stochastic components; and (ii) simulation models with optimization requirements. A discussion on how random search can be incorporated in simulation-optimization approaches is provided by Andradóttir (2006), while reviews and tutorials on simulation-optimization can be found in Fu, Glover and April (2005), Chau et al. (2014), and Jian and Henderson (2015). Since most human networks tend to be large-scale, the integration of simulation with metaheuristics (i.e., simheuristics) might become an ef-

fective way to include human factors inside *NP-hard* combinatorial optimization problems. Juan et al. (2018) provides a complete review of simheuristics (combination of simulation with metaheuristics), which facilitates to account for uncertainty in this kind of OR/IE problems. As discussed in Ferone et al. (2018), simheuristics allow for extendng traditional metaheuristic frameworks to solve large-scale complex problems with stochastic components, from transportation (Gonzalez-Martin et al., 2018) to telecommunication systems (Cabrera et al., 2014).

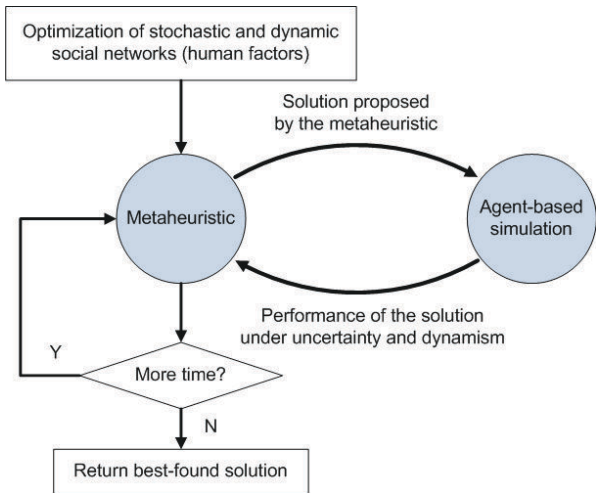


Figure 5: *Agent-based simheuristic framework.*

Accordingly, an open research line in modelling human factors inside large-scale human networks is the one related to exploring the fundamentals and potential applications of agent-based simheuristics (Panadero et al., 2018), where metaheuristic-driven algorithms make use of ABM to account for the uncertainty and dynamism present in these networks. As depicted in Figure 5, given an optimization problem involving human factors in human networks (i.e., a stochastic and dynamic large-scale system), the metaheuristic algorithm acts as an engine which proposes 'promising' solutions (one at a time) to the ABM module. Each of these solutions is then analysed by the ABM component, which provides estimates on the real performance of the proposed solution under the uncertainty and dynamic conditions associated with human factors. The feedback from the ABM module is used by the metaheuristic to guide the search process. This iterative process continues until a time-related ending condition is met. At that point, the best-found solution (or, alternatively, a set of top solutions with different properties) is offered to the decision maker.

Yet another interesting research area in this direction is that of 'learnheuristics' (Calvet et al., 2017), where metaheuristics are combined with machine learning in order to address variations in human behaviour due to changes in the environmental conditions. Thus, for instance, Calvet et al. (2016) propose a hybrid approach combining

metaheuristics with statistical learning in order to account for variations in the willingness to spend of consumers as the 'best-fit' shopping centres have been already assigned to other costumers.

6 Conclusions

This paper reviews how human behaviour in human networks is included in two of the most popular OR/IE techniques: simulation and optimization. The paper comprises an extended survey of related works in different types of human networks: supply chain management, public policies in emergency situations, and Internet-based human networks. Based on the literature review, different techniques typically employed to model human behaviour and social interactions are identified. Furthermore, the main research issues when modelling behavioural traits are depicted, including: cooperation among workers, workload balance, workers' initiative and autonomy, ergonomic conditions at work, personal attitudes, horizontal cooperation among carriers, disease propagation and dynamics, efficient and real-time communication during evacuations, crowd flow patterns, human interaction during evacuation of buildings, movements of vehicles and pedestrians under emergencies, viral marketing campaigns, pricing policies, discovery of network communities, identification of market target groups, information propagation, etc. Likewise, the pros and cons of each modelling technique have been highlighted. Thus, while agent-based simulation is the preferred methodology to modelling network systems dynamics and uncertainty, it is not a valid tool for optimization purposes. At the same time, metaheuristics are well suited to optimize large-scale human networks. However, they show severe limitations when human factors need to be fully considered. Accordingly, the paper argues in favour of hybridizing both techniques. One of these combinations is the so called 'agent-based simheuristics' approach. This integrated methodology benefits from the extraordinary capacity of metaheuristics to generate 'promising' solutions to large-scale combinatorial optimization problems. At the same time, stochastic and dynamic conditions that characterize human behaviour and social interaction can also be taken into account without compromising the resolvability of the corresponding optimization problem.

Acknowledgements

This work has been partially supported by the Erasmus+ (2018-1-ES01-KA103-049767) and the Jose Castillejo programs (CAS16/00201).

References

Abbassi, Z., Bhaskara, A. and Misra, V. (2015). Optimizing display advertising in online social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1–11.

Abu Naser, A. M. and Alshattnawi, S. (2014). An artificial bee colony (abc) algorithm for efficient partitioning of social networks. *International Journal of Intelligent Information Technologies*, 10, 24–39.

Ahmed, S. and Ezeife, C. I. (2013). Discovering influential nodes from trust network. In *Proceedings of the ACM Symposium on Applied Computing*, pp. 121–128.

Akhlaghpour, H., Ghodsi, M., Haghpanah, N., Mirrokni, V. S., Mahini, H. and Nikzad, A. (2010). Optimal iterative pricing over social networks. In *Internet and Network Economics: 6th International Workshop, Proceedings*, pp. 415–423.

Almagooshi, S. (2015). Simulation modelling in healthcare: Challenges and trends. *Procedia Manufacturing*, 3, 301 – 307. 6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences.

Alt, J. K. and Lieberman, S. (2010). Representing dynamic social networks in discrete event social simulation. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 1478–1491.

Andradóttir, S. (2006). An overview of simulation optimization via random search. *Handbooks in operations research and management science 13*, 617–631.

Baines, T. S., Asch, R., Hadfield, L., Mason, J. P., Fletcher, S. and Kay, J. M. (2004). Towards a theoretical framework for human performance modelling within manufacturing systems design. *Simulation Modelling Practice and Theory*, 13, 486–504.

Bakhshandeh, R., Samadi, M., Azimifar, Z. and Schaeffer, J. (2011). Degrees of separation in social networks. In *Proceedings of the 4th Annual Symposium on Combinatorial Search*, pp. 18–23.

Bandini, S., Manzoni, S. and Vizzari, G. (2009). Agent based modeling and simulation: An informatics perspective. *Journal of Artificial Societies and Social Simulation*, 12.

Ben-Zwi, O., Hermelin, D., Lokshtanov, D. and Newman, I. (2009). An exact almost optimal algorithm for target set selection in social networks. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, pp. 355–362.

Bendoly, E., Donohue, K. and Schultz, K. L. (2006). Behavior in operations management: assessing recent findings and revisiting old assumptions. *Journal of Operations Management*, 24, 737–752.

Benjelloun, A. and Crainic, T. G. (2009). Trends, challenges, and perspectives in city logistics. In *Proceedings of the Transportation and Land Use Interaction Conference*, Number 4, pp. 269–284.

Bisset, K., Feng, X., Marathe, M. and Yardi, S. (2009). Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In *Proceedings of the 2009 Winter Simulation Conference*, pp. 2020–2031.

Biswas, A., Gupta, P., Modi, M. and Biswas, B. (2015). An empirical study of some particle swarm optimizer variants for community detection. *Advances in Intelligent Systems and Computing*, 320, 511–520.

Blanco-Moreno, D., Fuentes-Fernández, R. and Pavón, J. (2011). Simulation of online social networks with krowdix. In *International Conference on Computational Aspects of Social Networks*, pp. 13–18.

Borodin, A., Filmus, Y. and Oren, J. (2010). Threshold models for competitive influence in social networks. In *Proceedings of the 6th International Conference on Internet and Network Economics*, pp. 539–550.

Boussier, J. M., Cucu, T., Ion, L., Estrailler, P. and Breuil, D. (2009). Goods distribution with electric vans in cities: towards and agent-based simulation. *World Electric Vehicle Journal*, 3, 1–9.

Brailsford, S. C. (2016). Healthcare: Human behavior in simulation models. In *Behavioral Operational Research: Theory, Methodology and Practice*, pp. 263–280. Palgrave Macmillan.

Bryant, J. (2003). *The Six Dilemmas of Collaboration: Inter-organisational Relationships as Drama* (1st ed.). New York, USA: Wiley.

Bryant, J. (2004). Drama theory as the behavioural rationale in agent-based models. In *IMA International Conference on Analysing Conflict and its Resolution*, pp. 99–102.

Cabrera, G., Juan, A. A., Lázaro, D., Marquès, J. M. and Proskurnia, I. (2014). A simulation-optimization approach to deploy internet services in large-scale systems with user-provided resources. *Simulation*, 90, 644–659.

Cai, Q., Gong, M., Ma, L., Ruan, S., Yuan, F. and Jiao, L. (2015). Greedy discrete particle swarm optimization for large-scale social network clustering. *Information Sciences*, 316, 503–516.

Cai, Q., Gong, M., Shen, B., Ma, L. and Jiao, L. (2014). Discrete particle swarm optimization for identifying community structures in signed social networks. *Neural Networks*, 58, 4–13.

Calvet, L., de Armas, J., Masip, D. and Juan, A. A. (2017). Learnheuristics: hybridizing metaheuristics with machine learning for optimization with dynamic inputs. *Open Mathematics*, 15, 261–280.

Calvet, L., Ferrer, A., Gomes, I., Juan, A. A. and Masip, D. (2016). Combining statistical learning with metaheuristics for the multi-depot vehicle routing problem with market segmentation. *Computers and Industrial Engineering*, 94, 93–104.

Candogan, O., Bimpikis, K. and Ozdaglar, A. (2010). Optimal pricing in the presence of local network effects. In *Proceedings of the 6th International Conference on Internet and Network Economics*, pp. 118–132.

Carnes, T., Nagarajan, C., Wild, S. M. and van Zuylen, A. (2007). Maximizing influence in a competitive social network: A follower's perspective. In *Proceedings of the 9th International Conference on Electronic Commerce*, pp. 351–360.

Chan, F. T. S. and Zhang, T. (2011). The impact of collaborative transportation management on supply chain performance: A simulation approach. *Expert Systems with Applications*, 38, 2319–2329.

Chang, H., Feng, Z. and Ren, Z. (2013). Community detection using ant colony optimization. pp. 3072–3078.

Chau, M., Fu, M. C., Qu, H. and Ryzhov, I. O. (2014). Simulation optimization: a tutorial overview and recent developments in gradient-based methods. In *Proceedings of the Winter Simulation Conference 2014*, pp. 21–35. IEEE.

Chen, W., Lu, P., Sun, X., Wang, Y. and Zhu, Z. A. (2010). Pricing in social networks: Equilibrium and revenue maximization. *CoRR abs/1007.1501*.

Chen, W., Wang, Y. and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 199–207.

Chen, Y. and Qiu, X. (2013). Detecting community structures in social networks with particle swarm optimization. *Communications in Computer and Information Science*, 401, 266–275.

Chu, M. and Law, K. (2013). Computational framework incorporating human behaviors for egress simulations. *Journal of Computing in Civil Engineering*, 27, 699–707.

Chu, M. L., Parigi, P., Latombe, J.-C. and Law, K. H. (2015a). Simulating effects of signage, groups, and crowds on emergent evacuation patterns. *AI Soc.*, 30, 493–507.

Chu, M. L., Parigi, P., Law, K. H. and Latombe, J.-C. (2015b). Simulating individual, group, and crowd behaviors in building egress. *Simulation*, 91, 825–845.

Crainic, T. G., Ricciardi, N. and Storchi, G. (2009). Models for evaluating and planning city logistics systems. *Transportation Science*, 43, 432–454.

Crespo Pereira, D., del Rio Vilas, D., Rios Prado, R. and Lamas Rodriguez, A. (2011). Experimental manufacturing system for research and training on human-centred simulation. In *The 23rd European Modeling and Simulation Symposium*, pp. 400–409.

Derigs, U. (2009). *Optimization and Operations Research*, Volume 2. Oxford, UK: EOLSS Publishers Co Ltd.

Digiesi, S., Kock, A. A., Mummolo, G. and Rooda, J. E. (2009). The effect of dynamic worker behavior on flow line performance. *International Journal of Production Economics*, 120, 368–377.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 57–66.

Duin, R., van Kolck, A., Anand, N., Tavasszy, L. A. and Taniguchi, E. (2011). Towards and agent-based modelling approach for the evaluation of dynamic usage of urban distribution centres. In *Proceedings of the 7th International Conference on City Logistics*.

Elkosantini, S. (2015). Toward a new generic behavior model for human centered system simulation. *Simulation Modelling Practice and Theory*, 52, 108–122.

Elkosantini, S. and Gien, D. (2009). Integration of human behavioural aspects in a dynamic model for a manufacturing system. *International Journal of Production Research*, 47, 2601–2623.

Eskandarpour, M., Ouelhadj, D., Hatami, S., Juan, A. A. and Khosravi, B. (2019). Enhanced multi-directional local search for the bi-objective heterogeneous vehicle routing problem with multiple driving ranges. *European Journal of Operational Research*.

Even-Dar, E. and Shapira, A. (2007). A note on maximizing the spread of influence in social networks. In X. Deng and F. C. Graham (Eds.), *Internet and Network Economics: Third International Workshop, Proceedings*, pp. 281–286.

Ferone, D., Gruler, A., Festa, P. and Juan, A. A. (2018). Enhancing and extending the classical grasp framework with biased randomisation and simulation. *Journal of the Operational Research Society*, 1–14.

Figueira, G. and Almada-Lobo, B. (2014). Hybrid simulation–optimization methods: A taxonomy and discussion. *Simulation Modelling Practice and Theory*, 46, 118–134.

Flake, G. W., Lawrence, S. and Giles, C. L. (2000). Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–160.

Flake, G. W., Tarjan, R. E. and Tsioutsiouliklis, K. (2003). Graph clustering and minimum cut trees. *Internet Mathematics*, 1, 385–408.

Fu, H., Liu, N., Liang, J., Pel, A. J. and Hoogendoorn, S. P. (2015). Modeling and simulation of evacuation route choice behavior using fuzzy set theory. In *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), 2015*, pp. 1327–1332.

Fu, M. C., Glover, F. W. and April, J. (2005). Simulation optimization: a review, new developments, and applications. In *Proceedings of the Winter Simulation Conference, 2005.*, pp. 13–pp. IEEE.

Gach, O. and Hao, J.-K. (2012). A memetic algorithm for community detection in complex networks. In *Lecture Notes in Computer Science*, Volume 7492, pp. 327–336.

Geary, S., Disney, S. M. and Towill, D. R. (2006). On bullwhip in supply chains-historical review, present practice and expected future impact. *International Journal of Production Economics*, 101, 2–18.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821–7826.

Gonzalez-Martin, S., Juan, A. A., Riera, D., Elizondo, M. G. and Ramos, J. J. (2018). A simheuristic algorithm for solving the arc routing problem with stochastic demands. *Journal of Simulation*, 12, 53–66.

Grosse, E. H., Glock, C. H., Jaber, M. Y. and Neumann, W. P. (2015). Incorporating human factors in order picking planning models: framework and research opportunities. *International Journal of Production Research*, 53, 695–717.

Hartline, J., Mirrokni, V. and Sundararajan, M. (2008). Optimal marketing strategies over social networks. In *Proceedings of the 17th International Conference on World Wide Web*, pp. 189–198.

Heath, B. L. and Hill, R. R. (2010). Some insights into the emergence of agent-based modellig. *Journal of Simulation*, 4, 163–169.

Hou, B., Yao, Y., Wang, B. and Liao, D. (2013). Modeling and simulation of large-scale social networks using parallel discrete event simulation. *Simulation*, 89, 1173–1183.

Huang, Y.-T., Lin, K.-H. and Wu, B. Y. (2011). A structural approach for finding real-friend links in internet social networks. In *Proceedings - 2011 IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, iThings/CPSCom 2011*, pp. 305–312.

Huerta, M. A., Fernandez, B. and Koutanoglu, E. (2007). Manufacturing multiagent system for scheduling optimization of production tasks using dynamic genetic algorithms. In *IEEE International Symposium on Assembly and Manufacturing*, pp. 245–250.

Javadi, S. H. S., Khadivi, S., Shiri, M. E. and Xu, J. (2014). An ant colony optimization method to detect communities in social networks. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 200–203.

Jian, N. and Henderson, S. G. (2015). An introduction to simulation optimization. In *2015 Winter Simulation Conference (WSC)*, pp. 1780–1794. IEEE.

Jin, D., Liu, D., Yang, B., Baquero, C. and He, D. (2011). Ant colony optimization with markov random walk for community detection in graphs. In *Lecture Notes in Computer Science*, Volume 6635, pp. 123–134.

Jin, D., Liu, D., Yang, B., Liu, J. and He, D. (2011). Ant colony optimization with a new random walk model for community detection in complex networks. *Advances in Complex Systems*, 14, 795–815.

Jin, D., Yang, B., Liu, J., Liu, D.-Y. and He, D.-X. (2012). Ant colony optimization based on random walk for community detection in complex networks. *Ruan Jian Xue Bao/Journal of Software*, 23, 451–464.

Jin, L., Chen, Y., Wang, T., Hui, P. and Vasilakos, A. V. (2013). Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine*, 51, 144–150.

Juan, A. A., Kelton, W. D., Currie, C. S. M. and Faulin, J. (2018). Simheuristics applications: dealing with uncertainty in logistics, transportation, and other supply chain areas. In *Proceedings of the 2018 Winter Simulation Conference*, pp. 3048–3059. IEEE Press.

Juan, A. A., Mendez, C., Faulin, J., de Armas, J. and Grasman, S. (2016). Electric vehicles in logistics and transportation: a survey on emerging environmental, strategic, and operational challenges. *Energies*,9.

Kagaya, S., Uchida, K., Hagiwara, T. and Negishi, A. (2005). An application of multi-agent simulation to traffic behavior for evacuation in earthquake disaster. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 4224–4236.

Kannabe, H., Noto, M., Morizumi, T. and Kinoshita, H. (2012). Agent-based social simulation model that accommodates diversity of human values. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1818–1823.

Kannan, R., Vempala, S. and Vetta, A. (2004). On clusterings: Good, bad and spectral. *J. ACM*, 51, 497–515.

Kasaie, P. and Kelton, W. D. (2013). Simulation optimization for allocation of epidemic-control resources. *IIE Transactions on Healthcare Systems Engineering*, 3, 78–93.

Kempe, D., Kleinberg, J. and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146.

Kempe, D., Kleinberg, J. and Tardos, É. (2005). Influential nodes in a diffusion model for social networks. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung (Eds.), *Automata, Languages and Programming: 32nd International Colloquium, Proceedings*.

Kennedy, W. G. (2010). Modelling human behaviour in agent-based models. In A. J. Heppenstall, A. T. Crooks, L. M. See, and M. Batty (Eds.), *Agent-Based Models of Geographical Systems*, pp. 167–181. Springer.

Khan, M., Jaber, M. Y. and Guiffrida, A. L. (2012). The effect of human factors on the performance of a two level supply chain. *International Journal of Production Research*, 50, 517–533.

Kimura, M., Saito, K., Nakano, R. and Motoda, H. (2010). Extracting influential nodes on a social network for information diffusion. *Data Mining and Knowledge Discovery*, 20, 70–97.

Knoke, D. and Yang, S. (2008). *Social Network Analysis (Quantitative Applications in the Social Sciences)* (2nd ed.). New York, USA: SAGE Publications, Inc.

Laskowski, M., Demianyk, B. C. P., Witt, J., Mukhi, S. N., Friesen, M. R. and McLeod, R. D. (2011). Agent-based modeling of the spread of influenza-like illness in an emergency department: A simulation study. *IEEE Trans. Inf. Technol. Biomed*, 15, 877–889.

Leskovec, J., Lang, K. J. and Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 631–640.

Li, J. and Chan, F. T. S. (2012). The impact of collaborative transportation management on demand disruption of manufacturing supply chains. *International Journal of Production Research*, 50, 5635–5650.

Li, Y., Zhang, X. and Zhang, S. (2011). Multi-agent Simulation System Study on Product Development Process. *Applied Mathematics & Information Sciences*, 5, 155–161.

Liu, Z. (2017). Modeling and simulation for healthcare operations management using high performance computing and agent-based model. *Journal of Computer Science & Technology*, 17.

Liu, Z., Jacques, C., Szyniszewski, S., Guest, J., Schafer, B., Igusa, T. and Mitrani-Reiser, J. (2016). Agent-Based Simulation of Building Evacuation after an Earthquake: Coupling Human Behavior with Structural Response. *Natural Hazards Review*, 17.

Luh, P. B., Wilkie, C. T., Chang, S. C., Marsh, K. L. and Olderman, N. (2012). Modeling and optimization of building emergency evacuation considering blocking effects on crowd movement. *IEEE Transactions on Automation Science and Engineering*, 9, 687–700.

Macal, C. M. and North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4, 151–162.

Mandala, S. R., Kumara, S. R. T., Rao, C. R. and Albert, R. (2013). Clustering social networks using ant colony optimization. *Operational Research*, 13, 47–65.

Nascimento, M. C. V. and Pitsoulis, L. (2013). Community detection by modularity maximization using grasp with path relinking. *Computers and Operations Research*, 40, 3121–3131.

Neumann, W. P. and Medbo, P. (2009). Integrating human factors into discrete event simulations of parallel flow strategies. *Production Planning & Control*, 20, 2–16.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.

Okdinawati, L., Simatupang, T. M. and Sunitiyoso, Y. (2014). A behavioral multi-agent model for collaborative transportation management (ctm). *Proceedings of T-LOG*, 62.

Okuda, Y., Nakamura, Y., Kishi, M., Ishikawa, N., and Hitomi, M. (1999). Simulation of human-oriented production systems considering workers' cooperation. In *8th IEEE International Workshop on Robot and Human Interaction*, pp. 381–386.

Panadero, J., Juan, A. A., Mozos, J. M., Corlu, C. G. and Onggo, B. S. (2018). Agent-based simheuristics: extending simulation-optimization algorithms via distributed and parallel computing. In *Proceedings of the 2018 Winter Simulation Conference*, pp. 869–880. IEEE Press.

Parikh, N., Swarup, S., Stretz, P. E., Rivers, C. M., Lewis, B. L., Marathe, M. V., Eubank, S. G., Barrett, C. L., Lum, K. and Chungbaek, Y. (2013). Modeling human behavior in the aftermath of a hy-

pothetical improvised nuclear detonation. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 949–956.

Pérez-Bernabeu, E., Juan, A. A., Faulin, J. and Barrios, B. B. (2015). Horizontal cooperation in road transportation: a case illustrating savings in distances and greenhouse gas emissions. *International Transactions in Operational Research*, 22, 585–606.

Putnik, G. D., Škulj, G., Vrabič, R., Varela, L. and Butala, P. (2015). Simulation study of large production network robustness in uncertain environment. *CIRP Annals - Manufacturing Technology*, 64, 439–442.

Qu, J. (2014). Fast PSO algorithm for community detection in graph [C]. In *International Conference of Information Science and Management Engineering*, pp. 529–536.

Quintero-Araujo, C. L., Gruler, A., Juan, A. A. and Faulin, J. (2019). Using horizontal cooperation concepts in integrated routing and facility-location decisions. *International Transactions in Operational Research*, 26, 551–576.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 2658–2663.

Renfro, R. S. (2001). *Modeling and analysis of social networks*. Ph. D. thesis.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 61–70.

Rico, F., Salari, E. and Centeno, G. (2007). Emergency departments nurse allocation to face a pandemic influenza outbreak. In *2007 Winter Simulation Conference*, pp. 1292–1298.

Riedel, R., Mueller, E., von der Weth, R. and Pflugradt, N. (2009). Integrating human behaviour into factory simulation- a feasibility study. In *IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 2089–2093.

Rivero, J., Cuadra, D., Calle, J. and Isasi, P. (2011). Using the aco algorithm for path searches in social networks. *Applied Intelligence*, 36, 899–917.

Robinson, S. (2014). *Simulation - The practice of model develoment and use* (2 ed.). London, UK: Palgrave-Macmilan.

Russel, S. and Norvig, P. (2003). *Artificial intelligence: A Modern Approach*. Englewood Cliffs, USA: Prentice-Hall.

Sabater, J. and Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, pp. 475–482.

Sarimveis, H., Patrinos, P., Tarantilis, C. D. and Kiranoudis, C. T. (2008). Dynamic modeling and control of supply chain systems: A review. *Computers & Operations Research*, 35, 3530–3561.

Schultz, K., Schoenherr, T. and Nembhard, D. (2010). An example and a proposal concerning the correlation of worker processing times in parallel tasks. *Management Science*, 56, 176–191.

Sercan, S., Sima, E.-U. and Sule, G.-O. (2009). Community detection using an colony optimization techniques. In *15th International Conference on Soft Computing*.

Shi, Z., Liu, Y. and Liang, J. (2009). PSO-based community detection in complex networks. Volume 3, pp. 114–119.

Siebers, P. O., Aickelin, U. and Menachof, D. (2008). Introduction to multi-agent simulation. In F. Adam and P. Humphreys (Eds.), *Encyclopedia of Decision Making and Decision Support Technologies*, pp. 554–564. Idea Group Publishing.

Siebers, P. O., Macal, C. M., Garnett, J., Buxton, D. and Pidd, M. (2010). Discrete-event simulation is dead, long live agent-based simulation! *Journal of Simulation*, 4, 204–210.

Silva, E., Donauer, M., Azevedo, A., PeÃ§as, P. and Henriques, E. (2013). A case study evaluating the impact of human behavior on a manufacturing process in-line with automatic processes by means of a simulation model. In *IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 145–149.

Silva, P. M. S. and Pinto, L. R. (2010). Emergency medical systems analysis by simulation and optimization. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 2422–2432.

Singh, A. and Singh, Y. N. (2012). Rumor spreading and inoculation of nodes in complex networks. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 675–678.

Song, F., Yang, X. and Du, L. (2010). The development of paramics based metropolitan emergency evacuation transportation simulation system - meetsim. In *ICCTP 2010*, pp. 388–400.

Spier, J. and Kempf, K. (1995). Simulation of emergent behavior in manufacturing systems. In *ASMC 95 Proceedings of the Advanced Semiconductor Manufacturing Conference and Workshop*, pp. 90–94.

Talbi, E.-G. (2006). *Metaheuristics: From Design to Implementation*. New York, USA: Wiley.

Tamagawa, D., Taniguchi, E. and Yamada, T. (2010). Evaluating city logistics measures using a multi-agent model. *Procedia - Social and Behavioral Sciences*, 2, 6002–6012.

Taniguchi, E., Thompson, R. G. and Yamada, T. (2012). Emerging techniques for enhancing the practical application of city logistics models. *Procedia - Social and Behavioral Sciences*, 39, 3–18.

Taniguchi, E., Yamada, T. and Okamoto, M. (2007). Multi-agent modelling for evaluating dynamic vehicle routing and scheduling systems. *Journal of the Eastern Asia Society for Transportation Studies*, 7, 933–948.

Teo, J. S. E., Taniguchi, E. and Qureshi, A. G. (2012). Evaluating city logistics measure in e-commerce with multiagent systems. *Procedia - Social and Behavioral Sciences*, 39, 349–359.

van Duin, J. H. R., van Kolck, A., Anand, N., Tavasszy, L. A. and Taniguchi, E. (2012). Towards an agent-based modelling approach for the evaluation of dynamic usage of urban distribution centres. *Procedia Social and Behavioral Sciences*, 39, 333–348. Seventh International Conference on City Logistics which was held on June 7-9, 2011, Mallorca, Spain.

Vazirani, V. V. (2012). *Approximation Algorithms*. Berlin Heidelberg, Germany: Springer Science & Business Media.

Wang, Z., Zhang, Z., Li, C., Xu, L. and You, C. (2015). Optimal ordering and disposing policies in the presence of an overconfident retailer: A stackelberg game. *Mathematical Problems in Engineering*, 1–12.

Wangapisit, O., Taniguchi, E., Teo, J. S. E. and Qureshi, A. G. (2014). Multi-agent systems modelling for evaluating joint delivery systems. *Procedia - Social and Behavioral Sciences*, 125, 472–483.

Wen, S., Zhou, W., Zhang, J., Xiang, Y., Zhou, W. and Jia, W. (2013). Modeling propagation dynamics of social network worms. *IEEE Transactions on Parallel and Distributed Systems*, 24, 1633–1643.

Weng, S., Cheng, B., Kwong, S. T., Wang, L. and Chang, C. (2011). Simulation optimization for emergency department resources allocation. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pp. 1231–1238.

Xiao, R. and Yu, T. (2011). A multi-agent simulation approach to rumor spread in virtual community based on social network. *Intelligent Automation & Soft Computing*, 17, 859–869.

Xu, K., Guo, X., Li, J., Lau, R. Y. K. and Liao, S. S. Y. (2012). Discovering target groups in social networking sites: An effective method for maximizing joint influential power. *Electronic Commerce Research and Applications*, 11, 318–334.

Xu, Y., Chen, L. and Zou, S. (2013). Ant colony optimization for detecting communities from bipartite network. *Journal of Software*, 8, 2930–2935.

Yang, Y., Song, L. and Zhang, X. (2007). Organization-oriented simulation of collaborative product development process based on designer's agent model. In *11th International Conference on Computer Supported Cooperative Work in Design*, pp. 309–314.

Yu, M., Ting, S. and Chen, M. (2010). Evaluating the cross-efficiency of information sharing in supply chains. *Expert Systems with Applications*, 37, 2891–2897.

Yuan, C. Y. and Shon, J. Z. (2008). The effects of collaborative transportation management on b2b supply chain inventory and backlog costs: A simulation study. In *IEEE International Conference on Service Operations and Logistics, and Informatics*, Volume 2, pp. 2929–2933.

Zhang, B., Chan, W. and Ukkusuri, S. V. (2009). Agent-based modeling for household level hurricane evacuation. In *Proceedings of the 2009 Winter Simulation Conference*, pp. 2778–2784.

Zhang, C., Hei, X., Yang, D. and Wang, L. (2016). A memetic particle swarm optimization algorithm for community detection in complex networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 30.

Zhang, X., Qiu, J., Zhao, D. and Schlick, C. M. (2015). A human-oriented simulation approach for labor assignment flexibility in changeover processes of manufacturing cells. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25, 740–757.

Zhang, Y., Wu, Z., Chen, H., Sheng, H. and Ma, J. (2008). Mining target marketing groups from users' web of trust on epinions. In *AAAI Spring Symposium - Technical Report*, pp. 116–121.

Zhou, X., Liu, Y., Zhang, J., Liu, T. and Zhang, D. (2015). An ant colony based algorithm for overlapping community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 427, 289–301.

SORT 43 (2) July-December 2019, 223-236 DOI: 10.2436/20.8080.02.86

Tail risk measures using flexible parametric distributions

José María Sarabia[1], Montserrat Guillen[2], Helena Chuliá[2,*] and Faustino Prieto[1]

Abstract

We propose a new type of risk measure for non-negative random variables that focuses on the tail of the distribution. The measure is inspired in general parametric distributions that are well-known in the statistical analysis of the size of income. We derive simple expressions for the conditional moments of these distributions, and we show that they are suitable for analysis of tail risk. The proposed method can easily be implemented in practice because it provides a simple one-step way to compute value-at-risk and tail value-at-risk. We show an illustration with currency exchange data. The data and implementation are open access for reproducibility.

MSC: 60E05, 62P05.

Keywords: Moments, multi-period risk assessment, value-at-risk

1 Introduction

Monitoring risk is one of the most difficult problems in many areas such as finance and insurance. When risk changes dynamically there is no guarantee that the distribution remains stable over time, for instance even if the same family of distributions can be assumed, there may be a drift and, moreover, dispersion may change. When the deviation from the mean is not constant over time, then we encounter the well-known concept of changing volatility.

We propose new risk measures that concentrate on the far-end tail of the distribution. We show that these new measures, under suitable mild regularity conditions, can be implemented easily because they have simple analytical (or numerical) expressions. This characteristic makes them suitable for monitoring risk, when a direct method is needed with the same protocol along time.

* Corresponding author: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain, Tel.: +34934021010 Fax: +34934021821. E-mail address: hchulia@ub.edu and Faustino Prieto. Department of Economics, University of Cantabria, 39005 Santander, Spain

[1] Department of Economics, University of Cantabria, 39005 Santander, Spain

[2] Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain

Received: September 2018

Accepted: May 2019

A rich variety of risk measures can be calculated under the approach that is presented here, when we consider flexible distributions for non-negative random variables. Our work is inspired by the analysis of the size of income distributions, for which there is a long tradition in economics. However, our main contribution is that we find straightforward formulas for the conditional moments of the distributions. Since we concentrate on estimating the tails of the distributions and we are also concerned about being able to implement these risk measures in practice using fast and direct computation, the simple moment expressions are very convenient, for instance to compute the expectation conditional on the variable exceeding the value at risk. We believe that these new measures have a large field of application and they offer an interesting new and tractable approach for practitioners.

2 Basic result

Since we aim at analysing the tail of the distribution, our first result is about moments and, in particular, on higher order moments beyond a certain value. Our interest on moments implies that we study the expectation of the transformation of a random variable through a power function and, just like it is done in conditional tail expectation, we condition on the domain beyond a certain level.

Theorem 1 *Let X be a non-negative and continuous random variable with PDF $f(x)$, CDF $F(x)$, and we assume that $E[X^r]$ is finite for some value $r > 0$. Let us denote by $F_{(r)}$ the CDF of the rth incomplete moments, that is, $F_{(r)}(x) = \frac{\int_0^x z^r dF(z)}{E[X^r]}$, e.g. defined in Kleiber and Kotz (2003). Then, if $t > 0$ we have,*

$$E[X^r|X > t] = E[X^r] \cdot \frac{1 - F_{(r)}(t)}{1 - F(t)}. \tag{1}$$

In particular, when $t = x_\alpha$ denotes the α quantile of X, that is $\Pr(X \leq x_\alpha) = F(x_\alpha) = \alpha$, formula (1) is then,

$$E[X^r|X > x_\alpha] = E[X^r] \cdot \frac{1 - F_{(r)}(x_\alpha)}{1 - \alpha}. \tag{2}$$

Proof: The result follows directly from the definition of incomplete moments given above and standard properties of the cumulative distribution function. ■

The interest of the previous result is that conditional tail higher-order moments can be easily derived if the assumed distribution has simple expressions for the (unconditional) moments, $E(X^r)$, and for the CDF of the rth incomplete moment. As we will see below, there are some distributions for which these expressions can easily be found.

3 McDonald's model

McDonald (1984) analysed distributions for the size of income and found a comprehensive framework that allows a straightforward estimation of parameters and additional features of many distributions for non-negative random variables. The generalized gamma (GG) distribution was proposed by Stacy (1962), while the generalized beta of the first kind (GB1) and the generalized beta of the second kind (GB2), sometimes termed Generalized Beta Prime, were proposed in this context by McDonald (1984) and they are defined in terms of their probability density functions ($a,b,p,q>0$) as follows:

$$f_{GG}(x;a,p,b) = \frac{ax^{ap-1}\exp(-(x/b)^a)}{b^{ap}\Gamma(p)},\ x>0, \tag{3}$$

$$f_{GB1}(x;a,p,q,b) = \frac{ax^{ap-1}[1-(x/b)^a]^{q-1}}{b^{ap}B(p,q)},\ 0\le x\le b, \tag{4}$$

$$f_{GB2}(x;a,p,q,b) = \frac{ax^{ap-1}}{b^{ap}B(p,q)[1+(x/b)^a]^{p+q}},\ x\ge 0, \tag{5}$$

and 0 otherwise. Here $\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\exp(-t)dt$ represents the gamma function and $B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt$ the beta function, where $\alpha,p,q>0$. Note that the parameter b is a scale parameter.

A random variable X with PDF (3)-(5) will be denoted by $X\sim GG(a,p,b)$, $X\sim GB1(a,p,q,b)$ and $X\sim GB2(a,p,q,b)$ respectively. These models include an important number of income distributions. As such, they have been wildly used in many applications. Here we present a few simple examples:

- The generalized gamma (**GG**) distribution includes: the exponential distribution ($a=p=1$), the classical gamma distribution ($a=1$); if $a=1$ and $p=n/2$, a chi-squared distribution with n degrees of freedom is obtained, the classical Weibull distribution ($p=1$), the half normal distribution ($a=2$ and $p=1/2$). Moreover, the two-parameter lognormal distribution is a limiting case of the generalized gamma distribution given by $a\to 0$, $p,b\to\infty$, $a^2\to\sigma^{-2}$ and $bp^{1/a}\to\mu$.

- The **GB1** distribution includes the three-parameter classical beta distribution with support $(0,b)$ if we set $a=1$ in (4). When letting $a=b=1$, we obtain the usual classical beta distribution of the first kind. Chapter 25 of the book of Johnson, Kotz, and Balakrishnan (1995) contains a careful study of beta distributions. See also Balakrishnan and Nevzorov (2004), Chapter 16.

- The **GB2** includes the usual second kind beta distribution ($a=1$), the Singh-Maddala distribution (Singh et al. (1976)) ($p=1$), the Dagum distribution (Dagum (1977)) ($q=1$), the Lomax or Pareto II distribution ($a=p=1$) and the Fisk or

log-logistic distribution ($p = q = 1$). The GB2 distribution was referred to as a Feller-Pareto distribution by Arnold (1983), including an additional location parameter.

One of the main advantages of the McDonald's family is the huge variety of particular or limiting cases that it contains. Many of the models that are basic in the analysis of size and income can be expressed in this framework. According to McDonald (1984), both of the generalized beta distributions include the generalized gamma as a limiting case.

3.1 Properties of the generalized function for the size distribution of income

In this section we describe several properties of the members of the McDonald family, which will be used in the rest of the paper. We aim at finding those characteristics that are useful to describe the tails, as we are mainly concentrated on measuring the risk.

In order to obtain the CDF of the GG distribution, we consider the incomplete gamma function ratio defined by,

$$G(x;\nu) = \frac{1}{\Gamma(\nu)} \int_0^x t^{\nu-1} \exp(-t) dt, \;\; x > 0, \tag{6}$$

with $\nu > 0$. Note that (6) corresponds to the CDF of the classical gamma distribution with shape parameter $\nu > 0$ and scale parameter $b = 1$. As a consequence,

$$x_\alpha = G^{-1}(\alpha;\nu) \tag{7}$$

represents the quantile of order α corresponding to the classical gamma distribution with shape parameter α, scale parameter $b = 1$ and PDF $f(x) = \frac{x^{\nu-1}e^{-x}}{\Gamma(\nu)}$.

Using (6), the CDF of (3) is given by,

$$F_{GG}(x;a,p,b) = G((x/b)^a;p), \;\; x \geq 0. \tag{8}$$

Now, we consider the incomplete beta function ratio defined by,

$$B(x;p,q) = \frac{1}{B(p,q)} \int_0^x t^{p-1}(1-t)^{q-1} dt, \;\; 0 \leq x \leq 1 \tag{9}$$

with $p,q > 0$. Function (9) corresponds to the CDF of the classical beta distribution with PDF $f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$. Therefore,

$$x_\alpha = B^{-1}(\alpha;p,q) \tag{10}$$

represents the quantile of order α of a classical beta distribution with parameter (p,q).

The CDF of the GB1 distribution is:

$$F_{GB1}(x;a,p,q,b) = B((x/b)^a;p,q),\ \ 0 \le x \le b, \tag{11}$$

where $B(\cdot;\cdot,\cdot)$ is defined in (9).

The CDF of the GB2 can be easily defined in terms of the incomplete beta function ratio (9) and their CDF is given by,

$$F_{GB2}(x;a,p,q,b) = B\left(\frac{(x/b)^a}{1+(x/b)^a};p,q\right),\ \ x \ge 0. \tag{12}$$

Butler and McDonald (1989) showed that many inequality measures depend upon the incomplete moments of the income distribution, see also Kleiber and Kotz (2003). They showed that they are easily calculated for a very broad family of distributions because they possess a closure property. This is the case of the GG, GB1 and the GB2 distributions. The CDF, the distribution of the rth incomplete moment $X_{(r)}$ and the moments of the GG, GB1 and GB2 distributions are summarized in Table 1.

Table 1: *The CDF, the distribution of the rth incomplete moment $X_{(r)}$ and the moments of the GG, GB1 and GB2 distributions. For the GB2 distribution $E[X^r]$ and $X_{(r)}$ exist if $q < r/a$.*

Distribution	GG	GB1	GB2
CDF	$G((x/b)^a;p)$	$B((x/b)^a;p,q)$	$B\left(\frac{(x/b)^a}{1+(x/b)^a};p,q\right)$
$X_{(r)}$	$GG(a,p+\frac{r}{a},b)$	$GB1(a,p+\frac{r}{a},q,b)$	$GB2(a,p+\frac{r}{a},q-\frac{r}{a},b)$
$E[X^r]$	$\frac{b^r\Gamma(p+\frac{r}{a})}{\Gamma(p)}$	$\frac{b^rB(p+\frac{r}{a},q)}{B(p,q)}$	$\frac{b^rB(p+\frac{r}{a},q-\frac{r}{a})}{B(p,q)}$

Summarized from Butler and McDonald (1989) and Kleiber and Kotz (2003)

3.2 Estimation of the GG, GB1 and GB2

In order to implement the calculation of the tail risk measures for the distributions of the McDonald family, it is necessary to provide a simple way to fit these distributions. These models can be estimated by maximum likelihood but, as already noted by Prentice (1974) among others, maximization can be difficult. Alternatively, moment estimates can be used.

For a given data set, the sample moments should be calculated and then the parameter estimates can be found, solving the expressions for the theoretical moments given in the last row of Table 1. All positive moments exist for the GG and the GB1. It is not the case for the GB2. Estimation by the method of moments up to four implies the existence of moments up to four in the GB2 case, which implies constraints of the parameters space.

3.3 Conditional moments

Using the results of the previous sections, we can obtain simple expressions for the tail moments. These results follow immediately.

3.3.1 Formulation for the GG Distribution

For the GG distribution, the conditional moments in formula (2) can be expressed as,

$$E[X^r|X>x_\alpha] = \frac{b^r\Gamma(p+r/a)}{(1-\alpha)\Gamma(p)}\cdot\left\{1-G\left(\left(\frac{x_\alpha}{b}\right)^a;p+\frac{r}{a}\right)\right\}, \tag{13}$$

where the quantile, also called value at risk (VaR) is

$$x_\alpha = b\cdot\left\{G^{-1}(\alpha;p)\right\}^{1/a}.$$

3.3.2 Formulation for the GB1 Distribution

For the GB1 distribution, the conditional moments in formula (2) are expressed as follows:

$$\begin{aligned} E[X^r|X>x_\alpha] &= \frac{b^rB(p+r/a,q)}{(1-\alpha)B(p,q)}\cdot\left\{1-B\left(\left(\frac{x_\alpha}{b}\right)^a;p+\frac{r}{a},q\right)\right\}, \\ &= \frac{b^r\Gamma(p+r/a)\Gamma(p+q)}{(1-\alpha)\Gamma(p+q+r/a)\Gamma(p)}\cdot\left\{1-B\left(\left(\frac{x_\alpha}{b}\right)^a;p+\frac{r}{a},q\right)\right\}, \end{aligned} \tag{14}$$

where

$$x_\alpha = b\cdot\left\{B^{-1}(\alpha;p,q)\right\}^{1/a}.$$

3.3.3 Formulation for the GB2 Distribution

For the GB2 distribution, formula (2) gives the following expression for the conditional moments,

$$\begin{aligned} E[X^r|X>x_\alpha] &= \frac{b^rB(p+r/a,q-r/a)}{(1-\alpha)B(p,q)}\cdot\left\{1-B\left(\frac{(x_\alpha/b)^a}{1+(x_\alpha/b)^a};p+\frac{r}{a},q-\frac{r}{a}\right)\right\}, \\ &= \frac{b^r\Gamma(p+r/a)\Gamma(q-r/a)}{(1-\alpha)\Gamma(p)\Gamma(q)}\cdot\left\{1-B\left(\frac{(x_\alpha/b)^a}{1+(x_\alpha/b)^a};p+\frac{r}{a},q-\frac{r}{a}\right)\right\}, \end{aligned} \tag{15}$$

if $q>r/a$ where

$$x_\alpha = b\cdot\left\{\frac{B^{-1}(\alpha;p,q)}{1-B^{-1}(\alpha;p,q)}\right\}^{1/a}.$$

4 Tail risk measures

One of the advantages of having obtained the expressions in the previous section is that it is straightforward to define tail risk measures. This means that we concentrate on the part of the distribution that exceeds a certain level, for instance a certain quantile. In fact, the expected shortfall is one of the easiest forms of tail risk measure, because in plain words, it measures the expected loss beyond a given quantile level and, as such, is only concerned about the size of losses in the worst-case part of the domain.

The different risk measures are given by,

$$E[X|X>x_\alpha] = m, \tag{16}$$

$$var[X|X>x_\alpha] = E[(X-m)^2|X>x_\alpha], \tag{17}$$

$$\gamma_1[X|X>x_\alpha] = \frac{E[(X-m)^3|X>x_\alpha]}{\{var[X|X>x_\alpha]\}^{3/2}}, \tag{18}$$

$$\gamma_2[X|X>x_\alpha] = \frac{E[(X-m)^4|X>x_\alpha]}{\{var[X|X>x_\alpha]\}^2} - 3. \tag{19}$$

These tail risk measures can be written in terms of the tail moments

$$m_r = E[X^r|X>x_\alpha],\ \ r=1,2,\ldots$$

as ($m_1 = m$),

$$var[X|X>x_\alpha] = m_2 - m^2, \tag{20}$$

$$\gamma_1[X|X>x_\alpha] = \frac{m_3 - 3\cdot m\cdot m_2 + 2\cdot m^3}{\{m_2-m^2\}^{3/2}}, \tag{21}$$

$$\gamma_2[X|X>x_\alpha] = \frac{m_4 - 4\cdot m\cdot m_3 + 6\cdot m^2\cdot m_2 - 3\cdot m^4}{\{m_2-m^2\}^2} - 3. \tag{22}$$

Note that the notion of tail value at risk (TVaR) corresponds to m_1. Risk measures other than the value at risk and the tail value at risk, such as GlueVaR proposed by Belles-Sampera, Guillén, and Santolino (2014) can also be calculated. Guillen, Prieto, and Sarabia (2011) analysed risk measures in tails that have a Pareto shape and Generalized beta-generated distributions were studied in Alexander et al. (2012).

5 Case study: tail measures in currency exchange series

Series of daily currency exchange are considered. An example using data from currency exchanges is suitable because exchanges always take a positive value. Three currency exchanges were selected: Australian to US dollars, US dollar to British pound sterling and US dollar to Yen. We only show here the results for the US dollar to British pound

sterling with a series ranging from January 1971 to July 2014. We have selected this particular time frame because it corresponds to a long interval covering several periods of crisis, and thus serves as a good illustration. The other exchange rates lead to similar conclusions, with the exception of the location in time of the periods of high risk, which do not necessarily coincide with those between US dollar and British pound. Results for the other currencies together with the R implementation can be obtained from the authors. Figure 1 displays the raw data for daily exchange rate series and Table 2 presents some summary statistics.

Table 2: *Descriptive summary of the observed exchange rate between the US dollar and the British pound from 1970 to 2014.*

Observed USD/GBP exchange	(N = 10921)
min	1.052
max	2.644
median	(IQR) 1.67 (1.56, 1.91)
mean	(95% CI) 1.77 (1.76, 1.77)
second moment	3.22
third moment	6.05
fourth moment	11.78

The second, third and fourth moments for the whole observed period (from 1970 to 2014) do not necessarily reflect the relative size with respect to the first moment at every window.

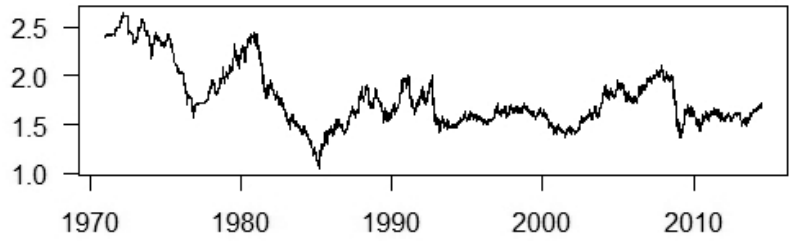


Figure 1: *US dollar/British pound exchange rate from 1971 to 2014.*

The fourth moment is much larger than the first, even if exchange rates means are smaller than 2, which implies that the importance of the fourth moment in the minimization procedure is the largest. A weighted method of moments that gives roughly the same order of magnitude to all four moments could be compared with the unweighted method. In the same vein, more recent observations could be weighted more that distant past observations in a rolling window. There are many possibilities on how to construct such weights and there is not a consensus in finance about this. We have preferred to leave this point as an open question for further research.

A rolling window is implemented, so that the tail risk is calculated using a window of 250 observations. Each new window drops the first observation and adds a new one at the end of the 250 observation days. In this way, a long daily series of tail risk measures can be obtained, using in each case a window of 250 days.

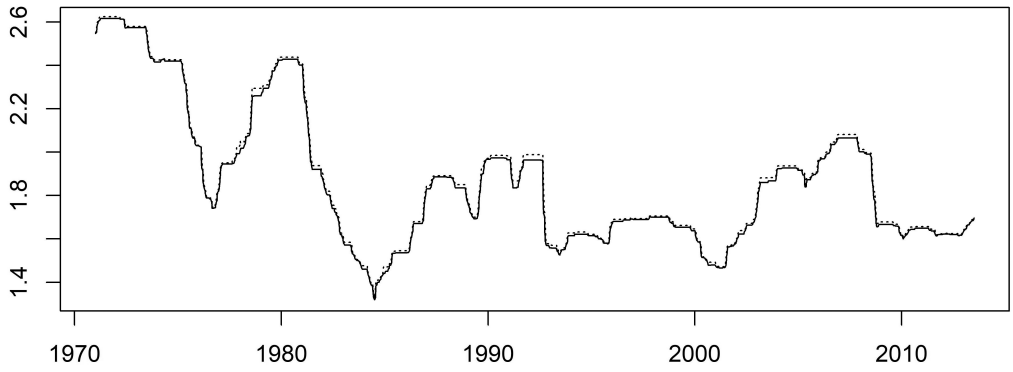


Figure 2: *Empirical estimates of 95% value at risk (solid line) and tail value at risk (dashed line) for the exchange rate between the US dollar and the British pound from 1971 to 2014, in 250-days windows.*

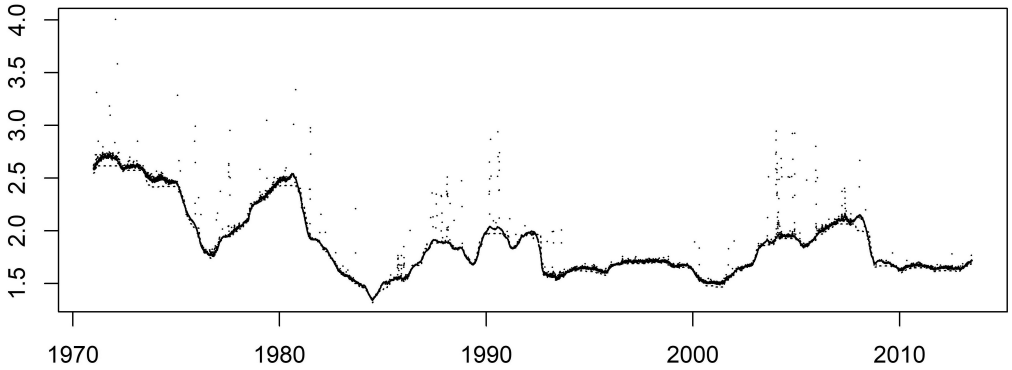


Figure 3: *McDonald model (GG) estimates of 95% value at risk for exchange rates between the US dollar and British pound from 1971 to 2014, in 250-days windows (dot points). Empirical estimates are presented in dashed lines.*

Our optimization method is based on minimizing the Euclidean distance between the theoretical moments and the empirical moments, where we always checked that distance was close enough to zero, less than 0.001. We have also compared parametric estimates versus empirical estimates as suggested by McDonald and Ransom (1979) . We always achieved convergence in our examples. However, as suggested by one of the reviewers, a useful recommendation when implementing this kind of optimization in a rolling window is to take the result of parameter estimation (in the previous window) as the seed in numerical optimization in the following one.

Figure 2 shows the results and compares the tail analysis for a 95% value at risk (VaR) and the 95% tail value at risk when using an Empirical CDF. Figures 3 and 4 present the analysis of the 95% VaR and the 95% TVaR of the McDonald generalized gamma model for the exchange rate, respectively. To save space, we only present the

graphical results for the GG distribution but the results (available upon request) are similar when we use de GB1 and GB2 distributions. The conclusion is that the proposed model is able to capture fluctuations of the risk in the exchange rate that the empirical analysis cannot capture. Note that the spikes in specially risky days are spotted much better with our method. There are periods of high risk around 1973 (oil crisis), 1985 (international intervention in the currency markets to depreciate the dollar), 1987 (market crash), 2004 (dot-com bubble) and 2008 (Lehman Brothers and global financial crisis). If the empirical conditional distribution function was used, the tail risk would have been substantially underestimated. The McDonald approach seems to provide values that are larger than those provided by the empirical approach and they seem to be much more sensitive to daily updates in the rolling window.

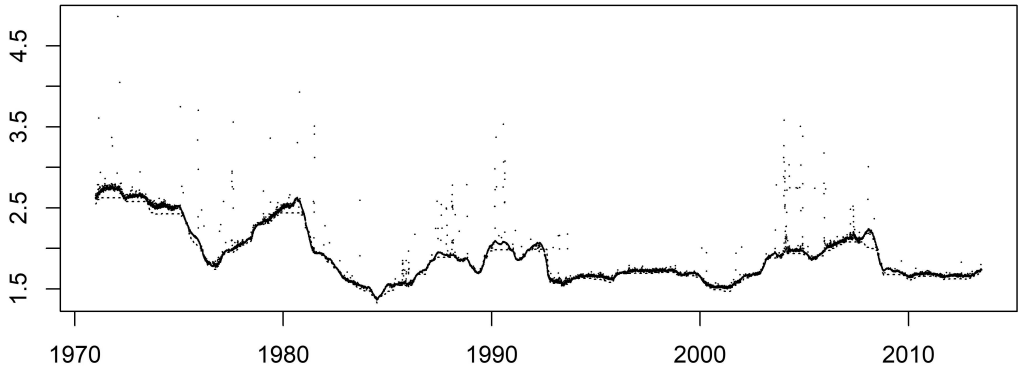


Figure 4: *McDonald model (GG) estimates of 95% tail value at risk for exchange rates between the US dollar and the British pound from 1971 to 2014, in 250-days windows (dot points). Empirical estimates are presented in dashed lines.*

Table 3 presents some summary statistics of the empirical and the estimated McDonald (GG, GB1 and GB2) 95% value at risk and tail value at risk in 250-days windows from 1971 to 2014 of the exchange rate between the US dollar and the British pound. As expected after inspection of Figures 3 and 4, the summary statistics of the value at risk and the tail value at risk are higher when using the GG distribution than when using the empirical CDF.

Figures 3 and 4 offer the comparative analysis for value at risk and tail value at risk to see the parametric estimates versus the empirical. The inferior stability of parametric estimates could indeed speak against the parametric method, but it could also show that fitting a parametric distribution requires to look at the whole domain, making inference about the tail more dependent on the location and shape than empirical risk measures. Empirical estimates of the quantiles differ from estimates based on the parametric fit because they only sort observations and choose the value (or an interpolation of two values) that corresponds to the chosen confidence level, here 95%. If there is a large

extreme suddenly appearing on the right tail, the quantile may not react to that phenomenon. This is the main disadvantage of working with quantiles. When looking at the empirical tail conditional expectation estimates (dashed line in Figure 2), there is only a slight increase of the tail value at risk compared to the value at risk. In contrast, parametric distributions are fitted with all the information in the data.

Table 3: *Summary of the empirical and the estimated McDonald (GG, GB1 and GB2) 95% value at risk and tail value at risk in 250-days windows from 1971 to 2014 of the exchange rate between the US dollar and the British pound.*

	Daily exchange rate USD/GBP (N = 10,671)			
		McDonald models		
	Empirical	GG	GB1	GB2
Value at Risk				
min	1.32	1.34	1.32	1.35
max	2.62	4.00	2.87	3.02
median	1.80	1.80	1.78	1.88
(IQR)	(1.62, 1.99)	(1.65, 2.03)	(1.63, 1.99)	(1.65, 2.08)
mean	1.87	1.89	1.86	1.92
(95% CI)	(1.86, 1.87)	(1.88, 1.90)	(1.85, 1.86)	(1.91, 1.92)
Tail Value at Risk				
min	1.33	1.36	1.23	1.39
max	2.63	4.86	2.96	3.29
median	1.82	1.83	1.79	1.93
(IQR)	(1.64, 2.00)	(1.67, 2.07)	(1.64, 2.01)	(1.68, 2.16)
mean	1.88	1.92	1.87	1.98
(95% CI)	(1.87, 1.88)	(1.91, 1.93)	(1.86, 1.87)	(1.97, 1.98)

In our case study, daily observations correspond to a different random variable, for which we only have exactly one observation. When we deploy a rolling window, our hypothesis is that the distribution remains stable during that window period and that observations are independent. Empirical risk estimates of value at risk and tail value at risk have been extensively used in the literature, knowing that they are very robust. But when analysing risk, and in our approach, we prefer a parametric approach that considers the size of all the observations.

In order to take into consideration sample size issues, we have tried wide windows observations of 500 and 750 daily data. The conclusions did not change. As noted by one of the reviewers, time series characteristics may indeed be interfering in the estimation. Standard errors may be affected by the existence of positive and significant correlation between subsequent daily observations, but our application does not address inference questions.

6 Conclusions

We conclude that the McDonald model is a suitable framework to analyse tail risk and we show that it can easily be implemented with moment estimates, it is fast and it does not require considerable computational effort. The main importance of our proposed approach is about the implementation.

Even if expressions for incomplete moments of income distributions had been analysed before, the focus there was on their link to inequality measures. Our added value here is about the analysis of conditional moments and their relationship with risk measures such as the tail conditional expectation. These exact expressions had not been implemented before. By finding the link between moments, incomplete moments of income distributions and risk measures we facilitate the task of risk analysts.

Tail risk analysis can be done as fast as when the empirical distribution is assumed, because parameter can be fitted using the first moments. Then, tail risk is computed immediately from the expressions presented above.

Butler and McDonald (1989) mentioned that in many fields of applications the entire shape of the distribution, not just its mean, is important and they gave an empirical example where they calculated normalized incomplete moments or moment distributions of the GB1, GB2 and GG in US income data for a series of years. They used maximum likelihood estimation on grouped data. They concluded that these income distribution moments characterize important properties of interest in an analysis of the distribution of economic data (see also Butler and McDonald (1987)). Our practical contribution concentrates on the tail. We provide a moment estimation procedure that is fast in practice, produces a quick answer (through the remark given by Theorem 1) and improves the results of empirical measures.

The proposed methodology is useful in the analysis of financial time series, since it has the capability to detect periods where the risk is high and the results are realistic in the most of cases. However, isolated points can suggest non-stability on parameter estimation.

We have not addressed the question of the relative merits of alternative estimation techniques in this paper. McDonald and Ransom (1979) noted that the techniques of maximum likelihood estimation and method of moments are not directly appropriate for the case in which grouped data is used. As a practical tool, these authors suggest to check the agreement between the implied, i.e. substituting the parameter estimates in the expression for the mathematical expectation, and empirical estimates of the mean. Their main concern is about the fact that they are using grouped rather than individual data. Since we are working on individual observations we believe that both maximum likelihood estimation and the method of moments estimation are suitable. However, when fitting a GG distribution, Prentice (1974) and earlier authors note that maximization of the likelihood function with Newton-Raphson method does not work well and that the existence of solutions to the log-likelihood equations is sometimes in doubt. For the GG distribution, the `flexsurv` R package (Jackson, 2016) could be used.

Acknowledgements

The support received from the Spanish Ministry of Science/FEDER ECO2016-76203-C2-1-P / C2-2-P is acknowledged. MG thanks ICREA Academia. We are grateful for the constructive comments and suggestions provided by the Editor and the reviewers, which have improved the paper.

References

Alexander, C., G. Cordeiro, E. Ortega, and J. Sarabia (2012). Generalized beta-generated distributions. *Computational Statistics and Data Analysis*, 56, 1880–1897.

Arnold, B. (1983). *Pareto Distributions*. International Co-operative Publishing House.

Balakrishnan, N. and V. Nevzorov (2004). *A Primer on Statistical Distributions*. John Wiley & Sons.

Belles-Sampera, J., M. Guillén, and M. Santolino (2014). Beyond value-at-risk: Gluevar distortion risk measures. *Risk Analysis*, 34, 121–134.

Butler, R. and J. McDonald (1987). Interdistributional income inequality. *Journal of Business and Economic Statistics*, 5, 13–18.

Butler, R. and J. McDonald (1989). Using incomplete moments to measure inequality. *Journal of Econometrics*, 42, 109–119.

Dagum, C. (1977). New model of personal income-distribution-specification and estimation. *Economie appliquée*, 30, 413–437.

Guillen, M., F. Prieto, and J. Sarabia (2011). Modelling losses and locating the tail with the pareto positive stable distribution. *Insurance: Mathematics and Economics*, 49, 454–461.

Jackson, C. (2016). flexsurv: a platform for parametric survival modeling in R. *Journal of Statistical Software*, 70.

Johnson, N., S. Kotz, and N. Balakrishnan (1995). Continuous univariate distributions, vol. 2 of wiley series in probability and mathematical statistics: applied probability and statistics.

Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*, Volume 470. John Wiley & Sons.

McDonald, J. (1984). Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society*, 647–663.

McDonald, J. and M. Ransom (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica: Journal of the Econometric Society*, 1513–1525.

Prentice, R. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, 61, 539–544.

Singh, S., G. Maddala, et al. (1976). A function for size distribution of incomes. *Econometrica*, 44, 963–970.

Stacy, E. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33, 1187–1192.

SORT 43 (2) July-December 2019, 237-258 DOI: 10.2436/20.8080.02.87

False discovery rate control for grouped or discretely supported p-values with application to a neuroimaging study

Hien D. Nguyen*, Yohan Yee, Geoffrey J. McLachlan and Jason P. Lerch

Abstract

False discovery rate (FDR) control is important in multiple testing scenarios that are common in neuroimaging experiments, and p-values from such experiments may often arise from some discretely supported distribution or may be grouped in some way. Two situations that may lead to discretely supported distributions are when the p-values arise from Monte Carlo or permutation tests are used. Grouped p-values may occur when p-values are quantized for storage. In the neuroimaging context, grouped p-values may occur when data are stored in an integer-encoded form. We present a method for FDR control that is applicable in cases where only p-values are available for inference, and when those p-values are discretely supported or grouped. We assess our method via a comprehensive set of simulation scenarios and find that our method can outperform commonly used FDR control schemes in various cases. An implementation to a mouse imaging data set is used as an example to demonstrate the applicability of our approach.

MSC: 62-07, 62F03, 62F35, 62N03, 62P10.

Keywords: Censored data, data quantization, discrete support, empirical-Bayes, false discovery rate control, grouped data, incompletely observed data, mixture model.

1 Introduction

Modern experiments in numerous fields of science now output the results of thousands to millions of hypothesis tests simultaneously. Recent accounts of the theoretical aspects of the phenomenon of simultaneous statistical inference with applications in the life sciences can be found in Dickhaus (2014). Further treatment of the topic can be found in Efron (2010).

We assume that we are operating in a scenario whereupon we (only) observe p-values from $n \in \mathbb{N}$ simultaneous tests of the hypotheses H_i ($i \in [n]$; $[n] = \{1, ..., n\}$), which may

* HDN is at the Department of Mathematics and Statistics, La Trobe University, Bundoora 3086, Victoria Australia (Corresponding author; email: h.nguyen5@latrobe.edu.au). GJM is at the School of Mathematics and Physics and Centre for Innovation in Biomedical Imaging Technology, University of Queensland, St. Lucia 4072, Queensland Australia. YY and JPL are at the Mouse Imaging Centre, Hospital for Sick Children, M5T 3H7 Toronto, Ontario Canada.

Received: December 2018
Accepted: April 2019

be either null or otherwise and may be related in some manner. Suppose that we are conducting well-specified standard significance tests at significance level $\alpha \in (0,1)$. If all of the hypotheses are null, then we can directly compute the expected number of tests declared significant as $n\alpha$. Taking n large (e.g. $n \geq 10^6$) and α at usual levels such as $\alpha \in (0.001, 0.1)$, the number of incorrectly declared hypotheses as not null can be greatly inflated. When there is a potential for large numbers of incorrectly rejected hypotheses, the outcome of using only standard significant tests can lead to spurious conclusions.

In recent years, the leading paradigm for the handling of large-scale simultaneous hypothesis testing scenarios is via the control of the false discovery rate (FDR) of an experiment. The control of FDR was first introduced by Benjamini and Hochberg (1995) and has since been developed upon by numerous other authors. The FDR of an experiment can be defined as FDR=$\mathbb{E}(N_{01}/N_R)\,\mathbb{P}(N_R > 0)$, where N_{01} and N_R denote the number of false positives and the number of rejected hypotheses (hypotheses declared significantly alternative) from the experiment, respectively.

The FDR control method of Benjamini and Hochberg (1995) was first developed to only take an input of n IID (identically and independently distributed) p-values. An extension towards the control of FDR in samples of correlated p-values was derived in Benjamini and Yekutieli (2001). Since these key publications, there have been numerous articles written on the topic of FDR control in various settings and under various conditions; see Benjamini (2010) and the comments therein for an account of the history and development of FDR control.

In most FDR control methods, there is an explicit assumption that the marginal distribution of the p-values of an experiment is uniform over the unit interval, if the hypothesis under consideration is null. This assumption arises via the classical theory of p-values of well-specified tests (cf. Dickhaus, 2014, Sect. 2). However, in practice, there are numerous ways for which the distribution of p-values under the null can deviate from uniformity. In Efron (2010, Sect. 6.4), several causes of deviation from uniformity are suggested. Broadly, these are: failed mathematical assumptions (e.g. incorrect use of distribution for computing p-values), correlation between p-values, and unaccounted covariates or misspecification of null hypotheses. A treatment on the effects of misspecification of the null hypotheses due to unaccounted covariates can be found in Barreto and Howland (2006, Chap. 7 Appendix and Chap. 18).

There are some FDR methods that account for deviation from uniformity in the null distribution. These include the methods of Yekutieli and Benjamini (1999), Korn et al. (2004), Pollard and van der Laan (2004), van der Laan and Hubbard (2006), and Habiger and Pena (2011). Unfortunately, the listed methods all require access to the original data of the experiment in order to compute permutation-based test statistics and thus permutation-based p-values. As mentioned previously, access to the original experimental data lies outside of the scope of this article as we only assume knowledge of the p-values. The empirical-Bayes (EB) paradigm provides a powerful framework under which the deviation of the null away from uniformity can be addressed with only access to the experimental p-values. The EB paradigm for FDR control was first introduced in

Efron et al. (2001). A relatively complete account of the EB paradigm appears in Efron (2010).

We largely follow the work of McLachlan, Bean and Ben-Tovim (2006) and Nguyen et al. (2014). Our novelty and development of the available literature is to present a methodology for addressing the problems that are introduced when p-values are distributed on a discrete support or when the p-values are grouped.

As in Nguyen et al. (2014), we particularly focus on the context of neuroimaging applications. In voxel-based morphometric neuroimaging studies (see, e.g., Ashburner and Friston, 2000), the number of simultaneously tested hypotheses often range in the tens of thousands to the tens of millions. Due to such inflated numbers, the risk of making false discoveries is often unacceptably high. Making inference without FDR control in such situations may lead to an overabundance of absurd conclusions. This is well demonstrated in the infamous results of Bennett et al. (2009), where neuronal activation in the brain of a dead fish was observed in a functional magnetic resonance imaging study, where the FDR was not controlled. Thus, FDR control is an important and ongoing area of research in the neuroimaging literature. A classic treatment regarding FDR control in neuroimaging can be found in Genovese, Lazar and Nichols (2002).

Grouped p-values may arise under incomplete observation; that is, under censoring, grouping, or quantization observation of p-values; see Turnbull (1976) for working definitions of the censored and grouped data and Gersho and Gray (1992) for quantization. We shall elaborate upon these definitions in the sequel.

Neuroimaging data such as MRI and functional MRI volumes are usually stored via one of a number of common storage protocols. Incomplete data may arise when data are compressed using one of these storage algorithms. Some common storage protocols under which neuroimaging data may be compressed include ANALYZE (Robb et al., 1989), DICOM Bidgood et al. (1997), MINC (Vincent et al., 2003), and NIFTI Cox et al., 2004). A good summary of these protocols is presented in Larobina and Murino (2014). In the pursuit of reduced storage sizes, it is not uncommon for neuroimaging data volumes to be stored at the minimum precision specification of any of the aforementioned formats. For example, DICOM volumes can only store data as integers, at a precision level as low as 8-bits (i.e. $2^8 = 256$ unique values). When p-values are stored in such a format, the true values are grouped into bins that are centered on a discrete number of possible values on the unit interval.

Discretely supported p-values may arise from Monte Carlo or permutation tests. In such cases, the p-values for a fixed number of permutations or Monte Carlo replications R, can only take on $R+1$ discrete value. Furthermore, Monte Carlo and permutation tests are both random approximations of exact tests. Such tests can again only output a discrete number of possible p-values that depend on the sample size of the data from which they are computed (cf. Phipson and Smyth, 2010). Monte Carlo and permutation tests are frequently used in neuroimaging studies; see, for example, Winkler et al. (2014).

It is known that grouped observations of real numbers can often lead to inaccuracies in statistical computations. Discussions of some aspects regarding the effects of grouping on statistical computation are discussed in Moschitta, Schoukens and Carbone (2015). The effects of quantization can particularly be ruinous when applying standard EB-based FDR control approaches. The effects of incompleteness in the observation of p-values qualifies as a failure in mathematical assumptions, under the taxonomy of Efron (2010, Sect. 6.4).

In this article, we address the problem of EB-based FDR control using p-values that are discretely supported or grouped, via the use of binned estimation. We demonstrate the effect of grouped p-values on the estimation of the EB model. Making use of the EM (expectation–maximization of Dempster, Laird and Rubin (1977) algorithm from the mix function in the mixdist package (MacDonald and Du, 2012) in the R programming language (R Core Team, 2016), we demonstrate that one can simply and rapidly maximum marginal likelihood (MML) estimation (cf. Varin, 2008) of the EB model. We further prove the consistency of the MML estimator for the EB model. A second numerical study is conducted to demonstrate the performance of our method under incomplete observation of p-values, where a comparison between our method is made against the commonly used methods of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), and Storey (2002). An example application to a mouse brain imaging dataset is then provided to demonstrate the usefulness of our approach in a real data scenario.

The article proceeds as follows. In Section 2, we introduce concepts relating to grouped and discretely observed p-values, and the EB model for p-values. We then demonstrate how the EB model can be used for FDR control. In Section 3, we present a demonstration of the effect of grouped p-values on the naive estimation of the EB model. In Section 4, a numerical study of the performance of our method is presented. In Section 5, the methodology is applied to control the FDR of a mouse imaging data set. Conclusions are drawn in Section 6. Further details regarding our methodology are included in the Supplementary Materials.

2 Binned estimation of the empirical Bayes model for grouped or discretely supported p-values

Let $0 = a_0 < a_1 < \cdots < a_{m-1} < a_m = 1$ be a set of m points along the line segment $[0,1]$. Suppose that we observe n p-values $P_i \in [0,1]$, for $i \in [n]$. Grouping may occur when $P_1, \ldots, P_n$ are subject to rounding (or quantization), such that each p-value to the nearest point a_j, for $j \in [m] \cup \{0\}$, where various measurements of closeness may be used for different applications. Observation of $P_1, \ldots, P_n$ may also be grouped when they are censored. That is, when we only observe the fact that each p-value $P_i \in (a_{j-1}, a_j)$, for some $j \in [n]$, and not its precise value. Under either quantization or censoring, the p-values P_i are each mapped to a discrete set of values, either the $m+1$ quantization

centers a_j or the m intervals (a_{j-1}, a_j), enumerated by the index j. In the case where $P_1, \ldots, P_n$ arise from a Monte Carlo or permutation tests, we may envisage that they are quantized approximations of p-values that arise from an asymptotically large population size and thus can be treated in the same manner as quantized p-values in practice.

2.1 The empirical Bayes model

For $i \in [n]$, let $Z_i = \Phi^{-1}(1 - P_i)$ be the probit transformation of P_i. We refer to Z_i as the z-scores. Here Φ is the cumulative distribution function of the standard normal distribution. Under the EB paradigm, we assume that some proportion $\pi_0 \in [0,1]$ of the n hypotheses are null and thus $\pi_1 = 1 - \pi_0$ are otherwise. Since an alternative (not null) hypothesis generates a p-value that is on average smaller than that of a null hypothesis, we can also assume that the z-scores of null hypotheses arise from some distribution with a mean $\mu_0 \in \mathbb{R}$, where $\mu_0 < \mu_1$ and $\mu_1 \in \mathbb{R}$ is the mean of the alternative z-scores. Under uniformity of the p-values, the z-scores have a standard normal distribution, we can approximate the density of the null z-scores by $f_0(z) = \phi\left(z; \mu_0, \sigma_0^2\right)$, where $\sigma_0^2 > 0$ and $\phi\left(\cdot; \mu, \sigma^2\right)$ is the normal density function with mean μ and variance σ^2. Likewise, we can approximate the density of the alternative z-scores by $f_1(z) = \phi\left(z; \mu_1, \sigma_1^2\right)$, where $\sigma_1^2 > 0$ (cf. Efron, 2004). The marginal density of any z-score, can be approximated by the two-component mixture model

$$f(z; \boldsymbol{\theta}) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{1}$$

where $\boldsymbol{\theta}^\top = \left(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\right)$ is the model parameter vector and $(\cdot)^\top$ is the transpose operator. We say that (1) is the EB model for p-values.

2.2 Statistical model for binned data

Let $-\infty = b_0 < b_1 < b_2 < \ldots < b_{m-1} < \infty$ for some $m \in \mathbb{N} \setminus \{1\}$. We define m bins B_j, for $j \in [m]$, where $B_j = (b_{j-1}, b_j]$ for $j \in [m-1]$ and $B_m = (b_{m-1}, \infty)$. Suppose that we observe n p-values P_i that are converted to z-scores Z_i, which may be infinite in value. Further, define $\mathbb{I}(A)$ as the indicator variable that takes value 1 if proposition A is true and 0 otherwise, and define a new random variable $X_i^\top = (X_{i1}, \ldots, X_{im})$, where $X_{ij} = \mathbb{I}(Z_i \in B_j)$, for each i and $j \in [m]$.

Suppose that the n p-values generate z-scores that are potentially correlated and marginally arise from a mixture model of form (1), with $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, for some valid $\boldsymbol{\theta}^0$. Using the bins and realizations $x_i^\top = (x_{i1}, \ldots, x_{im})$ of each X_i ($i \in [n]$), we can write the marginal likelihood and log-marginal likelihood functions under the mixture model approximation for the z-scores as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\int_{B_j} f(z; \boldsymbol{\theta}) \, \mathrm{d}z \right]^{x_{ij}}$$

and

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \log \int_{B_j} f(z;\boldsymbol{\theta})\,\mathrm{d}z. \tag{2}$$

Write the MML estimator for $\boldsymbol{\theta}^0$ that is obtained from n z-scores as $\hat{\boldsymbol{\theta}}_n$. We can define $\hat{\boldsymbol{\theta}}_n$ as a suitable root of the score equation $\nabla l = \mathbf{0}$, where ∇ is the gradient operator and $\mathbf{0}$ is the zero vector.

The marginal likelihood function is simply an approximation to the likelihood that is constructed under an assumption of independence between the observations X_i (cf. Varin, 2008). In light of not knowing what the true dependence structure between the observations is, the marginal likelihood function can be seen as a quasi-likelihood construction in sense of White (1982). The purpose of a quasi-likelihood construction is to make use of an approximation that is close enough to the true data generative process so that meaningful inference can be drawn. Here its use is to avoid the need to declare an explicit model for potential correlation structures between the observations.

The EM algorithm for MML estimation is the context of this article is provided in Supplementary Materials Section 1. The consistency of the MML estimator is also established in the same section.

2.3 Empirical Bayes-based FDR control

Upon estimation of the parameter vector $\boldsymbol{\theta}^0$ via the MML estimator $\hat{\boldsymbol{\theta}}_n$, we can follow the approach of McLachlan et al. (2006) in order to implement EB-based FDR control of the experiment. That is, consider the event $\{H_i \text{ is null } | Z_i = z_i\}$, for each $i \in [n]$. Via Bayes' rule and the MML estimator $\hat{\boldsymbol{\theta}}_n$, we can estimate the probability of the aforementioned event via the expression

$$\hat{\mathbb{P}}(H_i \text{ is null } | Z_i = z_i) = \frac{\hat{\pi}_0 \phi\left(z_i; \hat{\mu}_0, \hat{\sigma}_0^2\right)}{f\left(z_i; \hat{\boldsymbol{\theta}}_n\right)} = \tau\left(z_i; \hat{\boldsymbol{\theta}}\right). \tag{3}$$

Using (3), we can then define the rejection rule

$$r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = \begin{cases} 1, & \text{if } \tau\left(z_i; \hat{\boldsymbol{\theta}}_n\right) \leq c \\ 0, & \text{otherwise,} \end{cases}$$

where $c \in [0,1]$. Here $r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1$ if the null hypothesis of H_i is rejected (i.e. H_i is declared significant) and 0 otherwise.

Let the marginal FDR be defined as $m\text{FDR} = \mathbb{E}N_{01}/\mathbb{E}N_R$. We can estimate the mFDR of an experiment via the expression

$$\widehat{m\text{FDR}} = \frac{\sum_{i=1}^{n} \tau\left(z_i; \hat{\boldsymbol{\theta}}_n\right) \mathbb{I}\left(r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1\right)}{\sum_{i=1}^{n} \mathbb{I}\left(r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1\right)}, \tag{4}$$

which we can prove to converge to the mFDR in probability, under M-dependence (cf. Nguyen et al., 2014, Thm. 1). Subsequently, we can also demonstrate that for large n, the mFDR approaches the FDR (cf. Nguyen et al., 2014, Thm. 2).

Notice that $m\text{FDR} = m\text{FDR}(c)$ is a function of the threshold c. Using the thresholding value, we can approximately control the FDR at any desired level β by setting the threshold c using the rule

$$c_\beta = \arg\max \left\{ c \in [0,1] : \widehat{m\text{FDR}}(c) \leq \beta \right\}. \tag{5}$$

2.4 Choosing the binning scheme

Thus far in discussing the binned estimation of the z-score distribution f, we have assumed that the bin cutoffs $b_1, ..., b_{m-1}$ are predetermined. When the p-values are censored into intervals (a_{j-1}, a_j), for $j \in [m] \cup \{0\}$, as describe at the beginning of the section, we may take the values a_j to inform our bin cutoffs $b_1, \ldots, b_{m-1}$. This can be done by computing the probit transformation of each of the cutoffs. That is, we can set $b_0 = -\infty$, $b_1 = \Phi^{-1}(1 - a_{m-1})$, $b_2 = \Phi^{-1}(1 - a_{m-2}), \ldots, b_{m-1} = \Phi^{-1}(1 - a_1)$. Thus the bin cutoffs are implicitly given by the censoring and thus the problem does not require the user to make a choice regarding the binning scheme, similar to the situation originally encountered in McLachlan and Jones (1988).

When the p-values are quantized or when they are discretely distributed, we must make a non-trivial decision regarding the binning scheme to use. A simple approach to the choice of binning scheme is to use the techniques underlying optimal histogram smoothing on the finite z-scores. In R, there are several optimal histogram smoothing techniques that are deployed in the default hist function. These include the fixed bin width methods of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981).

Under the methods of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981), the number of bins is taken to be $m = \lceil \log_2 n \rceil + 1$,

$$m = \lceil (\text{Range}/h) \rceil \text{ with } h = 2 \times \text{IQR}/n^{1/3},$$

and

$$m = \lceil (\text{Range}/h) \rceil \text{ with } h = 3.5 \times s/n^{1/3},$$

respectively. Here, $\lceil \cdot \rceil$ is the ceiling operator, and Range, IQR, and s are the sample range, interquartile range, and standard deviation, respectively. We compare the effectiveness of each of the binning approaches in the next section.

The binning of data, or the approximation of density functions via histograms, is a nontrivial problem that extends beyond the scope of this article. There is an abundance of methods for data binning that are available within the statistical and machine learning literature. Any of such methods can be used in place of the ones that we have suggested. For example, see the papers of Wand (1997) and Birge and Rozenholc (2006) regarding alternative fixed bin width methods. Examples of variable bin width methods can be found in the works of Kontkanen and Myllymaki (2007) and Denby and Mallows (2009). Further approaches can be found within the references of the cited articles.

3 An integer encoding example

To demonstrate the effects of grouping on p-values, we use the effects of integer encoding of such values as an example. Table 1 of Larobina and Murino (2014) provides a summary of the possible data compression schemes that can be applied when storing data in the ANALYZE, DICOM, MINC, or NIFTI formats. The possible integer storage schemes available for ANALYZE are 8-bits unsigned, or 16 and 32-bits signed. For DICOM, the available schemes are 8, 16, and 32-bits signed or unsigned. For MINC, 8, 16, and 32-bits signed or unsigned, are available. Finally, NIFTI can store data as 8, 16, 32, or 64-bits signed or unsigned.

For reference, 8, 16, 32, and 64 binary bits unsigned can encode 256, 65536, 429496 7296, and 1.84E+19 ($a\text{E}b = a \times 10^b$) unique values, respectively. These numbers are doubled when signed encodings are used. In this article, we only consider integer compression in 8-bits or 16-bits signed and unsigned formats. This is because 32-bits and 64-bits can be used to encode single and double-precision floating points, respectively, which largely mitigate against the reduced precision problems that we discuss in this article.

3.1 Integer encoding of p-values

As noted earlier, we are largely concerned with large scale-hypothesis testing situations that arise from voxel-based experiments (cf. Ashburner and Friston, 2000). In such experiments, a hypothesis test is conducted at each voxel of an imaged volume. For statistical analyses, resulting volumes of p-values are generated. It is these volumes that are then stored, possibly in a reduced precision format, for dissemination or for storage.

Suppose that a γ-bits unsigned integer encoding is used, where $\gamma \in \mathbb{N}$. Note that a γ-bits signed integer encoding is effectively equivalent to a $(\gamma + 1)$-bits unsigned, for all intents and purposes. When the hypothesis testing data are stored as a p-value volume, we suppose that the data are stored such that the smallest integer value encodes the

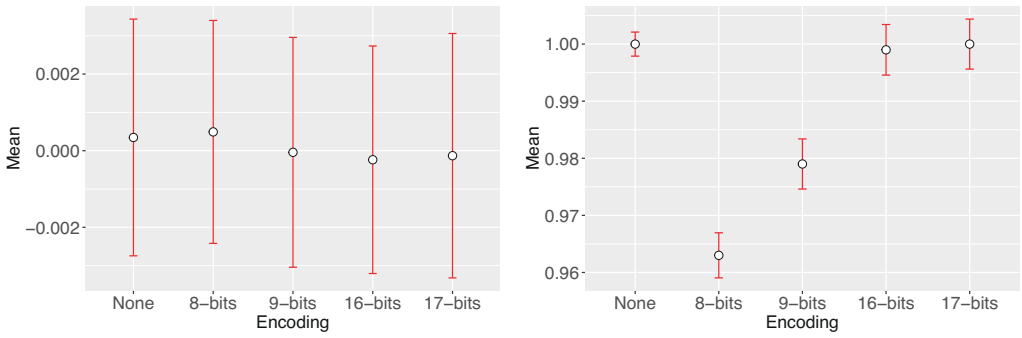
number zero and the largest integer value encodes the number one. The remainder of the integers are used to encode the unit interval at equally-spaced points. The encoding process then rounds the original p-values towards the nearest of these equally-spaced points. We refer to this approach as a γ-bits encoding. Under the storage protocols that we assess, $\gamma \in \{8, 9, 16, 17\}$ generate valid encodings. We note that our considered encoding scheme is only a simplified method of quantization. More complex encoding schemes are possible, such as those considered in Perlmutter et al. (1998).

3.2 The effect of integer encoding on the null distribution

Let $n = 10^6$, and for each $i \in [n]$, let H_i be a null hypothesis that is tested using a well-specified test resulting in a p-value P_i arising from a uniform distribution over the unit interval (cf. Dickhaus, 2014, Chap. 2). We simulate and encode the n p-values using γ-bits encodings, for all valid values of γ. The respective z-scores from each encoding scenario are computed, and the parameter elements of $f_0(z) = \phi\left(z; \mu_0, \sigma_0^2\right)$ are then estimated via ML estimation.

Here, we naively omit infinite z-scores. The process is repeated 100 times for each encoding rule. We also estimate the parameter elements of $f_0(z)$ for $n = 10^6$ z-scores that are obtained without encoding in order to provide a benchmark. All computations are conducted in R.

Figure 1 visualizes the results from the numerical study that is set up above. In the figure and elsewhere, we denote the estimate/estimator of any quantity θ as $\hat{\theta}$.



(a) Mean and standard errors from 100 ML estimates $\hat{\mu}_0$ of $\mu_0 = 0$.

(b) Mean and standard errors from 100 ML estimates $\hat{\sigma}_0^2$ of $\sigma_0^2 = 1$.

Figure 1: *Monte Carlo study regarding the estimation of $\mu_0 = 0$ and $\sigma_0^2 = 1$, in the presence of integer encodings of p-values. Means are represented by points and standard errors are equal to half the length of the error bars.*

Theoretically, we would anticipate that there is no deviation away from a standard normal distribution when no encoding is introduced. This is exactly what we observe in Figure 1(a), where neither the average of the mean nor variance estimates are outside of a 95% confidence interval (i.e., approximately Mean $\pm 2 \times$ SE, where SE is the

standard error). In fact, only two encoding schemes (8 and 9-bits encodings) resulted in significant differences of any kind, from the anticipated estimated values. Further notes regarding the interpretation of Figure 1 appears in Section 2.1 of the Supplementary Materials.

3.3 The effect on the z-score distribution

Now suppose that the hypotheses H_i are generated from two populations, a null one with probability $\pi_0 = 0.8$, and an alternative one with probability $\pi_1 = 0.2$. Under the null hypothesis, we generate test statistics T_i from a standard normal distribution, and under the alternative, we generate test statistics from a normal distribution with mean $\mu_1 = 2$ and variance $\sigma_1^2 = 1$, instead. The p-values $P_i = 1 - \Phi(T_i)$, for testing the null that the test statistics are standard normal, are also computed. Again, we let $n = 10^6$.

Encoding of the p-values is again conducted under the protocol that are described in Section 3.1. We then compute z-scores and discard any infinite values. The parameter vector $\boldsymbol{\theta}$ is then estimated via ML estimation. The process is again repeated 100 times for each encoding type. ML estimation is conducted via the usual EM algorithm for finite mixtures of normal distributions via the normalmixEM2comp function from the package mixtools (Benaglia et al., 2009). The result of this numerical study is visualized in Figure 1 of the Supplementary Materials.

The estimated parameter elements were uniformly significantly different from the generative values for the model. As γ increases, we observe that the estimated values appear to approach the nominal parameter values. However, this approach appears to be slow and still leads to significantly incorrect estimates, even for the largest considered γ. A quantification of this incorrectness appears in Section 2.2 of the Supplementary Materials.

4 Assessment of the binned estimator

4.1 Accuracy of z-score distribution

We first repeat the experiment from Section 3.3, except instead of ML estimation via the normalmixEM2comp function from the package mixtools, we conduct MML estimation via the mix function from the package mixdist. The results from the experiment, using binning schemes obtained via the histogram binning techniques of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981) are visualized in Figure 2 of the Supplementary Materials. Interpretation of appears in Section 3.1 of the Supplementary Materials.

We note that there is only one set of plots where we do not observe the uniform accuracy of the MML estimator, across the binning schemes that are applied. Under 8-bits encoding, we observe that only the Sturges-binned MML estimator yielded accurate es-

timates of the generative parameter elements. Both the Freedman-Diaconis (FD) and Scott-binned estimators resulted in significantly inaccurate estimates of the null proportion and alternative mean and variance parameters. We note that the Sturges binning leads to faster EM algorithm runtimes due to the fact that fewer numerical integrals are required in the E-step, as described in Section 1.1 of the Supplementary Materials. Since we do not observe any benefits from using FD or Scott-type binning in cases where all three methods yielded accurate estimates, we shall henceforth only consider the use of Sturges bins.

4.2 FDR control experiment

We perform a set of five numerical simulation scenarios, in order to assess the performance of the EB-based FDR control rule that is described in Section 2.3. These studies are denoted S1–S5, and will be described in the sequel.

In each of the scenarios, we generate $n = 10^6$ test statistics $T_1, \ldots, T_n$, with proportion $\pi_0 = 0.8$ that H_i is null ($i \in [n]$). The generative distribution of T_i given H_i is null or alternative differs by the simulation study. However, under each studied scenario, the null hypothesis is assumed to be that T_i is standard normal, and thus p-values are computed as $P_i = 1 - \Phi(T_i)$.

The p-values $P_1, \ldots, P_n$ then undergo the various valid encodings that were previously considered. The EB-based FDR control method is then used to decide which of the hypotheses H_i are significant, at the FDR control level $\beta \in \{0.05, 0.10\}$, based only on the encoded p-values. We compute the false discovery proportion (FDP) and true positive proportion (TPP) from the experiment as measures of performance of FDR control and testing power. The measures FDP and TPP are defined as $\text{FDP} = N_{01}/N_R$ and $\text{TPP} = N_{11}/N_1$, where N_{11} is the number of false positives, N_R is the number of rejected hypotheses (declared significantly alternative), N_{11}, is the number of true positives, and N_1 is the number of alternative hypotheses from the simulated experiment. For each simulation scenario, the experiment is repeated $\text{Reps} = 100$ times and the performance measurements are averaged over the repetitions.

For comparison, we also perform FDR control using the popular methods of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), which we denote as BH and BY, respectively. We also compare our EB-based FDR control to the EB-related FDR control technique of Storey (2002), which is commonly referred to as q-values. We implement the BH and BY methods via the base R `p.adjust` function. The q-values technique is implemented via the `qvalue` package (Storey et al., 2015). Scripts for conducting studies S1–S5 are available at `https://github.com/hiendn/FDR_for_grouped_P_values`.

4.3 Simulation scenarios

In Scenario S1, we independently generate T_i from a standard normal distribution, given that H_i is null, and from a normal distribution with mean 2 and variance 1, otherwise.

This scenario is identical to that which is studied Section 3.3.

Table 1: *Average FDP and TPP results (Reps* $= 100$*) for Scenario S5. The best outcome under each encoding for each value of* β *is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	**4.35E-02**	6.15E-02	**1.02E-01**	1.85E-01
	BH	*6.10E-02*	**9.62E-02**	1.82E-01	**3.20E-01**
	BY	2.48E-02	2.80E-02	1.51E-02	3.09E-02
	q-values	*7.14E-02*	1.16E-01	2.26E-01	3.85E-01
8-bits	EB	*1.03E-01*	*1.03E-01*	3.46E-01	3.46E-01
	BH	*1.56E-01*	*2.45E-01*	4.93E-01	6.58E-01
	BY	*1.03E-01*	*1.03E-01*	3.46E-01	3.46E-01
	q-values	*3.65E-01*	*5.62E-01*	8.00E-01	9.33E-01
9-bits	EB	*8.23E-02*	**8.23E-02**	2.70E-01	**2.70E-01**
	BH	*1.66E-01*	2.50E-01	5.15E-01	6.66E-01
	BY	*8.23E-02*	**8.23E-02**	2.70E-01	**2.70E-01**
	q-values	*3.63E-01*	5.59E-01	7.97E-01	9.32E-01
16-bits	EB	3.97E-02	5.67E-02	8.55E-02	1.62E-01
	BH	*1.55E-01*	*2.43E-01*	4.92E-01	6.56E-01
	BY	**4.45E-02**	**5.86E-02**	**1.06E-01**	**1.72E-01**
	q-values	*3.61E-01*	*5.60E-01*	7.97E-01	9.33E-01
17-bits	EB	4.12E-02	5.73E-02	8.53E-02	1.63E-01
	BH	*1.54E-01*	*2.42E-01*	4.90E-01	6.55E-01
	BY	**4.48E-02**	**5.88E-02**	**1.03E-01**	**1.70E-01**
	q-values	*3.60E-01*	*5.58E-01*	7.95E-01	9.32E-01

We consider hypothesis tests that generate dependent test statistics in Scenarios S2 and S3. In S2 two first-order autoregressive sequences of n observations are generated. The null sequence is generated with mean coefficient 0, autoregressive coefficient 0.5, and normal errors with variances scaled so that the overall variance of the sequence is 1. The second chain is the same, except that the mean coefficient is 2 instead of zero. If H_i is null, then T_i is drawn from the first chain; otherwise T_i is drawn from the second chain. See Amemiya (1985, Sect. 5.2) regarding autoregressive models. Scenario S3 is exactly the same as Scenario S2, except that the autoregressive coefficient is set to -0.5 instead of 0.5.

In Scenario S4, we independently generate T_i from a normal distribution with mean 0.5 and variance 1, given that H_i is null, and from a normal distribution with mean 2.5 and variance 1, otherwise. This scenario is misspecified in the sense that the p-values P_i are not computed under the correct null hypothesis. Thus, the distribution of the P_i will not be uniform and thus the well-specified testing assumption of BH, BY, and q-values is not met.

Lastly, in Scenario S5, we independently generate T_i from a Student t-distribution with mean 0.5 and variance 1 and degrees of freedom 25, given that H_i, and from a Student t-distribution with mean 2.5 and variance 1 and degrees of freedom 25, otherwise. Justifications regarding the choices for the five scenarios appear in Section 3.2 of the Supplementary Materials.

4.4 Results

The results for Scenarios S1–S5 are reported in Tables 1–5 of the Supplementary Materials, respectively. We provide the results for S5 in the main text, as it can be viewed as the scenario that is most difficult and is thus most interest.

From Table 1, we observe that q-values is anti-conservative uniformly over all encoding types and FDR control levels in Scenario S5. Furthermore, BH was also uniformly anti-conservative when used to control the FDR at $\beta = 0.05$. The BH method also yielded anti-conservative control of the FDR at $\beta = 0.10$, when the data were encoded using p-type encodings. Both EB and BY were equally anti-conservative for control of FDR at $\beta = 0.05$, when the data were encoded using 8-bits or 9-bits encodings. However, the control at the $\beta = 0.10$ level from both methods for the two aforementioned encoding schemes were both equal and approximately at the correct rate. For all other encoding types, both EB and BY correctly controlled the FDR, for both levels of β. BY appeared more powerful than EB although by only a small amount.

The results above demonstrate that EB along with BY were somewhat more robust to misspecification and data compression via integer encoding than the two other tested methods. Thus, as we had anticipated, there was an observable practical effect to FDR mitigation via conventional methods when p-value data were observed on a discrete support. However, our EB method, and to an extent, the BY method, were able to mitigate against the negative effects of discretization induced by censoring, grouping, and truncation, and thus should be preferred over the other assessed methods in such settings.

For a discussion of results regarding Scenarios S1–S4, we direct the reader to Section 4.4 of the Supplementary Materials. From the results of Scenarios S1–S5, we can conclude that the EB method can correctly control the FDR when the tests were well-specified, and are also somewhat robust to misspecification, otherwise.

5 Example application

5.1 Description of data

Correlations between the structural properties of brain regions, as measured over a sample of subjects, are being increasingly studied as a means of understanding neurological

development (Li et al., 2013) and diseases (Seeley et al., 2009, Wheeler and Voineskos, 2014, Sharda et al., 2016). These correlation patterns, which are often referred to as structural covariance in the neuroimaging literature, are widely studied in humans (Alexander-Bloch, Giedd and Bullmore, 2013, Evans, 2013), as well as in animal models such as mice (Pagani, Bifone and Gozzi, 2016).

For our example application, we study neurological magnetic resonance imaging (MRI) data from a sample of 241 mice. The MRI sample of both female and male adult mice were obtained by taking the control data from a phenotyping study (Ellegood et al., 2015) in order to create a representative wildtype population with variability. All mice were scanned ex-vivo after perfusion with a gadolinium-based contrast agent, and all images were obtained at the same location (i.e. the Mouse Imaging Centre). Scanning was performed on a Varian 7T small animal MR scanner that was adapted for multiple mouse imaging.

The preparation and image acquisition followed a standard pipeline that is similar to the one described in Lerch, Sled and Henkelman (2010). Specifically, a T2-weighted fast-spin echo sequence was used to produce whole-brain images that have an isotropic resolution of 56 micrometers. After images were acquired, the data were corrected for distortions and then registered together by deformation towards a common nonlinear average. The registration pipeline included corrections for nonuniformities that were induced by radio frequency inhomogeneities or gradient-related eddy currents (Sled, Zijdenbos and Evans, 1998). The registered images had a volume of $x \times y \times z = 225 \times 320 \times 152$ voxels, of which $n = 2818191$ voxels corresponded to neurological matter. The exported data were stored in the MINC format.

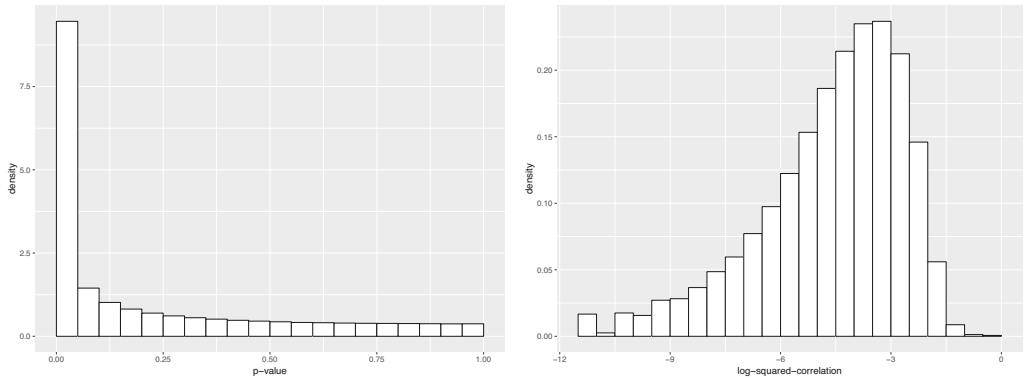
As an output, the registration process produces a set of Jacobian determinants that provide a measure of the extent in which a voxel from the average brain must expand or contract in order to match each of the individual brains of the sample. The Jacobian determinants field of each sample individual is thus a measure of local volume change. For further processing, the Jacobian determinants are log-transformed in order to reduce skewness.

5.2 Hypothesis testing

Upon attainment of the sample of 241 Jacobian determinant fields from the registered mice brain MRIs, we can assess whether or not the local volume change at any particular voxel is correlated with some region of interest. To do so, we select a "seed" voxel within the region of interest and compute the voxelwise sample (Pearson) correlation between the log-transformed Jacobian determinant of the seed voxel and those at every other voxel in the sample of MRIs. This correlation measure can then be used as a measure of structural covariance of the region of interest and the rest of the brain. In the past, structural covariance methods have been used to draw inference regarding a broad array of phenomena such as cortical thickness (Lerch et al., 2006), and cortical maturation and development (Raznahan et al., 2011).

Thus at each of the $n = 2818191$ voxels we computed a correlation coefficient. Using the correlation coefficients, we conducted voxelwise tests of the null hypothesis that the true correlation between the log-transformed Jacobian determinants of the seed voxel and voxel $i \in [n]$ is zero versus the two-sided alternative. The p-values of each test were computed using the Fisher z-transformation and normal approximation (Fisher, 1921).

Using the seed voxel at spatial location $(x,y,z) = (125,124,64)$ – within the bed nucleus of the stria terminalis – we conducted the hypothesis tests, as described above. Histograms of the p-values and log-squared correlation coefficients can be found in Figure 2. We note that the histogram of the log-squared correlation coefficients omits 35856 voxels that had zero correlation with the seed voxel. Further note that a correlation of one yields a log-squared coefficient of ≈ -0.69.



(a) Histogram of p-values. (b) Histogram of log-squared correlation coefficients.

Figure 2: *Histograms of p-values and log-squared correlation coefficients for the structural covariance experiment with seed voxel* $(x,y,z) = (125,124,64)$ *are presented in subplots (a) and (b), respectively.*

An inspection of Figure 2 reveals that the p-value distribution from the experiment deviates significantly from a uniform distribution. The magnitude of the deviation indicates that there may be a potentially large number of voxels that are strongly correlated with the seed voxel, and thus with the region of interest that the seed voxel represents. Using FDR control, we can attempt to identify these correlated voxels in a manner that limits the potential number of false discoveries that are made.

Using the unique function in R, we observed that there were only 66249 discrete and unique numerical values that made up the sample of p-values. These discrete values include zero and one, making up 311575 and 6 voxels of the p-value sample, respectively. Our observations indicate that the data were censored and grouped, at some stage in processing pipeline. It is difficult to tell how such incompleteness were induced, since there may have been multiple encodings of the data along the pipeline that has resulted in the final reported outputs. As such, from our earlier discussions, it would be prudent to apply our EB-based FDR control methodology, since it explicitly accounts for the encoded nature of the data. Furthermore, due to the mathematical approximation via the

use of the Fisher z-transformation as well as the omission of other variables that may contribute to the analysis such as covariates describing the mice (e.g. gender and model strain), the null hypothesis that the population correlation is equal to zero is likely to be misspecified. From Section 4.3, we have observed that the EB-based method is effective in such a setting.

5.3 FDR control

We firstly transform the p-values p_i to the z-scores $p_i = \Phi^{-1}(1-p_i)$, for each $i \in [n]$. A histogram of the z-scores that is obtained is presented in Figure 3. We note that the z-scores that are obtained from the 311581 with p-values equal to zero or one are omitted in this plot. There is a clear truncation of the histogram at the z-score value of 4.169 which corresponds to the smallest non-zero p-value of 1.53E-05.

Using the methods from Section 2, we fit the EB mixture model and obtain the parameter vector

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\top} &= \left(\hat{\pi}_0, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2\right) \\ &= \left(0.5035, 0.5141, 1.200^2, 2.9568, 1.785^2\right), \end{aligned} \tag{6}$$

which corresponds to the mixture model,

$$f\left(z;\hat{\boldsymbol{\theta}}\right) = 0.5035 \times \phi\left(z; 0.5141, 1.200^2\right) + 0.4965 \times \phi\left(z; 2.9568, 1.785^2\right). \tag{7}$$

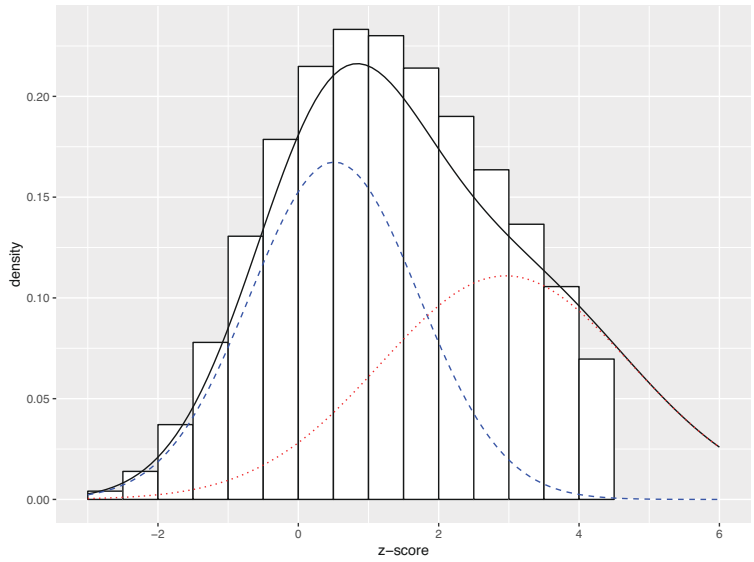


Figure 3: *The functions $f\left(\cdot;\hat{\boldsymbol{\theta}}\right)$, $\hat{\pi}_0\hat{f}_0$, and $\hat{\pi}_1\hat{f}_1$ are plotted with solid, dashed, and dotted lines, respectively.*

As in Sect. 4, we use the Sturges binning scheme that was previously described in Section 2.4. Let $\hat{f}_0(z) = \phi\left(z; 0.5141, 1.200^2\right)$ and $\hat{f}_1(z) = \phi\left(z; 2.9568, 1.785^2\right)$ be the estimates of f_0 and f_1, respectively. We visualize $f\left(\cdot; \hat{\boldsymbol{\theta}}\right)$, $\hat{\pi}_0\hat{f}_0$, and $\hat{\pi}_1\hat{f}_1$ together in Figure 3. A discussion regarding the goodness-of-fit of (7) is provided in Section 4.1 of the Supplementary Materials.

Upon inspection of Figure 3, we observe that mixture model (7) provides a good fit to the suggested curvature of the histogram. The estimated parameter vector from (6) indicates that the null distribution is significantly shifted to the right. This may be due to a combination of the effects of encoding and the effects of mathematical misspecification of the test and omission of covariates. We further observe that there is a large proportion (almost 50%) of potentially alternative hypotheses. Given such a high number, there is potentially for numerous false positives if we were to reject the null using the p-value (or z-score) alone. Thus, we require FDR control in order to make more careful inference.

Using Eqs (4) and (5), we controlled the estimated *m*FDR at the $\beta = 0.1$ level by setting the threshold $c_{0.1} = 0.09986$. This resulted in 608685 of the voxels being declared significantly correlated with the seed, under FDR control, which equates to 21.60%.

For comparison, using BH, BY, and q-values to control the FDR at the same $\beta = 0.1$ level, we obtain 1314429, 727102, and 1718143 significant voxels, respectively. Correspondingly, these numbers respectively translate to 46.64%, 25.80%, and 60.97% of the total number of hypotheses tested. Given the similarity of this testing scenario to simulation study S4, we can expect that the BH and q-values methods are grossly anti-conservative in their control and are would therefore would yield a greater FDR level than that which is desired. We observe, as in our simulations, that our method and BY tend to result in similar numbers of rejections. Whether one method or the other is overly conservative or anti-conservative in this case cannot be deduced without further assessment of the true significance of the rejected hypotheses.

Figure 4 displays visualizations of the significant voxels using our EB method at the perpendicular cross-sections intersecting the seed point $(x, y, z) = (125, 124, 64)$. Upon inspection of Figure 4 we observe that significant correlation with the seed vector appears to be exhibited across the brain. The displays A2 and A3 in Figure 4 further show that the correlation appears to be symmetric between the two hemispheres. Furthermore, the correlation patterns appear in contiguous and smooth regions.

The observations of whole-brain correlation with the bed nucleus of the stria terminalis are well supported in the literature. For example, similar connectivity observations were made by Dong et al. (2001) and Dong and Swanson (2006) in mouse studies, and by McMenamin and Pessoa (2015) and Torrisi et al. (2015) in human studies.

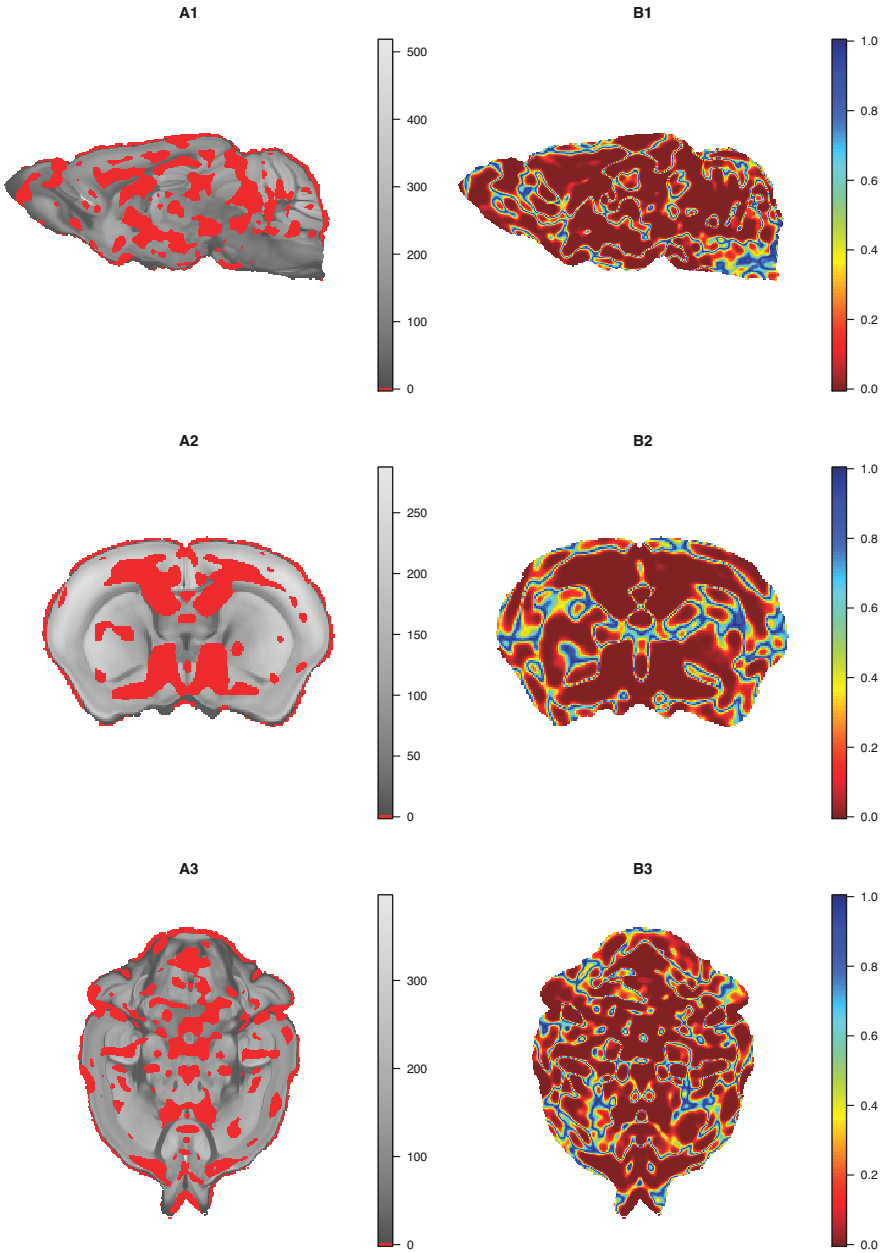


Figure 4: *A1 and B1 display the anatomic background MRI intensities and p-values for the $x = 125$ slice, respectively. Similarly A2 and B2 display the respective quantities for the $y = 124$ slice, and A3 and B3 display the respective quantities for the $z = 64$ slice. In A1–A3, red voxels indicate those that are significant when controlled at the $\beta = 0.1$ FDR level.*

6 Conclusions

We have presented an EB-based FDR control method for the mitigation of false positive results in multiple simultaneous hypothesis testing scenarios where only p-values are available from the hypothesis tests, and when these p-values are distributed on a discrete support. Due to the nature of the construction of our method, it is robust to situations where the hypothesis tests are also misspecified or when there may be omitted covariates that have not been included in the testing procedures for regression models.

In order to handle the discretization induced by censoring, grouping, or quantization of p-value data, we utilized a finite mixture model that can be estimated from binned data. We proved that the parameter vector of the mixture model can also be estimated consistently, even when the testing data may be correlated. A simulation study was used to demonstrate that our methodology was competitive with some popular methods in well-specified testing scenarios, and outperformed these methods when the testing data arise from misspecified tests.

Finally a brain imaging study of mice was conducted to demonstrate our methodology in practice. The study constituted a whole-brain voxel-based study of connectivity to the bed nucleus of the stria terminalis, consisting of $n = 2818191$ tests. The p-values for the study were obtained from a complex pipeline that resulted in a set of quantized values, which included zeros and ones. Furthermore, the p-values were correlated (due to the spatial nature of imaging and subsequent processing) and the hypothesis tests were conducted under mathematical assumptions that may have lead to misspecification. As such, the use of our methodology was most suitable for the study. As a result of the study, we found whole-brain correlation patterns that were consistent with those found in the literature.

Conducting FDR control when p-values are distributed on a discrete support, such as when the values are incompletely observed or when tests are conducted via Monte Carlo or permutation schemes, is an interesting inferential problem and requires careful attention. Our developed methodology provides a simple and robust solution when performing inference with such p-value data.

Acknowledgements

HDN is funded by Australian Research Council project DE170101134. GJM was funded partially by the Australian Government through the Australian Research Council (project numbers IC170100035, DP170100907). The authors are also thankful for the many enlightening and useful comments from the Editorial Board and from the Referees that have greatly improved the expositional quality of the paper.

Supplementary Materials

The Supplementary Materials for the article can be found online at `https://github.com/hiendn/FDR_for_grouped_P_values`.

References

Alexander-Bloch, A., Giedd, J. N. and Bullmore, E. (2013). Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience*, 14, 322–336.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry—the methods. *NeuroImage*, 11, 805–821.

Barreto, H. and Howland, F. M. (2006). *Introductory Econometrics Using Monte Carlo Simulation with Microsoft Excel*. Cambridge: Cambridge University Press.

Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32, 1–29.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society B*, 72, 405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165–1188.

Bennett, C. M., Baird, A. A., Miller, M. B. and Wolford, G. L. (2009). Neuro correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. In *15th Annual meeting of the Organization for Human Brain Mapping*.

Bidgood, W. D., Horri, S. C., Prior, F. W. and Van Syckle, D. E. (1997). Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4, 199–212.

Birge, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10, 24–45.

Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., Lancaster, J. L., Rex, D. E., Smith, S. M., Woodward, J. B. and Strother, S. C. (2004). A (sort of) new image data format standard: Nifti-1. *Neuroimage*, 22, e1440.

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*, 39, 1–38.

Denby, L. and Mallows, C. (2009). Variations on the histogram. *Journal of Computational and Graphical Statistics*, 18, 21–31.

Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. New York: Springer.

Dong, H.-W., Pretrovich, G. D., Watts, A. G. and Swanson, L. W. (2001). Basic organization of projections from the oval and fusiform nuclei of the bed nuclei of the stria terminalis in adult rat brain. *Journal of Comparative Neurology*, 436, 430–455.

Dong, H.-W. and Swanson, L. R. (2006). Projections from bed nuclei of the stria terminalis, anteromedial area: cerebral hemisphere integration of neuroendocrine, autonomic, and behavioral aspects of energy balance. *Journal of Comparative Neurology*, 494, 142–178.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99, 96–104.

Efron, B. (2010). *Large-scale Inference*. Cambridge: Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151–1160.

Ellegood, J., Anagnostou, E., Babineau, B., Crawley, J., Lin, L., Genestine, M., Dicicco-Bloom, E., Lai, J., Foster, J., Penagarikano, O., Geshwind, H., Pacey, L. K., Hampson, D. R., Laliberte, C. L., Mills, A. A., Tam, E., Osborne, L. R., Kouser, M., Espinosa-Becerra, F., Xuan, Z., Powell, M., Raznahan, A., Robins, D. M., Nakai, N., Nakatani, J., Takumi, T., van Eede, M. C., Kerr, T. M., Muller, C., Blakely, R. D., Veenstra-VanderWeele, J., Henkelman, R. M. and Lerch, J. P. (2015). Clustering autism: using neuroanatomic difference in 26 mouse models to gain insight into the heterogeneity. *Molecular Psychiatry*, 20, 118–125.

Evans, A. C. (2013). Networks of anatomical covariance. *Neuroimage*, 80, 489–504.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3–32.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 453–476.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15, 870–878.

Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. New York: Springer.

Habiger, J. D. and Pena, E. A. (2011). Randomised P-values and nonparametric procedures in multiple testing. *Journal of Nonparametric Statistics*, 23, 583–604.

Kontkanen, P. and Myllymaki, P. (2007). MDL histogram density estimation. In *Artificial Intelligence and Statistics* (pp. 219–226).

Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Statistical Planning and Inference*, 124, 379–398.

Larobina, M. and Murino, L. (2014). Medical image file formats. *Journal of Digital Imaging*, 27, 200–206.

Lerch, J. P., Sled, J. G. and Henkelman, R. M. (2010). *Magnetic Resonance Neuroimaging*, chapter MRI phenotyping of genetically altered mice, (pp. 349–361). Springer: New York.

Lerch, J. P., Worsley, K., Shaw, W. P., Greenstein, D. K., Lenroot, K. L., Giedd, J. and Evans, A. C. (2006). Mapping anatomic correlations across cerebral cortex (MACACC) using cortical thickness from MRI. *NeuroImage*, 31, 993–1003.

Li, X., Pu, F., Fan, Y., Niu, H., Li, S. and Li, D. (2013). Age-related changes in brain structural covariance networks. *Frontiers in Human Neuroscience*, 7, 98.

MacDonald, P. D. M. and Du, J. (2012). *mixdist: Finite Mixture Distribution Models*. Comprehensive R Archive Network.

McLachlan, G. J., Bean, R. W. and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22, 1608–1615.

McLachlan, G. J. and Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44, 571–578.

McMenamin, B. W. and Pessoa, L. (2015). Discovering networks altered by potential threat ("anxiety") using quadratic discriminant analysis. *Neuroimage*, 116, 1–9.

Moschitta, A., Schoukens, J. and Carbone, P. (2015). Information and statistical efficiency when quantizing noisy DC values. *IEEE Transactions on Instrumentation and Measurement*, 64, 308–317.

Nguyen, H. D., McLachlan, G. J., Cherbuin, N. and Janke, A. L. (2014). False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Transactions on Medical Imaging*, 33, 1735–1748.

Pagani, M., Bifone, A. and Gozzi, A. (2016). Structural covariance networks in the mouse brain. *Neuroimage*, 129, 55–63.

Perlmutter, S. M., Cosman, P. C., Tseng, C.-W., Olshen, R. A., Grey, R. M., Li, K. C. P. and Bergin, C. J. (1998). Medical image compression and vector quantization. *Statistical Science*, 13, 30–53.

Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values whenn permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9, 1–12.

Pollard, C. S. and van der Laan, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *Statistical Planning and Inference*, 125, 85–100.

R Core Team (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Raznahan, A., Lerch, J. P., Lee, N., Greenstein, D., Wallace, G. L., Stockman, M., Clasen, L., Shaw, P. W. and Giedd, J. N. (2011). Patterns of coordinated anatomical change in human cortical development: a longitudinal neuroimaging study of maturational coupling. *Neuron*, 72, 873–884.

Robb, R. A., Hanson, D. P., Karwoski, R. A., Larson, A. G., Workman, E. L. and Stacy, M. C. (1989). Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis. *Computerized Medical Imaging and Graphics*, 13, 433–454.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66, 605–610.

Seeley, W. W., Zhou, R. K. C. J., Miller, B. L. and Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62, 42–52.

Sharda, M., Khundrakpam, B. S., Evans, A. C. and Singh, N. C. (2016). Disruption of structural covariance networks for language in autism is modulated by verbal ability. *Brain Structure and FUnction*, 221, 1017–1032.

Sled, J. G., Zijdenbos, A. P. and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17, 87–97.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64, 479–498.

Storey, J. D., Bass, A. J., Dabney, A. and Robinson, D. (2015). *qvalue: Q-value estimation for false discovery rate control*.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21, 65–66.

Torrisi, S., O'Connell, K., Davis, A., Reynolds, R., Balderston, N., Fudge, J. L., Grillon, C. and Ernst, M. (2015). Resting state connectivity of the bed nucleus of the stria terminalis at ultra-high field. *Human Brain Mapping*, 36, 4076–4088.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society B*, 38, 290–295.

van der Laan, M. J. and Hubbard, A. E. (2006). Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology*, 5, 14.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92, 1–28.

Vincent, R. D., Janke, A., Sled, J. G., Baghdadi, L., Neelin, P. and Evans, A. C. (2003). MINC 2.0: a modality independent format for multidimensional medical images. In *10th Annual Meeting of the Organization for Human Brain Mapping*.

Wand, M. P. (1997). Data-Based Choice of Histogram Bin Width. *The American Statistician*, 51, 59–64.

Wheeler, A. L. and Voineskos, A. N. (2014). A review of structural neuroimaging in schizophrenia: from connectivity to connectomics. *Frontiers in Human Neuroscience*, 8, 653.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92, 381–397.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Statistical Planning and Inference*, 82, 171–196.

SORT 43 (2) July-December 2019, 259-288 DOI: 10.2436/20.8080.02.88

Kernel distribution estimation for grouped data

Miguel Reyes[1], Mario Francisco-Fernández[2], Ricardo Cao[2]
and Daniel Barreiro-Ures[2]

Abstract

Interval-grouped data appear when the observations are not obtained in continuous time, but monitored in periodical time instants. In this framework, a nonparametric kernel distribution estimator is proposed and studied. The asymptotic bias, variance and mean integrated squared error of the new approach are derived. From the asymptotic mean integrated squared error, a plug-in bandwidth is proposed. Additionally, a bootstrap selector to be used in this context is designed. Through a comprehensive simulation study, the behaviour of the estimator and the bandwidth selectors considering different scenarios of data grouping is shown. The performance of the different approaches is also illustrated with a real grouped emergence data set of *Avena sterilis* (wild oat).

MSC: 62G05, 62N99, 62G09.

Keywords: Bootstrap bandwidth, cumulative distribution function estimator, interval data, plug-in bandwidth

1 Motivation

In the experimental sciences, data usually come from measurements of continuous variables such as temperature, mass, weight, time, length, etc. However, for several reasons, measurements are always obtained in finite precision; i.e., all observed data are rounded or grouped to some extent.

A typical situation in which grouped data clearly appear (and the degree of grouping can be considerable) is when researchers observe variables not continuously, but periodically, thus obtaining time to event data distributed along a set of consecutive intervals. Situations like this appear very frequently in areas such as engineering, economics, social sciences, epidemiology, medicine, agriculture and more (Coit and Dey, 1999, Guo, 2005, Minoiu and Reddy, 2009, Pipper and Ritz, 2007, Rizzi et al., 2016). Especially in these cases, data uncertainty should be taken into account to avoid serious mistakes when making inferences.

[1] Departamento de Actuaría, Física y Matemáticas. Universidad de las Américas-Puebla, Puebla, México. miguel.reyes@udlap.mx
[2] Research Group MODES. Departamento de Matemáticas, Facultade de Informática, CITIC, ITMATI. Universidade da Coruña, A Coruña, Spain. mariofr@udc.es, ricardo.cao@udc.es, daniel.barreiro.ures@udc.es
Received: February 2019
Accepted: July 2019

One of these situations that partially motivated this work was a real problem from weed science, where weed emergence is coded as a set of non-equally spaced grouped data. In this framework, a key variable to study seedling emergence is the cumulative hydrothermal time (CHTT), which is a mix of time of exposure to certain temperature and humidity conditions. CHTT is typically available in k inspections and the number of emerged seedlings at each one is registered. In more restrictive situations, only the cumulative proportion of emerged seedlings recorded at every monitoring date are reported. Most of the statistical methods used in this context tackle the problem of modeling weed emergence (the so-called emergence curve) from a regression point of view. Parametric models such as Gompertz and logistic have been widely used to define the relationship between the CHTT and weed emergence. However, due to the limitations of this approach, in Cao et al. (2013), this problem has been dealt with through nonparametric estimation of the distribution function of the CHTT at emergence. In that paper, a simple kernel distribution estimator adapted to deal with grouped data, based on a modification of the standard kernel estimator of the distribution function, was proposed and applied to analyse a weed emergence data set. This nonparametric approach has recently been proven to outperform the classical regression methods in terms of prediction error (González-Andújar et al., 2016). However, a deeper statistical analysis of this new nonparametric distribution estimator is required. In the present paper, we study the asymptotic properties of this estimator. Additionally, a plug-in and a bootstrap bandwidth selector are proposed and compared in different scenarios through a comprehensive simulation study.

The organization of this paper is as follows. In Section 2 the notation used throughout the paper and the kernel distribution estimator for grouped data are presented. In Section 3, under some assumptions, the asymptotic bias, variance, and mean integrated squared error (MISE) of this estimator are obtained. In Section 4, using the asymptotic MISE expression, a plug-in bandwidth selector is proposed. Additionally, closed forms for the MISE and its bootstrap version, MISE*, are presented and a bootstrap bandwidth selector is derived. In Section 5, a simulation study with different sample sizes is presented to show the consistency of the estimator under different grouping scenarios. In Section 6, the nonparametric estimator and both bandwidth selection methods are applied to a grouped emergence data of *Avena sterilis* (wild oat). Finally, Section 7 summarizes the main conclusions. Proofs are included in Appendix A. Supplementary materials completing the simulation study and with an additional empirical study based on real data are available online.

2 Kernel distribution estimator for interval-grouped data

Let us introduce the notation for grouped data. Suppose that X is the random variable of interest, with density function f and distribution function F, and let $(X_1, X_2, \ldots, X_n)$ be a random sample of X. Consider a set of intervals $[y_{j-1}, y_j)$, $j = 1, 2, \ldots, k$, where

the j-th interval length is $l_j = y_j - y_{j-1}$, its midpoint is $t_j = \frac{1}{2}(y_{j-1} + y_j)$, and denote the number of observations within each interval by $(n_1, n_2, \ldots, n_k)$. Sometimes, only the sample proportions $(w_1, w_2, \ldots, w_k)$ are available, where $w_j = F_n(y_j-) - F_n(y_{j-1}-)$ is the actual observed random quantity, and $F_n(y-)$ is the left-hand limit of the empirical distribution function F_n.

For example, using this notation and focusing on the weed emergence problem that motivated this research, X would be the random variable measuring the CHTT at emergence of a particular weed. Moreover, denoting by n the number of seedlings that have emerged at the end of the monitoring process, since the inspections carried out to count the number of emerged seedlings are performed at a limited number of k instants, the values $X_1, X_2, \ldots, X_n$, measuring the CHTT at emergence of every single seedling, cannot be observed. In this case, what is observed is the CHTTs at inspections (the limits of the intervals, previously denoted by y_i, $i = 0, 1, \ldots, k$) and the total number of seedlings that have emerged in the intervals between consecutive inspection times, $n_1, n_2, \ldots, n_k$, (or the corresponding sample proportions, $w_1, w_2, \ldots, w_k$, with $w_i = n_i/n$).

It is worth mentioning that there exists a parallelism between grouped data and the so called interval-censored data (see the book by Klein and Moeschberger, 1997, for an introduction about interval-censored data in survival analysis). The main similarity is that the exact value of the interest random variable data X_i is not observed and one is only able to know the interval in which every datum of the interest population belongs. There are two main differences between grouped data and interval-censored data. The first one is that the intervals $[y_{j-1}, y_j)$ are typically fixed (not random) for grouped data, while the interval endpoints are random variables for interval-censored data. As a consequence, for interval-censored data there are, in principle, as many different intervals as the sample size, n, while for grouped data the number of different intervals, k, is known beforehand and is smaller than the sample size ($k < n$). General estimation methods applicable for interval-censored data, as Turnbull's estimator (Turnbull, 1976), can also be used for grouped data. In our grouped data setup, Turnbull's estimator of the cumulative distribution function just gives the empirical cumulative distribution function for grouped data.

First, let us consider the ideal continuous case, where $(X_1, X_2, \ldots, X_n)$ are supposed to be observed. From the well-known Parzen-Rosenblatt kernel density estimator (Parzen, 1962, Rosenblatt, 1956), defined as

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i), \tag{1}$$

where $K_h(u) = h^{-1} K(u/h)$, with K a kernel function (typically an auxiliary density function) and h the bandwidth parameter, it is straightforward to obtain a kernel estimator of the cumulative distribution function (cdf) as

$$\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)\,dt = \frac{1}{n}\sum_{i=1}^{n} \mathbb{K}\left(\frac{x-x_i}{h}\right), \tag{2}$$

where $\mathbb{K}(t) = \int_{-\infty}^{x} K(t)\,dt$. Some theoretical properties of (2) can be found in Hill (1985), Nadaraya (1964) and Reiss (1981). Although the choice of the kernel function is of secondary importance, the bandwidth h plays a crucial role in the shape of estimator (2). Differently to the case of kernel density estimation, there are not many contributions addressing the bandwidth selection problem in kernel distribution estimation. Different cross-validation methods were studied in Bowman, Hall, and Prvan (1998) and Sarda (1993); and plug-in selectors were considered in Altman and Leger (1995) and Polanski and Baker (2000). The interested reader can find more theoretical details and an extended discussion on the previous cross-validation and plug-in selectors in Quintela-del-Río and Estévez-Pérez (2012). More recently, a bootstrap bandwidth selector for the estimator (2) has been developed in Dutta (2015).

When working with grouped data, the issue of density estimation has been widely addressed employing different approaches. Using nonparametric methods, in Reyes, Francisco-Fernandez, and Cao (2016) a simple modification of the estimator (1) was proposed and studied, while in Reyes, Francisco-Fernández, and Cao (2017) two different bandwidth selectors for this estimator were analysed. Also in a nonparametric context, Wang and Wertelecki (2013) proposed a bootstrap type kernel density estimator for binned data, and Blower and Kelsall (2002) proposed a nonlinear binned kernel estimator. In this setting, Rizzi et al. (2016) compared the performance of different nonparametric density estimators for grouped data via a simulation study and also using some empirical cancer data. Other approaches in this framework consist in converting the density estimation problem to a regression problem using the root-unroot algorithm (Brown et al., 2010) or using parametric methods (Wang and Wang, 2016). Parametric methods can be useful for heavy grouping if the assumed model is correct, but if this is not true, the results obtained can be wrong.

Studies on estimation methods for the distribution function for grouped data are much scarcer and they are mainly based on the empirical distribution function (Turnbull, 1976). Although the distribution function is closely connected with the density function, in some situations, the data are collected in an accumulated way, making the distribution function the element of interest. This is the case, for example, in the weed emergence problem previously described. Therefore, it is of special concern to develop and study specific distribution function estimators for grouped observations.

Starting with the related density estimation problem, in Scott and Sheather (1985) and Titterington (1983), the kernel density estimator (1) was redefined to be used with binned data, assuming a constant binwidth, as

$$\tilde{f}_h(x) = n^{-1}\sum_{i=1}^{k} n_i K_h(x-t_i). \tag{3}$$

In Reyes et al. (2016), a modified version of (3) considering the general case of different interval lengths, given by

$$\hat{f}_h^g(x) = \frac{1}{h}\sum_{i=1}^{k} w_i K\left(\frac{x-t_i}{h}\right), \tag{4}$$

was studied. Its asymptotic properties were obtained, and a plug-in bandwidth selector was proposed and analysed.

From (4), it is straightforward to obtain a kernel distribution estimator for binned or grouped data as

$$\hat{F}_h^g(x) = \int_{-\infty}^{x} \hat{f}_h^g(u)\,du = \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x-t_i}{h}\right), \tag{5}$$

where $\mathbb{K}(x) = \int_{-\infty}^{x} K(z)\,dz$. Note that (5) is a simple modification of (2) for the context of interval-grouped data.

3 Theoretical results

In this section, a closed form for the MISE of the kernel distribution estimator for grouped data (5) is obtained, and its asymptotic properties are derived. Using standard calculations and assuming that $F(y_k) = 1$ and $F(y_0) = 0$, it is easy to prove that the expectation and the variance of $\hat{F}_h^g(x)$ are, respectively,

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) p_i \tag{6}$$

and

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) p_i(1-p_i) - \frac{2}{n}\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right) p_i p_j, \tag{7}$$

where $p_i = F(y_i) - F(y_{i-1})$.

From (6) and (7), it is straightforward to obtain a closed expression for the MISE of the estimator defined in (5):

$$\mathrm{MISE}\left(\hat{F}_h^g\right) = \mathbb{E}\left\{\int \left[\hat{F}_h^g(x) - F(x)\right]^2 dx\right\} = B + V, \tag{8}$$

where,

$$B = \int \left\{\mathbb{E}\left[\hat{F}_h^g(x)\right] - F(x)\right\}^2 dx \tag{9}$$

denotes the integrated squared bias and

$$V = \int \mathbb{V}\left[\hat{F}_h^g(x)\right] dx \tag{10}$$

is the integrated variance.

The asymptotic bias and variance of (5) are stated in Theorem 3.1, whose proof is included in Appendix A. The following assumptions are needed.

Assumption 3.1 *The kernel K is a symmetric probability density function with support in $[-1,1]$, at least 3-times differentiable and such that $K^{(3)}$ is bounded.*

Assumption 3.2 *The distribution F has compact support $[\mathscr{L},\mathscr{U}]$, it is 4-times differentiable and $F^{(4)}$ is continuous.*

Assumption 3.3 *The bandwidth $h = h_n$ is a non random sequence of positive numbers such that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.*

Assumption 3.4 *Given a set of $k = k_n$ intervals $[y_{j-1}, y_j)$, $j = 1,2,\ldots,k$, $y_0 \leqslant \mathscr{L}$ and $y_k \geqslant \mathscr{U}$, the average interval length is $\bar{l} = \bar{l}_n = \frac{1}{k}\sum_{i=1}^{k} l_i$, where l_i is the abbreviated notation for the i-th interval length $l_{i,n}$. It is assumed that $\lim_{n\to\infty} \bar{l} = 0$, $\lim_{n\to\infty} n\bar{l} = \infty$, $\bar{l} = o\left(h^{5/3}\right)$, and $\max_i \left|l_i - \bar{l}\right| = \max_{1\leqslant i\leqslant k} \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.*

Assumptions 3.1 and 3.2 are just smoothness and differentiability conditions about the kernel K and the distribution function F. Assumption 3.3 is the typical one used in kernel estimation concerning the sample size n and the bandwidth h. However, Assumption 3.4 is of special importance and deserves some comments.

Condition $\lim_{n\to\infty} \bar{l} = 0$ simply states that, as the sample size increases, the average interval length shrinks. This means that, taking into account the condition $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$, all intervals are shrinking as well. However, $\lim_{n\to\infty} n\bar{l} = \infty$ states that n should increase faster than $\bar{l}$ decreases. This is an important condition from a theoretical point of view, as if the intervals shrink faster than n increases, at some point there would be more intervals than data points, and some of the intervals would be empty or there would not be enough data points in each interval.

Condition $\bar{l} = o\left(h^{5/3}\right)$ states an intuitive idea: as the sample size n increases, the average length $\bar{l}$ must vanish faster than, at least, h (concretely, faster than $h^{5/3}$). This condition has a practical basis. Since the average distance between points is $\bar{l}$, the bandwidth must be greater than $\bar{l}$ at all times to gather information from the surroundings. In other words, as n increases, h must vanish, but always behind $\bar{l}$.

Regarding the condition about $\max_i \left|l_i - \bar{l}\right|$, at first, this is necessary from the strictly mathematical viewpoint, but in practice it is a way for controlling the variability of the intervals. This assumption means that the lengths of the intervals are not very different. In other words, in our assumptions we unquestionably accept different interval lengths

in order to generalize the binned estimator, but within certain limits, and these limits of maximum variability are controlled by $\bar{l}$ via $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.

Theorem 3.1 *Under Assumptions 3.1 to 3.4,*

$$\mathrm{MSE}\left[\hat{F}_h^g(x)\right] = \frac{h^4}{4}\mu_2(K)^2 F''(x)^2 + \frac{1}{n}F(x)[1-F(x)] - \frac{h}{n}F'(x)C_0 + o\left(\frac{h}{n}\right) + o\left(h^4\right)$$

and

$$\mathrm{MISE}\left(\hat{F}_h^g\right) = \mathrm{AMISE}\left(\hat{F}_h^g\right) + o\left(\frac{h}{n}\right) + o\left(h^4\right),$$

where

$$\mathrm{AMISE}\left(\hat{F}_h^g\right) = \frac{h^4}{4}\mu_2(K)^2 A(f') + \frac{1}{n}\int F(x)[1-F(x)]\,dx - \frac{h}{n}C_0 \tag{11}$$

with $A(f') = \int f'(x)^2 dx$, *and*

$$C_0 = 2\int zK(z)\,\mathbb{K}(z)\,dz > 0.$$

Remark 3.1 *Since the distribution function F has compact support (Assumption 3.2) then the integral $\int F(x)(1-F(x))\,dx$ is finite. This needs not be the case for a general cdf F.*

Remark 3.2 *Taking care of higher order terms in the asymptotic expansions for the MSE and the MISE, the resulting approximations show the impact of the average interval length, $\bar{l}$, in these error criteria:*

$$\begin{aligned}
\mathrm{MSE}\left[\hat{F}_h^g(x)\right] &= \left(\frac{h^2}{2}\mu_2(K) + \frac{\bar{l}^2}{12}\right)^2 F''(x)^2 + \frac{1}{n}F(x)[1-F(x)] - \frac{h}{n}F'(x)C_0 \\
&+ \frac{\bar{l}^2}{24n}F''(x) + o\left(\frac{h}{n}\right) + o\left(h^4\right) + o\left(\bar{l}^4\right) + o\left(h^2\bar{l}^2\right) + o\left(\frac{\bar{l}^2}{n}\right) \\
\mathrm{MISE}\left(\hat{F}_h^g\right) &= \left(\frac{h^2}{2}\mu_2(K) + \frac{\bar{l}^2}{12}\right)^2 A(f') + \frac{1}{n}\int F(x)[1-F(x)]\,dx - \frac{h}{n}C_0 \\
&+ o\left(\frac{h}{n}\right) + o\left(h^4\right) + o\left(\bar{l}^4\right) + o\left(h^2\bar{l}^2\right) + o\left(\frac{\bar{l}^2}{n}\right).
\end{aligned}$$

Under Assumptions 3.1 - 3.4, these two expressions reduce to the asymptotic expressions given in Theorem 3.1.

4 Bandwidth selectors

As pointed out in Section 1, the kernel distribution estimator (2) heavily depends on the bandwidth h. Obviously, the same occurs for the estimator adapted for grouped data (5), since too small bandwidths give estimates that are too close to the empirical cdf, and too large selections tend to provide oversmoothed estimators. In this sense, it is very important to have an automatic bandwidth selection method producing reliable estimates for a real data set. In this section, two bandwidth selectors (plug-in and bootstrap) are proposed for (5) in the context of interval-grouped data.

4.1 Plug-in bandwidth selector

From Eq. (11), it is immediate to get an asymptotically optimal global bandwidth. Taking the first derivative of (11), equating to zero and solving for h, it follows that

$$h_{AMISE} = \left[\frac{C_0}{n\mu_2 (K)^2 A(f')} \right]^{\frac{1}{3}}. \tag{12}$$

Note that Eq. (12) is the same as that for continuous data (see, e.g., Azzalini, 1981, Hill, 1985, Mack, 1984). However, it is important to keep in mind that (12) holds as an asymptotic optimal bandwidth for grouped data only as long as Assumptions 3.1 to 3.4 hold. Otherwise, some other important terms of the asymptotic expansion of MISE $\left(\hat{F}_h^g\right)$ remain non-negligible, thus making (11) fall short as a MISE $\left(\hat{F}_h^g\right)$ approximation.

In Eq. (12), an estimate of $A(f')$ is required to have a practical bandwidth. To estimate $A(f')$, we used the proposal of Polansky and Baker (2000) adapted for grouped data. Other approaches could be used here, but we preferred the Polansky and Baker method for computational reasons and because it gave stable results when using grouped data. In the continuous data case, Polansky and Baker (2000) proposed to estimate $A(f')$ by $-\hat{\psi}_{\eta,2}$, where

$$\hat{\psi}_{\eta,2} = \frac{1}{n^2\eta^3} \sum_{i=1}^{n} \sum_{j=1}^{n} L'' \left(\frac{X_i - X_j}{\eta} \right), \tag{13}$$

L being a kernel function (possibly different from K) and $\eta > 0$ an auxiliary smoothing parameter. The bandwidth η can be selected using a plug-in procedure. For this, it would be necessary to obtain the asymptotic MSE of $\hat{\psi}_{\eta,2}$, that depends on $\psi_4 = \int f^{(4)}(x) f(x) dx$, and then estimate ψ_4. Clearly, the problem still remains, since estimating ψ_4 will depend on an initial bandwidth, which in turn will depend on $\psi_6 = \int f^{(6)}(x) f(x) dx$, and so on. A common strategy is to estimate ψ_u with some quick and simple rule, like the normal scale rule (Wand and Jones, 1995). Once $\hat{\psi}_{\eta,u}$ is obtained, it is possible to select a bandwidth for estimating ψ_{u-2}. Then, having estimated $\hat{\psi}_{\eta,u-2}$, a bandwidth for estimating ψ_{u-4} can be selected, and so forth. Polansky and Baker (2000) suggest using the same iterative method.

In the context of grouped data, we propose to estimate $A(f')$ with $\hat{A}_{PB_g} = -\hat{\psi}^g_{\eta,2}$, where $\hat{\psi}^g_{\eta,2}$ is an appropriate version of (13), given by:

$$\hat{\psi}^g_{\eta,2} = \frac{1}{\eta^3} \sum_{i=1}^{k} \sum_{j=1}^{k} L'' \left(\frac{t_i - t_j}{\eta} \right) w_i w_j. \tag{14}$$

Similar steps to those described previously for continuous data can be followed now to select the bandwidth η. It should be noted that in this case, to obtain a plug-in bandwidth for η it is necessary to derive the asymptotic MSE of $\hat{\psi}^g_{\eta,2}$ using grouped data. In Reyes et al. (2017), both the asymptotic variance and bias of $\hat{\psi}^g_{\eta,u}$ were derived for $u > 0$. Based on those, a way of selecting the plug-in bandwidth for $\hat{\psi}^g_{\eta,u}$ was proposed. Using that approximation with $u = 2$ and plugging $\hat{A}_{PB_g}$ into (12) gives a practical plug-in bandwidth selector for $\hat{F}^g_h(x)$,

$$\hat{h}_{PB_g} = \left[\frac{C_0}{n\mu_2(K)^2 \hat{A}_{PB_g}} \right]^{\frac{1}{3}}. \tag{15}$$

Note that using similar arguments to those employed in Theorem 2 of Reyes et al. (2017), the relative rate of convergence for the plug-in bandwidth $\hat{h}_{PB_g}$ can be derived.

4.2 Bootstrap bandwidth selector

The bootstrap method can be used to produce an estimator of the MISE. In the grouped data setup, this has been already proposed by Reyes et al. (2017) for density estimation. These authors have proved that there exists a closed expression for the bootstrap version of the MISE in that context. This implies that Monte Carlo is not needed to obtain a bootstrap approximation of the MISE in density estimation for grouped data. This will be also the case for cdf estimation for grouped data.

To build a bootstrap version of the MISE, we consider a pilot bandwidth, ζ, and construct the grouped-data smooth estimator of F as defined in (5), but replacing h by ζ. The idea is to draw resamples from $\hat{F}^g_\zeta$, to group the data and to compute the estimator $\hat{F}^g_h$ with those bootstrap samples. The bootstrap resampling plan proceeds as follows.

1. Fix some pilot bandwidth, ζ, and consider the grouped-data smooth cdf estimator, $\hat{F}^g_\zeta$.
2. Draw $(n_1^*, \ldots, n_k^*)$ from a multinomial distribution $\mathscr{M}_k(n; \tilde{p}_1^\zeta, \ldots, \tilde{p}_k^\zeta)$, with $\tilde{p}_i^\zeta = \hat{F}^g_\zeta(y_i) - \hat{F}^g_\zeta(y_{i-1})$, $i = 1, \ldots, k$, and define $w_i^* = n_i^*/n$.

3. Compute the grouped-data smooth cdf estimator based on this bootstrap resample:

$$\hat{F}_h^{g*}(x) = \sum_{i=1}^{k} w_i^* \mathbb{K}\left(\frac{x-t_i}{h}\right).$$

4. Define the bootstrap version of MISE:

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{\int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx\right\},$$

where, $\mathbb{E}^*$ denotes the bootstrap expectation (with respect to $\hat{F}_\zeta^g$).

Remark 4.1 *Since, under Assumption 3.1, the support of $\hat{F}_\zeta^g$ is $[t_1-\zeta, t_k+\zeta]$, it may happen that this interval is not contained in $[y_0, y_k]$. This only happens if $\zeta \leq \frac{1}{2}\min\{l_1, l_k\}$. So, in order to resample from a distribution with support contained in $[y_0, y_k]$, we consider the conditional distribution corresponding to $\hat{F}_\zeta^g$ restricted to the interval $[y_0, y_k]$.*

The previous remark implies that it may happen that $\sum_{i=1}^k \tilde{p}_i^\zeta < 1$. If this is the case, we define

$$\hat{p}_i^\zeta = \frac{\tilde{p}_i^\zeta}{\sum_{j=1}^k \tilde{p}_j^\zeta},\ i = 1, 2, \ldots, k, \tag{16}$$

and we draw the bootstrap resamples in Step 2 from a multinomial distribution with probabilities $\hat{p}_i^\zeta$.

Substituting p_i by $\hat{p}_i^\zeta$ in (6) and (7), the bootstrap version of the mean integrated squared error admits the following closed expression:

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{\int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx\right\} = B^* + V^*,$$

where

$$B^* = \int \left\{\mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] - \hat{F}_\zeta^g(x)\right\}^2 dx$$

and

$$V^* = \int \mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] dx,$$

with

$$\mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) \hat{p}_i^\zeta$$

and

$$\mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] = \frac{1}{n}\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) \hat{p}_i^\zeta\left(1-\hat{p}_i^\zeta\right) - \frac{2}{n}\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)\hat{p}_i^\zeta \hat{p}_j^\zeta.$$

This approach is computationally efficient since there is no need to use Monte Carlo to approximate the bootstrap resampling distribution. Finally, the bootstrap bandwidth is defined as the minimizer of MISE$^*\left(\hat{F}_h^{g*}\right)$, in the smoothing parameter, h:

$$h^*_{MISE} = \arg\min_{h>0} MISE^*\left(\hat{F}_h^{g*}\right). \tag{17}$$

An important step in this bootstrap procedure is that of selecting the pilot bandwidth ζ. After performing some empirical experiments, we have used a method inspired by the idea of smoothing splines, based on selecting the pilot parameter that minimizes the squared distance between the nonparametric cdf estimator and the empirical distribution function, plus a penalty term to avoid obtaining very small bandwidths. The idea consists in finding the parameter, denoted by ζ^{λ}_{emp}, such that,

$$\zeta^{\lambda}_{emp} = \arg\min_{h>0} \sum_{i=0}^{k} \left[F_n(y_i) - \hat{F}_h^g(y_i)\right]^2 + \lambda \int \hat{f}_h^{g\prime}(x)^2\, dx,$$

where $\lambda \geq 0$ determines the penalty degree over the global slope of the nonparametric density estimator, defined in (4). To select an "optimal" penalty degree, λ_{opt}, we have used the rule of finding the penalty allowing to obtain a pilot bandwidth that best approximates the overall slope of the population density, that is,

$$\lambda_{opt} = \arg\min_{\lambda\geq 0} \left| A\left(\hat{f}^{g\prime}_{\zeta^{\lambda}_{emp}}\right) - A(f')\right|.$$

In practice, λ_{opt} can be estimated by

$$\hat{\lambda}_{opt} = \arg\min_{\lambda\geq 0} \left| A\left(\hat{f}^{g\prime}_{\zeta^{\lambda}_{emp}}\right) - A(\hat{f}'_{\theta})\right|,$$

where $\hat{f}'_{\theta}$ represents a parametric estimator of the first derivative of the density function, fitted with the grouped data sample and flexible enough to capture, at least partially, the global slope of f. It was checked that fitting normal mixture models with a maximum number of $r = 5$ components provided, in general, very good results. In practice, the Expectation Maximization (EM) method (Mclachlan and Peel, 2000) was used to estimate the parameters of these models, using the BIC criterion to select the best fit.

Other simpler alternatives to select the pilot bandwidth, ζ, were also explored, producing in general worse results than those obtained with the algorithm previously described. For that reason (and for reason of space), only the results obtained when using the previous approach to select the pilot bandwidth are shown in the paper. In the Supplementary Materials, some simulations experiments comparing the performance of the bootstrap bandwidth (17) when using as a pilot bandwidth ζ^{λ}_{emp} and when this auxiliary parameter is derived using the plug-in technique, $\hat{h}_{PB_g}$, are presented. A better per-

formance of the bootstrap selector is clearly observed when using the pilot bandwidth obtained by the method described above.

5 Simulations

To have an idea of the effectiveness of the estimator (5) when using (15) and (17) as bandwidth selectors, some simulation studies were performed under different grouping scenarios. For this, the free statistical software `R` and packages `nor1mix` and `binnednp` were used (Barreiro et al., 2019, Mächler, 2017, R Core Team, 2019).

As the population density, we used a normal mixture $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}$, with $\phi_{\mu,\sigma}$ a $N\left(\mu,\sigma^2\right)$ density, $\alpha = (0.70, 0.22, 0.06, 0.02)$, $\mu = (207, 237, 277, 427)$ and $\sigma = (25, 20, 35, 50)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively. This normal mixture was used in weed science to model the relationship between weed emergence of *Bromus diandrus* and hydrothermal time (Cao et al., 2011). A total number of 1000 trials were considered throughout all simulations.

Trying to mimic the asymptotic conditions on $\bar{l}$ in Assumption 3.4, in a first simulation experiment, the behaviour of the MISE for grouped data, denoted by MISE_g, was studied depending on the bandwidth h, for sample sizes 60, 240, and 960. Two different scenarios were considered based mainly on Assumption 3.4.

S1. $n^{\frac{5}{9}}\bar{l} \to 0$

S2. $n^{\frac{5}{9}}\bar{l} \to \infty$

In Scenario S1, condition $\bar{l} = o\left(h^{5/3}\right)$ is confirmed for $h \sim n^{-\frac{1}{3}}$ (classical optimal rate in the case of distribution estimation without grouping), for example, h_{AMISE} or $\hat{h}_{PB_g}$; while in Scenario S2 occurs the opposite. It is important to note that in both scenarios $\bar{l}$ tends to zero as n increases.

Note that MISE_g can be approximated by numerical integration in an interval $[a,b]$ using (8), (9) and (10), jointly with the expressions of the expectation and the variance of $\hat{F}_h^g$ in (6) and (7). In practice, we considered $a = 0$, $b = 509.25$. With these values for a and b, the area under the reference normal mixture in $[a,b]$ is 0.999.

To simulate the set of intervals as n increases, the next steps were followed:

1. Consider $\bar{l} = En^{-\alpha}$ and $a_n = Fn^{-\beta}$, where E, α, F and β are positive constants.
2. Take initial interval lengths $\{l_i\}$ for $i = 1, 2, \ldots, 5$: $l_1 = \bar{l} - 4a_n$, $l_2 = \bar{l} + 0.5a_n$, $l_3 = \bar{l} - 1.5a_n$, $l_4 = \bar{l} + 3a_n$, $l_5 = \bar{l} + 2a_n$.
3. For $i > 5$, $l_i = l_{[(i-1) \bmod 5]+1}$, where $[m \bmod \ell]$ stands for m modulo ℓ. Then, the initial set of intervals is repeated one after another, as many times as necessary.

Constants E and F are just selected considering the distribution support. To choose the positive constants α and β, note that according to the initial set of intervals in Step 2, it follows that $\max_i \left|l_i - \bar{l}\right| = 4a_n = 4Fn^{-\beta}$.

Assumption 3.4 and Step 1 impose that $4Fn^{-\beta} = o\left(\bar{l}\right) = o\left(En^{-\alpha}\right)$, which basically is

$$n^{-\beta} = o\left(n^{-\alpha}\right). \tag{18}$$

So, for (18) to hold, $n^{\alpha-\beta} \to 0$, which only occurs when $\alpha - \beta < 0$, i.e., when $\beta > \alpha$.

Now, recall that in Scenario S1, $\bar{l} = o\left(h^{5/3}\right) = o\left(n^{-5/9}\right)$ must hold. Thus, according to Step 1, $\bar{l} = En^{-\alpha} = o\left(n^{-\frac{5}{9}}\right)$, which basically is

$$n^{-\alpha} = o\left(n^{-\frac{5}{9}}\right). \tag{19}$$

This only occurs when $\frac{5}{9} - \alpha < 0$; i.e., when $\alpha > 5/9$.

In brief, to simulate Scenario S1, (18) and (19) must hold, i.e., $\beta > \alpha > 5/9$ must be true. On the other hand, to simulate S2, (18) must hold but (19) must not. It is required that $n^{\frac{5}{9}}\bar{l} \to \infty$, so both $\beta > \alpha$ and $\alpha < 5/9$ must be true. Specifically, in our simulations, we chose $(E, \alpha, F, \beta) = (800, 4/5, 150, 1)$ for S1, and $(E, \alpha, F, \beta) = (37.1, 1/20, 150, 1)$ for S2.

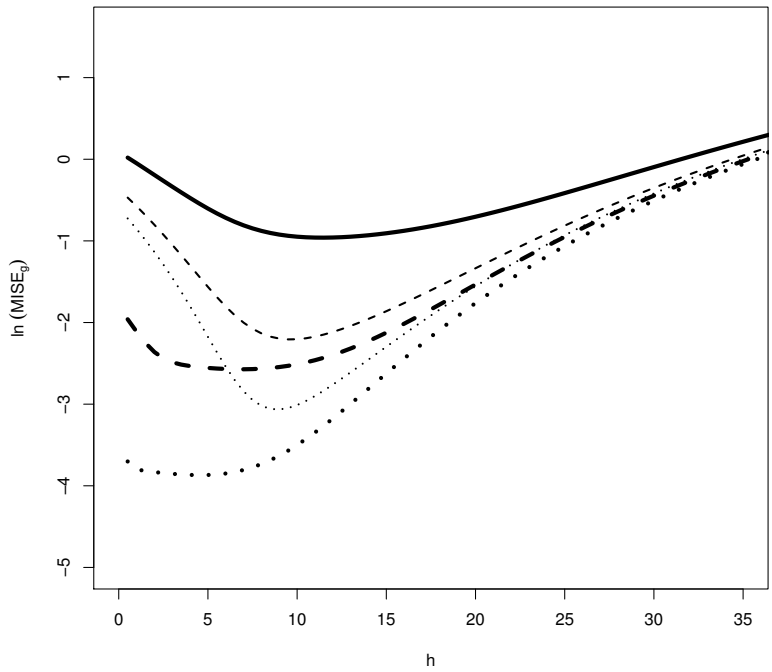


Figure 1: $\ln\left(\text{MISE}_g\right)$ *curves by scenario and sample size. Solid lines are for* $n = 60$*, dashed lines for* $n = 240$ *and dotted lines for* $n = 960$*. Thick lines represent curves in S1, while thin lines represent curves in S2 (note that curves for* $n = 60$ *are practically the same in both scenarios).*

Firstly, for each sample size and each scenario, MISE_g was approximated over a grid of values of h. Figure 1 shows the MISE_g curves, as a function of h, for the three

different sample sizes in both scenarios. Note that a logarithmic scale was used in the vertical axis in order to better appreciate minimal values. This is because very small differences were found in the MISE_g curves for small values of h, particularly for the largest sample size. This suggests that even in the case of grouped data, small deviations from the optimal bandwidth may still give quite good distribution estimates, making the distribution estimation relatively resistant to grouping effects. On the other hand, it is important to note in Figure 1 that MISE_g decreases as the sample size increases, which seems to confirm consistency of the estimator defined in (5). It is also clear that optimal bandwidths for S2 are larger than for S1.

Next, a second simulation experiment was performed to analyse the behaviour of the plug-in bandwidth (15) and the bootstrap selector (17). We compared the sampling distribution of $\hat{h}_{PB_g}$ and h^*_{MISE} with respect to the target values of the bandwidths minimizing MISE_g, denoted by h_{MISE_g}, for each sample size and scenario. The process performed was the following:

1. Simulate an n-size sample from the normal mixture reference density f.
2. Divide the data range into intervals $[y_{i-1}, y_i)$ of length l_i (according to the previous guidelines).
3. Considering the interval midpoints, estimate $A(f')$ by means of $\hat{A}_{PB_g}$ and calculate $\hat{h}_{PB_g}$ using (15).
4. Select ζ as described in Section 4.2 and approximate h^*_{MISE}.
5. Compute $\hat{h}_{PB_g}/h_{MISE_g}$ and h^*_{MISE}/h_{MISE_g}.
6. Repeat Steps 1 to 5 $B = 1000$ times.

Figure 2 shows the results as box-plots. Regarding $\hat{h}_{PB_g}$ (yellow left box-plots for each sample size), it can be observed that starting from the same grouping conditions and sample size, the sampling distribution gets more precise as the sample size increases under both scenarios, S1 and S2. However, in both situations $\hat{h}_{PB_g}$ is far from the target value. In S1, when the sample size increases, although the sampling distribution gets more accurate, the plug-in bandwidths seem to be in general excessively large. In S2, we observe the same pattern, but now the bandwidths become too small for large sample sizes. This biased performance of $\hat{h}_{PB_g}$ may be due, mainly, to two factors. On the one hand, the remaining terms of the bias of (5), depending on $\bar{l}$, do not vanish as fast as required for (15) to be a good bandwidth selector. On the other hand, the method used here (see Reyes et al., 2017) to select the pilot bandwidth, η, requires estimating $A(f')$. This is done by canceling the sum of the two main bias terms of $\hat{\psi}^g_{\eta,2}$. This could be not able to produce good pilot smoothing parameters because, opposite to the complete data case, some second order terms depending on $\bar{l}$ could have a significant impact on the MSE of $\hat{\psi}^g_{\eta,2}$.

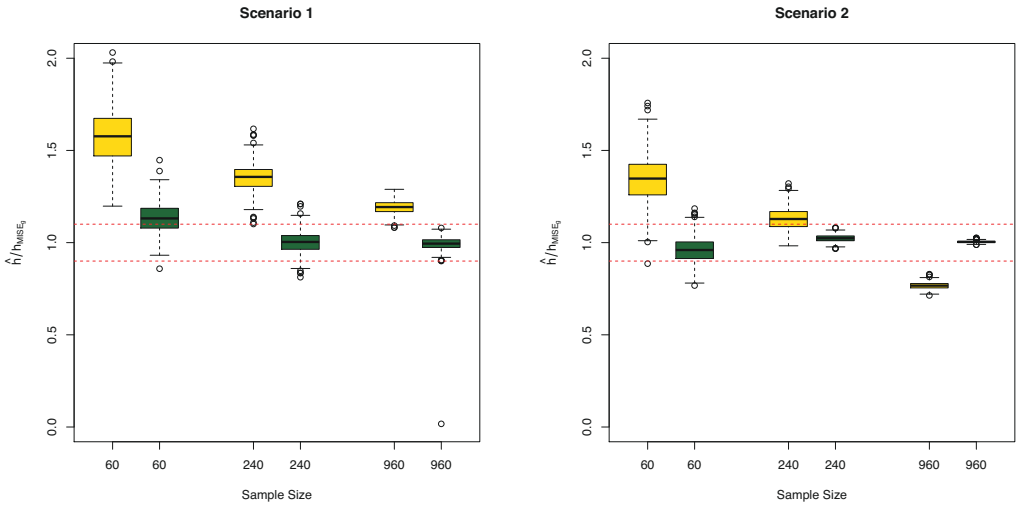


Figure 2: *Box-plots for $\hat{h}_{PB_g}/h_{MISE_g}$ (yellow left box-plots for each sample size) and box-plots for h^*_{MISE}/h_{MISE_g} (green right box-plots for each sample size) for both scenarios. Red dotted lines are plotted at values 0.9 and 1.1 for reference.*

Regarding Figure 2, it can be observed that h^*_{MISE} (green right box-plots for each sample size) outperforms $\hat{h}_{PB_g}$ in approaching h_{MISE_g}. The bootstrap selector shows more stability under any sample size and scenario. This means that it is preferable in both cases of light or heavy grouping and for any sample size.

Figure 3 shows the effect on the distribution estimator (5) of the bandwidth selectors (15) and (17), respectively, in both scenarios. Clearly, when using $\hat{h}_{PB_g}$ (yellow left box-plots for each sample size), while in S1 the quality of $\hat{F}^g_h(x)$ gets better as the sample size increases, in S2, poor distribution estimates for large n are obtained. The impact of poor bandwidth selection is evident in the quality of the distribution estimator, whose error increases by up to three times. However, it should be noted that it does not impact so negatively in the corresponding estimates as in the case of density estimation for grouped data (see Reyes et al., 2017). In opposition, when using the bootstrap bandwidth (green right box-plots for each sample size), the quality of the distribution estimates improves as the sample size increases in both scenarios, clearly outperforming the plug-in bandwidth selector.

It is of interest to study situations in which it is ideally observed the sample size increasing and the average length decreasing at different rates, but in practice that seldom really occurs. For that reason, we also performed some simulations (not shown here for reasons of space, but included in the Supplementary Materials) dealing with a more factual situation in which there is a given sample size and a given set of fixed intervals. In that simulation, a sample size of $n = 240$, a fixed set of average lengths, $\bar{l}$, and a grid of values for h were considered. Those experiments show that the bootstrap smoothing parameter, h^*_{MISE}, seems to be very stable, always centred somewhere around h_{MISE_g}

and with decreasing variability when the average length $\bar{l}$ increases. On the other hand, the plug-in selectors are larger than the target value for small or moderate values of $\bar{l}$, and smaller than the optimal bandwidth for large values of $\bar{l}$. However, as pointed out previously, it was also observed that bandwidth selection is not so critical in distribution as in density estimation, since slightly different bandwidths produced very similar distribution estimates.

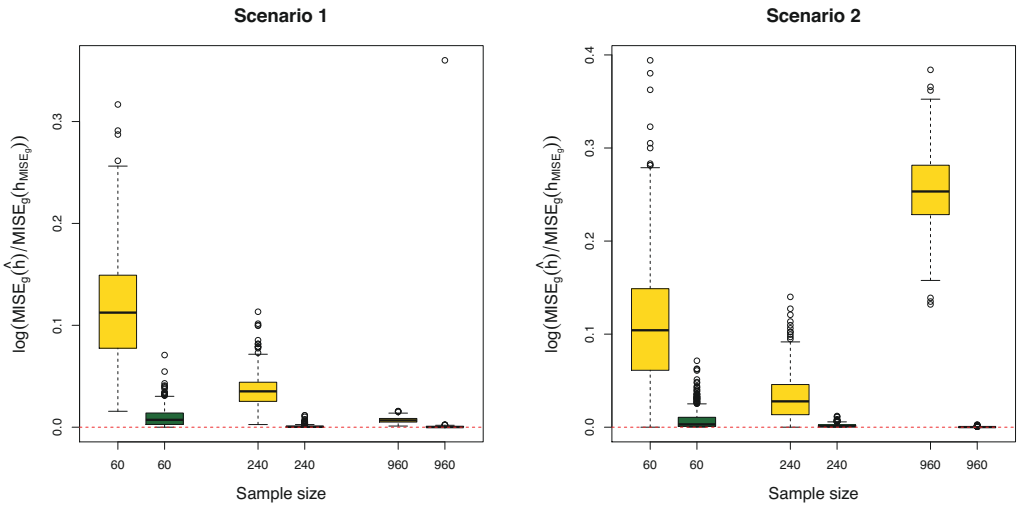


Figure 3: *Box-plots for* $\ln\left[\mathrm{MISE}_g\left(\hat{h}_{PB_g}\right)/\mathrm{MISE}_g\left(h_{MISE_g}\right)\right]$ *(yellow left box-plots for each sample size) and* $\ln\left[\mathrm{MISE}_g\left(h^*_{MISE}\right)/\mathrm{MISE}_g\left(h_{MISE_g}\right)\right]$ *(green right box-plots for each sample size) for both scenarios.*

6 An empirical study from real data

In this section, a real data set of wild oat (*Avena sterilis* L.) emergence is considered to illustrate the performance of the kernel distribution estimator for grouped data (5), when using the plug-in (15) and bootstrap (17) bandwidth selectors. To do this, the `binnednp` `R` package (Barreiro et al., 2019) is employed. This package, developed by the authors of the present paper, jointly with a weed scientist and two computer engineers, contains some functions implementing most of the nonparametric methods for grouped data (and related problems), studied by the authors in this and in previous papers.

The data of *Avena sterilis* were taken from an experiment performed during Winter-Spring 2006-2007 in Gibraleon (37º 22'N, 6º 54'W; altitude 26 m), located in the province of Huelva (Andalucia, South of Spain). Four polyvinylchloride cylinders (250 mm diameter 50 mm height) placed 1 m apart were considered and, for each one of them, 200 seeds of *A. sterilis* were mixed thoroughly with the soil and distributed over the 0-100 mm depth. Numbers of emerged weed seedlings were recorded once or twice

a week and then removed by cutting seedling stems at ground level with minimum disturbance of the substrate. All the data for the cumulative numbers of seedling emergence from the field were converted to a square meter basis. The CHTT at emergence in the different inspection days, at three depths (10, 20 and 50 mm), were calculated, using the same methodology as that described in Cao et al. (2011).

The observed emergence data are shown in Table 1. As it can be seen, the cumulative hydrothermal time at emergence can not be observed for every individual seed, but just in an aggregated way.

Table 1: *Seedling emergence data of A. sterilis.*

	CHTT			**Nº Seedlings**				
	Depth			**Cylinder**				
Date	**10 mm**	**20 mm**	**50 mm**	**1**	**2**	**3**	**4**	**Pooled**
27 November 2006	100	92	67	0	0	0	0	0
4 December 2006	160	146	105	0	0	0	0	0
12 December 2006	218	199	143	2	6	8	3	19
14 December 2006	218	217	155	1	0	0	1	2
19 December 2006	218	217	185	2	1	1	3	7
22 December 2006	218	217	199	2	1	1	0	4
26 December 2006	218	217	204	1	1	0	0	2
28 December 2006	218	217	204	0	0	0	0	0
2 January 2007	218	217	204	0	0	0	0	0
5 January 2007	218	217	204	0	2	0	0	2
9 January 2007	218	217	204	2	2	9	2	15
12 January 2007	218	217	204	3	7	18	11	39
18 January 2007	218	217	204	12	7	19	22	60
25 January 2007	218	217	204	6	5	8	13	32
1 February 2007	265	261	232	2	5	7	7	21
9 February 2007	352	340	287	13	12	5	8	38
15 February 2007	405	421	343	7	12	13	4	36
23 February 2007	459	505	421	0	0	1	0	1
5 March 2007	509	571	538	0	0	0	0	0
19 March 2007	509	571	538	0	0	0	0	0
Nº Emerged seedlings				53	61	90	74	$n = 278$

Before computing the kernel distribution estimator (5) to obtain approximations of the corresponding weed emergence curves, some preliminary analyses were performed. Firstly, the function `anv.binned` included in the `binnednp` package was employed to test whether the "cylinder factor" does not have a significant effect on the emergence curve. If so, we could considered the pooled sample of the four cylinders. In the function `anv.binned`, a bootstrap approach using a Cramér-von Mises type distance is implemented to carry out this type of hypothesis testing. The experimentation conditions seem to support the idea of having a "non-significant cylinder effect" and, after applying

the function anv.binned, this hypothesis was corroborated for the three depths. Secondly, an interesting issue for the weed researchers is to find out what the best soil depth is among the three possibilities available in this case, 10, 20 and 50 mm, to measure the CHTT in order to have more prediction power. To address this problem, moment-based indices and probability density-based indices were proposed in Cao et al. (2011). Estimates of these indices are implemented in the function emergence.indices of the binnednp package. After applying this function to the pooled sample of *Avena sterilis*, it is concluded that the best soil depth to measure the CHTT is 10 mm. Therefore, in what follows, only the CHTT measured at 10 mm and the pooled sample are considered for the subsequent analyses.

After these previous analyses, the emergence curve of *Avena sterilis*, using the CHTT measured at 10 mm and the pooled sample, is estimated computing the kernel distribution estimator (5). To do this, we used the functions bw.dist.binned and bw.dist.binned.boot of the binnednp package, returning the plug-in (15) and bootstrap (17) bandwidths, respectively. Arguments in these functions allow to control, among other things, the pilot bandwidths needed in both selectors. For example, in the case of the plug-in bandwidth, $\hat{h}_{PB_g}$, in bw.dist.binned, different types of models can be used in the last step of the iterative method explained in Section 4.1: assuming a normal distribution, using a complete nonparametric approach or considering a normal mixture model. In the case of the bootstrap bandwidth, h^*_{MISE}, in bw.dist.binned.boot, the user can employ as a pilot bandwidth that selected using the method inspired by the idea of smoothing splines, described in Section 4.2, or the one derived using the plug-in technique, $\hat{h}_{PB_g}$. The default pilot bandwidths in these functions are those described in Sections 4.1 and 4.2, respectively. Other parameters of bw.dist.binned and bw.dist.binned.boot allow to plot the corresponding nonparametric distribution estimators and to compute bootstrap confidence bands for the distribution function. It is important to highlight that the functions of this library have been efficiently programmed, using integration of C++ in the R code, and applying parallel computing methods to speed up the running time of the algorithms. This is especially important in those methods making use of bootstrapping to obtain numerical results in a very short time.

Using the default pilot bandwidths, the plug-in and bootstrap smoothing parameters obtained are, respectively, 9.83 and 13.74. The corresponding kernel distribution estimates of the emergence curves computed using (5) are shown in the left panel of Figure 4 (in green when using the plug-in bandwidth and in red when using the bootstrap bandwidth). The empirical distribution of the grouped sample (black line) is also shown in this plot. As indicated in the previous section, it can be seen that the effect of the bandwidth on the estimator's behaviour is not substantial, since slightly different bandwidths produce very similar distribution estimates.

As pointed out in Section 1, parametric regression models have been widely used to model the relationship between the CHTT and weed emergence. For the sake of comparison, the function bw.dist.binned also allows to fit Weibull and logistic parametric regression functions to describe seedling emergence, with parameters estimated by ma-

ximum likelihood. The corresponding fits using the *Avena sterilis* data set are shown in the right panel of Figure 4, using a green line for the Weibull and a blue line for the logistic estimators. The nonparametric distribution estimator (5) with bootstrap bandwidth (red line) and the empirical distribution of the grouped sample data (black line) are also included in this plot. It can be observed that none of both classical parametric (distribution) models fits the data well, possibly leading to wrong emergence estimations. On the other hand, the nonparametric approach does not assume any particular distribution for the variable under consideration. As a consequence, it provides more flexible estimators capable of capturing complex features in the HTT distribution and producing more reliable results.

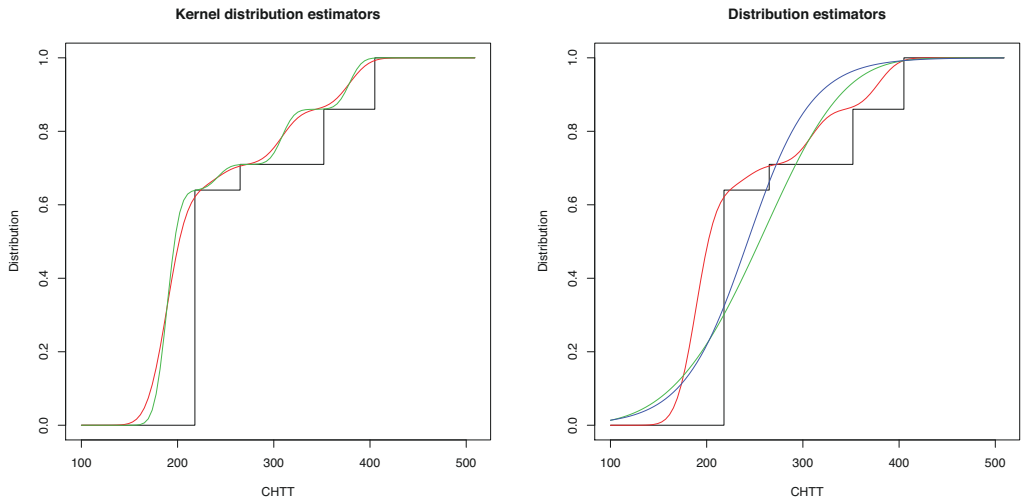


Figure 4: *Left panel: Kernel distribution estimates considering plug-in (green line) and bootstrap (red line) bandwidths. Right panel: parametric regression fits, Weibull (green line) and logistic (blue line), and nonparametric kernel distribution estimate using the bootstrap bandwidth (red line). The empirical distribution of the grouped sample (black lines) is also shown.*

7 Conclusions

In short, it has been shown that under realistic assumptions, the kernel distribution estimator is an effective tool for modeling grouped data due to the good performance of the bootstrap smoothing parameter selector proposed in this paper. This bandwidth selector, using an appropriate criterion to select the corresponding pilot bandwidth, presents a stable and unbiased sampling distribution under any scenario or sample size in the simulation studies performed. Regarding the Polansky and Baker plug-in bandwidth, although theoretically it is a consistent estimator of the optimal bandwidth, in practice, it only has an appropriate behaviour when there is a fixed sample size and a given set

of intervals for certain degree of grouping. This can be due to the fact that this plug-in bandwidth is focused on minimizing the AMISE and some neglected terms of less order, depending on $\bar{l}$, can have a substantial influence under certain grouping conditions (see Remark 3.2). Something similar could occur in the process of selecting the pilot bandwidth needed to estimate $A(f')$. On the other hand, h^*_{MISE} targets directly the MISE, producing much better results.

In any case, the different simulations performed show that the kernel distribution estimator is a somewhat robust procedure, in the sense that bandwidth selections slightly different from the optimal bandwidth do not seem to heavily influence the distribution estimation. From another viewpoint, it was shown that really high values of the ratio between the average length $\bar{l}$ and the data range have to be considered in order to actually notice a severe impact of the grouping effect.

These findings leave some insights about kernel distribution estimation for grouped data as well as some possible future work. Since distribution estimation seems to be resistant to grouping effect, a possible future topic of research could be the design of a plug-in bandwidth selector that could work well in different grouping scenarios. This would imply to find out the real influence of second-order terms in the MISE of $\hat{F}^g_h(x)$ and somehow incorporate these effects in the plug-in bandwidth expression. Moreover, a deeper study about the pilot bandwidth selection problem to estimate $A(f')$ would also be necessary. These two issues would transform the usual simple plug-in bandwidth selection method in a much more complicated problem. Fortunately, the bootstrap bandwidth approach proposed in this paper provides a selector that covers any case of grouping, thus controlling or reducing the increase of the error of the estimates. Moreover, it is important to note that this bootstrap procedure does not need Monte Carlo and, therefore, it is also an efficient computing time approach. Facing applications, this implies a substantial improvement in the estimation of data structure, allowing smart inferences even when data are heavily grouped.

Appendix A. Proof of Theorem 3.1

Proof Applying the expectation operator to (5), it is easy to prove that

$$\mathbb{E}\left[\hat{F}^g_h(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) \mathbb{E}[w_i] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) p_i \tag{A.1}$$

where $p_i = F(y_i) - F(y_{i-1})$.

Using a Taylor expansion of p_i around t_i and substituting into (A.1), and the fact that

$$\alpha_{ji} = \left(\frac{l_i}{2}\right)^j - \left(-\frac{l_i}{2}\right)^j = \begin{cases} 0 & \text{for j even} \\ 2\left(\frac{l_i}{2}\right)^j & \text{else} \end{cases}, \tag{A.2}$$

gives

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} l_i H_1(t_i) + \frac{1}{24}\sum_{i=1}^{k} l_i^3 H_2(t_i) + \frac{1}{4!}\sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right)\delta, \tag{A.3}$$

where $\delta = F^{(4)}(\xi_i)\left(\frac{l_i}{2}\right)^4 - F^{(4)}(\xi_{i-1})\left(-\frac{l_i}{2}\right)^4$, with $\xi_i \in [t_i, y_i]$ and $\xi_{i-1} \in [y_{i-1}, t_i]$, $H_1(t) = F'(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$ and $H_2(t) = F'''(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$. As long as $F^{(4)}$ is Lipschitz, δ can be easily bounded leading to $\left|\sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right)\delta\right| = O\left(\bar{l}^4\right)$, so that (A.3) becomes

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} l_i H_1(t_i) + \frac{1}{24}\sum_{i=1}^{k} l_i^3 H_2(t_i) + O\left(\bar{l}^4\right). \tag{A.4}$$

Considering the first term on the right hand side of (A.4), taking the integral over the i-th interval, using a Taylor expansion with $s = t - t_i$, by (A.2) and summing over all k intervals gives

$$\sum_{i=1}^{k} l_i H_1(t_i) = \int H_1(t)\,dt - \frac{1}{24}\sum_{i=1}^{k} l_i^3 H_1''(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(4)}(\xi_i)(t-t_i)^4\,dt. \tag{A.5}$$

Bounding the third term on the right hand side of (A.5) gives

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(4)}(\xi_i)(t-t_i)^4\,dt\right| = O\left(\frac{\bar{l}^4}{h^4}\right). \tag{A.6}$$

Now, working on the second term on the right hand side of (A.5), it can be expressed as

$$\sum_{i=1}^{k} l_i^3 H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) + \overline{l^2}\sum_{i=1}^{k} l_i H_1''(t_i). \tag{A.7}$$

The second term on the right hand side of (A.7) can be expressed as

$$\overline{l^2}\sum_{i=1}^{k} l_i H_1''(t_i) =$$

$$\overline{l^2}\int H_1''(t)\,dt - \frac{\overline{l^2}}{24}\sum_{i=1}^{k} l_i^3 H_1^{(4)}(t_i) - \frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)(t-t_i)^4\,dt. \tag{A.8}$$

Bounding the third and second terms on the right hand side of (A.8) gives

$$\left|\frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)(t-t_i)^4\,dt\right| = O\left(\frac{\bar{l}^6}{h^6}\right)$$

and $\left|\sum_{i=1}^{k} l_i^3 H_1^{(4)}(t_i)\right| \leqslant O\left(\frac{\bar{l}^4}{h^4}\right)$. In turn, bounding $\overline{l^2}\int H_1''(t)\,dt$ results in

$$\left|\overline{l^2}\int H_1''(t)\,dt\right| \leqslant O\left(\bar{l}^2\right)\int \left|H_1''(t)\right|dt. \tag{A.9}$$

By Assumption 3.1, it is easy to check that $H_1''(t) = 0$ when $\frac{x-t}{h} < -1$, and $H_1''(t) = F'''(t)$ when $\frac{x-t}{h} < -1$. As a consequence,

$$\begin{aligned}\int_{-\infty}^{\infty}\left|H_1''(t)\right|dt &= \int_{-\infty}^{x-h}\left|F'''(t)\right|dt + \int_{x-h}^{x+h}\left|\mathbb{K}\left(\frac{x-t}{h}\right)F'''(t)\right. \\ &- \left.\frac{1}{h}2F''(t)K\left(\frac{x-t}{h}\right) + \frac{1}{h^2}F'(t)K'\left(\frac{x-t}{h}\right)\right|dt. \end{aligned}\tag{A.10}$$

Hence, solving and bounding the right hand side of (A.10) gives

$$\int_{-\infty}^{\infty}\left|H_1''(t)\right|dt = O\left(\frac{1}{h}\right),$$

which implies that

$$\left|\overline{l^2}\int H_1''(t)\,dt\right| = O\left(\frac{\bar{l}^2}{h}\right).$$

Updating (A.7), gives

$$\sum_{i=1}^{k} l_i^3 H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) + O\left(\frac{\bar{l}^2}{h}\right). \tag{A.11}$$

For bounding the first term on the right hand side of (A.11), realize that by previous elaborations,

$$\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1''(t)\,dt -$$

$$\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)(t-t_i)^4\,dt. \tag{A.12}$$

Under Assumption 3.4, the last two terms can be bounded as

$$\left|\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i)\right| \leqslant \max_i\left|l_i^2 - \overline{l^2}\right| k l_{max}^3 \left\|H_1^{(4)}\right\|_\infty = o\left(\frac{\bar{l}^4}{h^4}\right) \tag{A.13}$$

and

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1^{(6)}(\xi_i)(t-t_i)^4\,dt\right| \leqslant$$

$$\frac{1}{4!80}\max_i\left|l_i^2-\overline{l^2}\right|kl_{max}^5\left\|H_1^{(6)}\right\|_\infty = o\left(\frac{\overline{l}^6}{h^6}\right), \tag{A.14}$$

and $\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt$ as

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| &\leqslant \max_i\left|l_i^2-\overline{l^2}\right|\sum_{i=1}^{k}\left|\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| \\ &\leqslant o\left(\overline{l}^2\right)\int\left|H_1''(t)\right|dt. \end{aligned}\tag{A.15}$$

Using the same arguments as above,

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| = o\left(\frac{\overline{l}^2}{h}\right). \tag{A.16}$$

Considering (A.12), (A.13), (A.14), (A.15) and (A.16),

$$\sum_{i=1}^{k}l_i^3H_1''(t) = O\left(\frac{\overline{l}^2}{h}\right), \tag{A.17}$$

thus leading to

$$\sum_{i=1}^{k}l_iH_1(t_i) = \int H_1(t)\,dt + O\left(\frac{\overline{l}^2}{h}\right). \tag{A.18}$$

Integrating by parts, a change of variable, using a Taylor expansion on F and by kernel properties, lead to

$$\sum_{i=1}^{k}l_iH_1(t_i) = F(x)+\frac{h^2}{2}F''(x)\mu_2(K)+O\left(h^4\right)+O\left(\frac{\overline{l}^2}{h}\right). \tag{A.19}$$

Regarding the second term on the right hand side of (A.4),

$$\sum_{i=1}^{k}l_i^3H_2(t_i) = \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_iH_2(t_i)+\overline{l^2}\sum_{i=1}^{k}l_iH_2(t_i). \tag{A.20}$$

Proceeding as above,

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_iH_2\left(t_i\right)\right|=o\left(\bar{l}^2\right). \tag{A.21}$$

As to the second term, note that by Ostrowski's inequality (Anastasiou, Kechriniotis, and Kotsos, 2006; Ostrowski, 1938)

$$\left|l_iH_2\left(t_i\right)-\int_{y_{i-1}}^{y_i}H_2\left(t\right)dt\right|\leqslant\frac{1}{4}\mathfrak{L}_{H_2}l_i^2,$$

where $\mathfrak{L}_{H_2}$ is the H_2 Lipschitz constant. Summing up over all k intervals and considering Assumption (3.4) lead to

$$\sum_{i=1}^{k}\left|l_iH_2\left(t_i\right)-\int_{y_{i-1}}^{y_i}H_2\left(t\right)dt\right|=O\left(\frac{\bar{l}}{h}\right)$$

which in turn implies that

$$\overline{l^2}\sum_{i=1}^{k}l_iH_2\left(t_i\right)=\bar{l}^2\int H_2\left(t\right)dt+o\left(\bar{l}^2\right).$$

Integrating by parts, a change of variable, by a Taylor expansion on F'' and simplifying due to the kernel K properties lead to

$$\int H_2\left(t\right)dt=F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right),$$

so that

$$\overline{l^2}\sum_{i=1}^{k}l_iH_2\left(t_i\right)=\bar{l}^2\left[F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right)\right]+o\left(\bar{l}^2\right). \tag{A.22}$$

Using (A.22) and (A.21), Eq. (A.20) becomes

$$\sum_{i=1}^{k}l_i^3H_2\left(t_i\right)=\bar{l}^2\left[F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right)\right]+o\left(\bar{l}^2\right). \tag{A.23}$$

So, joining (A.23) and (A.19) into (A.3), and by Assumption 3.4,

$$\mathbb{E}\left[\hat{F}_h^g\left(x\right)\right]=F\left(x\right)+\frac{h^2}{2}F''\left(x\right)\mu_2\left(K\right)+o\left(h^2\right),$$

from which, the bias is

$$\mathbb{B}\left[\hat{F}_h^g(x)\right] = \frac{1}{2}h^2 F''(x)\,\mu_2(K) + o\left(h^2\right). \tag{A.24}$$

Regarding the variance, considering that $(n_1, n_2, \ldots, n_k)$ is a multinomial random vector and $w_i = n_i/n$, applying this operator to (5), it gives

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) - \frac{2}{n}\sum_{i<j}\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)p_i p_j. \tag{A.25}$$

Since $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i and by (A.2), the first term on the right hand side of (A.25) (except a factor $1/n$) can be written as

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) = \sum_{i=1}^{k} l_i H_3(t_i) + O\left(\bar{l}\right), \tag{A.26}$$

where $H_3(t) = \mathbb{K}^2\left(\frac{x-t}{h}\right)F'(t)$. Integrating H_3 over the i-th interval, by a Taylor expansion, using $s = t - t_i$, by parity conditions (A.2), summing over all k intervals and reordering gives

$$\sum_{i=1}^{k} l_i H_3(t_i) = \int H_3(t)\,dt - \frac{1}{24}\sum_{i=1}^{k} l_i^3 H_3''(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_3^{(4)}(\xi_i)(t-t_i)^4\,dt. \tag{A.27}$$

As done for (A.5), it is easy to check that

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_3^{(4)}(\xi_i)(t-t_i)^4\,dt\right| = O\left(\frac{\bar{l}^4}{h^4}\right) \tag{A.28}$$

and

$$\sum_{i=1}^{k} l_i^3 H_3''(t_i) = O\left(\frac{\bar{l}^2}{h}\right). \tag{A.29}$$

Considering (A.29), (A.28) and (A.27), Eq. (A.26) transforms into

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) = \int H_3(t)\,dt + O\left(\bar{l}\right). \tag{A.30}$$

As above, using integration by parts, the change of variable $u = (x-t)/h$ and a Taylor expansion give

$$\int H_3(t)\,dt = F(x) - hF'(x)C_0 + O\left(h^2\right),$$

where $C_0 = 2\int \mathbb{K}(u) K(u) u du$. Substituting the last expression into (A.30) and by Assumption 3.4 it gives

$$\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) p_i(1-p_i) = F(x) - hF'(x) C_0 + O\left(h^2\right). \tag{A.31}$$

Let us turn back to eq. (A.25). Because $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i, by (A.2), the second term on the right hand side of (A.25) (except a factor $-2/n$) can be written as

$$\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right) \mathbb{K}\left(\frac{x-t_j}{h}\right) p_i p_j = \sum_{i<j} H_4(t_i,t_j)\, l_i l_j + O\left(\bar{l}^2\right), \tag{A.32}$$

where $H_4(z_1,z_2) = \mathbb{K}\left(\frac{x-z_1}{h}\right) \mathbb{K}\left(\frac{x-z_2}{h}\right) F'(z_1) F'(z_2)$.

Considering the second order Taylor expansion around (t_i,t_j) and by parity conditions (A.2),

$$\int_{y_{i-1}}^{y_i} \int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\, dz_2 dz_1 \quad = \quad H_4(t_i,t_j)\, l_i l_j + \frac{\mathfrak{T}_0}{2}, \tag{A.33}$$

where

$$\begin{aligned} \mathfrak{T}_0 \quad = \quad & \int_{y_{i-1}}^{y_i} \int_{y_{j-1}}^{y_j} \left[\frac{\partial^2 H_4}{\partial z_1^2}(\xi_1,\xi_2)(z_1-t_i)^2 + 2\frac{\partial^2 H_4}{\partial z_1 \partial z_2}(\xi_1,\xi_2)(z_1-t_i)(z_2-t_j) \right. \\ & \left. + \frac{\partial^2 H_4}{\partial z_2^2}(\xi_1,\xi_2)(z_2-t_j)^2 \right] dz_2 dz_1. \end{aligned}$$

Summing over all $k(k-1)/2$ terms of the form (A.33) and reordering,

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) \quad = \quad \sum_{i<j} \int_{y_{i-1}}^{y_i} \int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\, dz_2 dz_1 - \frac{1}{2} \sum_{i<j} \mathfrak{T}_0. \tag{A.34}$$

The second term on the right hand side of (A.34) can be easily bounded by Assumption 3.4, since $\left|\frac{1}{2}\sum_{i<j} \mathfrak{T}_0\right| = O\left(\frac{\bar{l}^2}{h^2}\right)$.

As a consequence,

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) = \sum_{i<j} \int_{y_{i-1}}^{y_i} \int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\, dz_2 dz_1 + O\left(\frac{\bar{l}^2}{h^2}\right). \tag{A.35}$$

On the other hand, it is straightforward to prove that

$$\sum_{i<j} \int_{y_{i-1}}^{y_i} \int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\, dz_2 dz_1 = \frac{1}{2} \int\int H_4(z_1,z_2)\, dz_2 dz_1 + O\left(\bar{l}\right). \tag{A.36}$$

Now, using (A.36) and (A.35),

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) = \frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2 dz_1 + O\left(\bar{l}\right) + O\left(\frac{\bar{l}^2}{h^2}\right). \tag{A.37}$$

Integration by parts, two changes of variable $[u_1 = (x-z_1)/h,\, u_2 = (x-z_2)/h]$ and a Taylor expansion around x give

$$\frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2 dz_1 = \frac{1}{2}\left[F^2(x) + O\left(h^2\right)\right] \tag{A.38}$$

so that, considering (A.38), (A.37) and Assumption 3.4, Eq. (A.32) becomes

$$\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right) p_i p_j = \frac{1}{2}F^2(x) + O\left(h^2\right). \tag{A.39}$$

Now, putting back (A.39) and (A.31) in (A.25) and simplifying,

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 + O\left(\frac{h^2}{n}\right). \tag{A.40}$$

Collecting (A.40) and (A.24), one obtains

$$\begin{aligned}\mathrm{MSE}\left[\hat{F}_h^g(x)\right] &= \frac{1}{4}h^4 F''(x)^2\mu_2(K)^2 + \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 \\ &\quad + O\left(\frac{h^2}{n}\right) + o\left(h^4\right).\end{aligned} \tag{A.41}$$

Finally, dealing with the integrated versions of the terms coming up in the proof of (A.41), one can obtain the following asymptotic expression for AMISE,

$$\mathrm{AMISE}\left[\hat{F}_h^g\right] = \frac{1}{4}h^4\mu_2(K)^2 A(f') + \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \frac{h}{n}C_0,$$

which corresponds with just integrating the leading terms in (A.41). ∎

Acknowledgements

The authors thank three anonymous referees and the Editor for numerous useful comments that significantly improved this article. The authors also thank Dr. Fernando Bastida and Dr. José Luis González-Andújar for providing the *Avena sterilis* L. emergence data employed in Section 6.

This research has been supported by MINECO grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva

ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

Altman, N. and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195–214.

Anastasiou, K., Kechriniotis A. and Kotsos, B. (2006). Generalizations of the Ostrowski's inequality. *Journal of Interdisciplinary Mathematics*, 9, 49–60.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68, 326–328.

Barreiro, D., Fraguela, B., Doallo, R., Cao, R., Francisco-Fernandez, M. and Reyes, M. (2019). *binnednp: Nonparametric estimation for interval-grouped data*. https://cran.r-project.org/package=binnednp R package version 0.4.0.

Blower, G. and Kelsall, J. E. (2002). Nonlinear kernel density estimation for binned data: convergence in entropy. *Bernoulli*, 8, 423–449.

Bowman, A., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85, 799–808.

Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via waved block thresholding. *Probability Theory and Related Fields*, 146, 401–433.

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F. and González-Andújar, J. L. (2011). Computing statistical indices for hydrothermal times using weed emergence data. *Journal of Agricultural Science*, 149, 701–712.

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F. and González-Andújar, J. L. (2013). Modeling *Bromus diandrus* seedling emergence using nonparametric estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 64–86.

Coit, D. and Dey, K. (1999). Analysis of grouped data from field-failure reporting systems. *Reliability Engineering & System Safety*, 65, 95–101.

Dutta, S. (2015). Local smoothing for kernel distribution function estimation. *Communications in Statistics, Simulation and Computation*, 44, 878–891.

González-Andújar, J. L., Francisco-Fernández, M., Cao, R., Reyes, M., Urbano, J. M., Forcella, F. and Bastida, F. (2016). A comparative study between nonlinear regression and nonparametric approaches for modeling *Phalaris paradoxa* seedling emergence. *Weed Research*, 56, 367–376.

Guo, S. (2005). Analysing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, 27, 637–652.

Hill, P. (1985). Kernel estimation of a distribution function. *Communications in Statistics, Theory and Methods*, 14, 605–620.

Klein, J. P. and Moeschberger, M. (1997). *Survival Analysis*. New York: Springer Verlag.

Mächler, M. (2017). *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*. https://CRAN.R-project.org/package=nor1mix. R package version 1.2-3.

Mack, Y. (1984). Remarks on some smoothed empirical distribution functions and processes. *Bulletin of Informatics and Cybernetics*, 21, 29–35.

Mclachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Minoiu, C. and Reddy, S. (2009). Estimating poverty and inequality from grouped data: How well do parametric methods perform? *Journal of Income Distribution*, 18, 160–178.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Applications*, 10, 186–190.

Ostrowski, A. (1938). Über die Absolutabweichung einer differenzierbaren Funktion von ihrem Integralmittelwert. *Commentarii Mathematici Helvetici*, 10, 226–227.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pipper, C. and Ritz, C. (2007). Checking the grouped data version of the Cox model for interval-grouped data survival data. *Scandinavian Journal of Statistics*, 34, 405–418.

Polanski, A. and Baker, E. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, 65, 63–80.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012). Nonparametric kernel distribution function estimation with kerdiest: An R package for bandwidth choice and applications. *Journal of Statistical Software*, 50, 1–21. http://www.jstatsoft.org/v50/i08/.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reiss, R. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8, 116–119.

Reyes, M., Francisco-Fernandez, M. and Cao, R. (2016). Nonparametric kernel density estimation for general grouped data. *Journal of Nonparametric Statistics*, 28, 235–249.

Reyes, M., Francisco-Fernández, M. and Cao, R. (2017). Bandwidth selection in kernel density estimation for interval-grouped data. *TEST*, 26, 527–545.

Rizzi, S., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T., Vaupel, J. and Jacobsen, R. (2016). Comparison of non-parametric methods for ungrouping coarsely aggregated data. *BMC Medical Research Methodology*, 16, 59.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27, 832–837.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution function. *Journal of Statistical Planning and Inference*, 35, 65–75.

Scott, D. and Sheather, S. (1985). Kernel density estimation with binned data. *Communications in Statistics, Theory and Methods*, 27, 832–837.

Titterington, D. (1983). Kernel-based density estimation using censored, truncated or grouped data. *Communications in Statistics, Theory and Methods*, 12, 2151–2167.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38, 290–295.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall/CRC.

Wang, B. and Wang, X.-F. (2016). Fitting the generalized lambda distribution to pre-binned data. *Journal of Statistical Computation and Simulation*, 86, 1785–1797.

Wang, B. and Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65, 4–12.

SORT 43 (2) July-December 2019, 289-316 DOI: 10.2436/20.8080.02.89

Detecting outliers in multivariate volatility models: A wavelet procedure*

Aurea Grané[1], Belén Martín-Bagarrán[2] and Helena Veiga[3]

Abstract

It is well known that outliers can affect both the estimation of parameters and volatilities when fitting a univariate GARCH-type model. Similar biases and impacts are expected to be found on correlation dynamics in the context of multivariate time series. We study the impact of outliers on the estimation of correlations when fitting multivariate GARCH models and propose a general detection algorithm based on wavelets, that can be applied to a large class of multivariate volatility models. Its effectiveness is evaluated through a Monte Carlo study before it is applied to real data. The method is both effective and reliable, since it detects very few false outliers.

MSC: 62M10, 91B84, 62G35, 62H10, 65C05.

Keywords: Correlations, multivariate GARCH models, outliers, wavelets.

1 Introduction

The correlation structure of security returns is the keystone of both portfolio allocation and risk management decisions. In the literature, there are several models to estimate correlations, the multivariate GARCH being the most popular class of models. Oil and financial series of returns often exhibit excess of kurtosis that can be caused by large unexpected observations. In the univariate context, some authors tried to capture this excess of kurtosis by estimating volatility models with fat tail distributed errors. However, it was observed that the estimated residuals of these models still registered excess kurtosis (see Baillie and Bollerslev, 1989, Teräsvirta, 1996). Furthermore, it is well known that these observations can affect the estimation of the GARCH parameters

* The authors acknowledge financial support from BRU-IUL, FEDER funds, Spanish Ministry of Economy and Competitiveness (MTM2014-56535-R, MTM2012-36163-C06-03, ECO2015-70331-C2-2-R) Spanish Ministry of Science, Innovation and Universities (PGC2018-096977-B-l00), FCT grant UID/GES/00315/2019 and Junta de Andalucía (FQM-329).

[1] Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Spain. *Corresponding author*. Email: aurea.grane@uc3m.es.

[2] University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, United Kingdom, Belen.Martin@ed.ac.uk.

[3] Instituto Flores de Lemus and BRU-IUL (Instituto Universitario de Lisboa). Email: mhveiga@est-econ.uc3m.es.

Received: October 2018
Accepted: August 2019

(Fox, 1972, Van Dijk, Franses and Lucas, 1999, Verhoeven and McAleer, 2000, Bali and Guirguis, 2007, Charles and Darné, 2014), the tests of conditional homoscedasticity (Carnero, Peña and Ruiz, 2007, Grossi and Laurini, 2009), the out-of-sample volatility forecasts (Ledolter, 1989, Chen and Liu, 1993, Franses and Ghijsels, 1999, Grané and Veiga, 2010, Boudt, Daníelsson and Laurent, 2013), the volatility estimates (Carnero, Peña and Ruiz, 2012, Behmiri and Manera, 2015) and the risk measures (Grané and Veiga, 2014). For a recent survey on the effects of the outliers on the specification, on the parameter estimation, on the volatility estimation and prediction see Hotta and Trucíos (2018). Analogously, when volatilities and correlations are estimated using multivariate GARCH models similar effects might be expected (see, for instance, Boudt and Croux, 2010).

Portfolios are often composed of commodities that can exhibit negative correlations with stock price returns as, for example, oil returns. Oil is a strategic commodity used as an input in all economic activities, therefore, turmoils in the oil market can propagate to stock markets and affect correlations making them quite negative (see Ramos, Martín-Barragán and Veiga, 2015). Therefore, the first objective of this paper is to study the effect of outliers, in particular, isolated level outliers, patches of level outliers and volatility outliers, on the estimated correlations of multivariate GARCH models (see Hotta and Tsay, 2012, Galeano and Peña, 2013, for a resume on the different types of outliers). We focus on the diagonal Baba-Engle-Kraft-Kroner (D-BEKK) by Engle and Kroner (1995), the conditional constant correlation (CCC) model by Bollerslev (1990) and the dynamic conditional correlation (DCC) model by Engle (2002). We have chosen these models because they are often used in empirical works (see Bauwens, Laurent and Rombout, 2006, Silvennoinen and Teräsvirta, 2009, for excelent surveys on these models). Moreover, as a multivariate GARCH model, the D-BEKK is easier to estimate than the full BEKK because it involves less parameters. Yet, the conclusions and the procedures of this work can be extended to other more sophisticated multivariate volatility models. The second aim is to propose a procedure able to detect outliers in multivariate volatility models that is based on the residuals. The detection of outliers may warn the researcher to use more sophisticate models, filter for outliers or to use robust methods as those proposed, for instance, in Muler and Yohai (2008) for univariate GARCH models and Boudt and Croux (2010) for multivariate GARCH models. Robust estimators work well in the case of additive outliers but lose their efficiency when there is an autoregressive high order dependence. Furthermore and regarding the M-estimators, their asymptotic theory is often based on the hypothesis that the errors are homoscedastic, which is not the case when dealing, for instance, with the DCC model. On the other hand, there are M-estimators that do not depend on homoscedastic errors but they are not so efficient. Therefore, our recommendation is to start by applying a detection procedure and in case outliers are detected, the researcher can either filter them, go for a robust estimation method or use a more sophisticated model that can accommodate outliers.

The Monte Carlo study leads us to conclude that outliers affect the estimated correlations and the effect is stronger for the conditional correlation models (CCC and DCC). Second, our detection procedure is very reliable, not only because the percentage of correct detections is quite high, but also because it detects very few false outliers. This property ensures that when one observation is detected as a possible outlier, it is indeed an outlier.

The advantages of our method are several: first, it can be applied to any multivariate volatility model given that the errors follow a known distribution, second it is well suited for detecting several types of outliers, such as isolated single/multiple outliers and patches of outliers; third, the method is easy and quick to apply, which makes it an attractive tool for academic communities and/or practitioners; fourth, it can be applied to a high number of series, and finally, it is reliable since it detects very few false outliers.

The organization of this paper is as follows. In Section 2 we present the volatility models used in the paper and review two particular types of additive outliers. In Section 3 we study the effect of outliers on the estimated correlations via several simulation studies. In Section 4 we present and evaluate the performance of the detection algorithm and we apply it to several daily series of returns in Section 5. Finally, we conclude in Section 6. Additional Monte Carlo experiments are reported in the Appendix.

2 Outliers in multivariate volatility models

Multivariate financial time series of returns exhibit similar patterns to those of univariate series, such as persistent time-varying volatilities. Additionally, they display time-varying correlations that are often modeled by multivariate GARCH models. One advantage of these models is that they are flexible enough to represent the dynamics of the volatilities and correlations. We start this section describing the models we are going to evaluate. Next we present the type of outliers we are going to consider.

2.1 Models under evaluation

The models that we consider have been pioneer in the financial econometrics literature and are often applied empirically to many fields such as volatility spillover transmission, contagion, portfolio management, asset allocation, etc. However, the methodology developed in this paper is not restricted to them.

In particular, the models under evaluation are the diagonal Baba-Engle-Kraft-Kroner (D-BEKK) model defined in Engle and Kroner (1995), the constant conditional correlation (CCC) model by Bollerslev (1990), and the dynamic conditional correlation (DCC) model by Engle (2002).

Let $\{\mathbf{y}_t\}$ be a vector stochastic process with dimension $N \times 1$ such that $E(\mathbf{y}_t) = \mathbf{0}$ and $\mathscr{F}_{t-1}$ is the information set till time $t-1$. We consider that $\mathbf{y}_t = \boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_t = \mathbf{H}_t^{1/2}\boldsymbol{\eta}_t$,

where $\mathbf{H}_t$ is the conditional covariance matrix of $\mathbf{y}_t$ and $\boldsymbol{\eta}_t$ is an iid vector error process such that $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = \mathbf{I}$, the identity matrix of order N. We assume that there is no linear dependence in $\mathbf{y}_t$.

Alternative approaches in the literature propose different models for the dependence of $\mathbf{H}_t$ on past information $\mathscr{F}_{t-1}$. In the D-BEKK, the dependence of $\mathbf{H}_t$ on past information is modeled directly. In contrast, in the CCC and DCC models, first the conditional variances are modeled using univariate specifications and then $\mathbf{H}_t$ is obtained by using these conditional standard deviations together with some specifications of the correlations (constant for CCC and time-varying for DCC).

We now proceed to a more detailed description of these three models.

DIAGONAL BEKK MODEL The D-BEKK of first order is a restricted version of the model defined in Engle and Kroner (1995), where the dependence of $\mathbf{H}_t$ on past information is modeled as follows:

$$\mathbf{H}_t = \mathbf{C}\mathbf{C}^{\mathsf{T}} + \mathbf{A}^{\mathsf{T}}\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}^{\mathsf{T}}\mathbf{A} + \mathbf{B}^{\mathsf{T}}\mathbf{H}_{t-1}\mathbf{B}, \tag{1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ diagonal matrices and $\mathbf{C}$ is a $N \times N$ lower triangular matrix. The D-BEKK is covariance stationary if and only if $a_{ii}^2 + b_{ii}^2 < 1$ for all i, where a_{ii} and b_{ii} are, respectively, the diagonal elements of $\mathbf{A}$ and $\mathbf{B}$. The conditional covariance matrix is positive definite by construction.

CONDITIONAL CORRELATION MODELS The CCC model is given by

$$\mathbf{H}_t = \mathbf{D}_t\mathbf{R}\mathbf{D}_t = \left(\rho_{ij}\sqrt{h_{ii,t}\,h_{jj,t}}\right)_{ij,t},$$

where $\mathbf{D}_t = \mathrm{diag}(h_{11,t}^{1/2}, \ldots h_{NN,t}^{1/2})$. Here $h_{ii,t}$ is defined in a univariate GARCH–type context such as $h_{ii,t} = \alpha_{0i} + \alpha_{1i}\varepsilon_{i,t-1}^2 + \beta_{1i}h_{ii,t-1}$ and $\mathbf{R} = (\rho_{ij})_{1\le i,j\le N}$ is a correlation matrix, that is symmetric and positive definite, with $\rho_{ii} = 1$, $-1 \le \rho_{ij} \le 1$, $\rho_{ij} = \rho_{ji}$ for $i, j = 1, \ldots, N$. If the N conditional variances are positive, since $\mathbf{R}$ is a positive definite matrix, then $\mathbf{H}_t$ is positive definite. The number of parameters to be estimated are $N(N+5)/2$. Furthermore, univariate GARCH models require that $\alpha_{0i} > 0$, $\alpha_{1i} \ge 0$ and $\beta_{1i} \ge 0$ to guarantee positive conditional variances and $\alpha_{1i} + \beta_{1i} < 1$ to enforce stationary (see Duan et al., 2006).

On the other hand, the dynamic conditional correlation model, DCC, by Engle (2002) is defined as

$$\mathbf{H}_t = \mathbf{D}_t\mathbf{R}_t\mathbf{D}_t, \tag{2}$$

with $\mathbf{D}_t$ defined as before and $\mathbf{R}_t = \left(q_{ij,t}/\sqrt{q_{ii,t}\,q_{jj,t}}\right)_{ij,t}$, where $\mathbf{Q}_t = (q_{ij,t})$ is a $N \times N$ symmetric positive definite matrix given by:

$$\mathbf{Q}_t = (1 - \alpha - \beta)\,\bar{\mathbf{Q}} + \alpha\,\mathbf{u}_{t-1}\mathbf{u}_{t-1}^{\mathsf{T}} + \beta\,\mathbf{Q}_{t-1}, \tag{3}$$

where $\mathbf{u}_t = (u_{1,t}, \ldots, u_{N,t})^\top$ with $u_{i,t} = \varepsilon_{i,t}/\sqrt{h_{ii,t}}$, $\bar{\mathbf{Q}}$ is the unconditional variance matrix of $\mathbf{u}_t$ and α and β are non-negative scalar parameters that satisfy $\alpha + \beta < 1$ (see Bauwens et al., 2006).

We now proceed to define the type of outliers we are going to study. Following Hotta and Tsay (2012), we distinguish two type of additive outliers, level and volatility, and propose a simple extension to the multivariate case.

2.2 Additive level outliers

Additive level outliers (ALOs) can be caused by institutional changes or market corrections that do not affect volatility. In this case, the conditional mean equation of the multivariate volatility model becomes:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\omega} \cdot I_T(t) + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &= \mathbf{H}_t^{1/2} \boldsymbol{\eta}_t, \end{aligned} \tag{4}$$

where $\boldsymbol{\eta}_t$ is as before, that is, an iid vector error process such that $E(\boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top) = \mathbf{I}$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)^\top$ is a vector containing the ALOs' sizes and $I_T(t) = 1$ for $t \in T$ and 0 otherwise, representing the presence of ALOs at a given set of times T. ALOs can occur simultaneously at the same instant t or not and their sizes can coincide or not.

Note that the conditional covariance matrix $\mathbf{H}_{t+1}$ depends only on the past information through $\boldsymbol{\varepsilon}_t$ and $\mathbf{H}_t$. Since the effect of the outlier is only in $\mathbf{y}_t$, the conditional covariance matrix will not be affected by this type of outliers. Indeed ALOs only affect the level of the series. This is true for all multivariate GARCH models.

2.3 Additive volatility outliers

Additive volatility outliers (AVOs) affect both the level of the returns and their volatilities (see Carnero et al., 2007, Grané and Veiga, 2010, Hotta and Tsay, 2012, Hotta and Trucíos, 2018). In this context, the conditional mean equation of the multivariate GARCH model becomes:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &= \boldsymbol{\omega} \cdot I_T(t) + \mathbf{H}_t^{1/2} \boldsymbol{\eta}_t, \end{aligned} \tag{5}$$

where $\boldsymbol{\eta}_t$, $\boldsymbol{\omega}$ and $I_T(t) = 1$ are defined as in section 2.2.

In contrast to ALOs, the effect of AVOs in $\mathbf{y}_t$ is through the term $\boldsymbol{\varepsilon}_t$ which indeed affects the conditional covariance matrix $\mathbf{H}_{t+1}$. This means that the values of the returns following the outlier occurrence will also be affected, since the conditional covariance matrix has been modified by the outlier. In order to highlight this behavior we are going to focus in the D-BEKK model.

Let $\{\mathbf{y}_t\}$ be a vector stochastic process following the D-BEKK model described by equation (1) and $\{\mathbf{y}_t^*\}$ a vector stochastic process following the D-BEKK model contaminated with an AVO at time s. Let $\mathbf{H}_t^*$ and $\boldsymbol{\varepsilon}_t^*$ denote the conditional covariance matrix and the vector of errors for the contaminated process $\{\mathbf{y}_t^*\}$, respectively. Using this notation, equation (1) for the process $\{\mathbf{y}_t^*\}$ is:

$$\mathbf{H}_t^* = \mathbf{C}\mathbf{C}^\top + \mathbf{A}^\top \boldsymbol{\varepsilon}_{t-1}^* \boldsymbol{\varepsilon}_{t-1}^{*,\top} \mathbf{A} + \mathbf{B}^\top \mathbf{H}_{t-1}^* \mathbf{B}.$$

At time s, when the outlier occurs, $\boldsymbol{\varepsilon}_s^* = \boldsymbol{\omega} + \boldsymbol{\varepsilon}_s$. Hence, at time $s+1$, the conditional covariance matrix of the process $\{\mathbf{y}_t^*\}$ will get contaminated. That is:

$$\mathbf{H}_{s+1}^* = \mathbf{H}_{s+1} + \mathbf{A}^\top (\boldsymbol{\omega}\boldsymbol{\omega}^\top + \boldsymbol{\omega}\boldsymbol{\varepsilon}_s^\top + \boldsymbol{\varepsilon}_s\boldsymbol{\omega}^\top)\mathbf{A}.$$

Note that, after time $s+1$, the conditional covariance matrix of the process $\{\mathbf{y}_t^*\}$ is different than that of the non–contaminated process $\{\mathbf{y}_t\}$, since it is affected by both the second and the third terms of equation (1). It is easy to see that the third term is affected by the outlier since it ultimately depends on $\mathbf{H}_{t-1}^*$. The second term depends on $\boldsymbol{\varepsilon}_t^*$, whose covariance is actually $\mathbf{H}_t^*$, which is hence different from the non-contaminated vector of errors $\boldsymbol{\varepsilon}_t$.

Regarding, the CCC and DCC models the conditional covariance matrix $\mathbf{H}_t$ is also affected by the AVO, since it depends on the conditional variances obtained with the univariate GARCH models, that are as well affected by the AVO (see Carnero et al., 2007, Grané and Veiga, 2010, Carnero et al., 2012).

3 Effects of outliers on the correlations: Simulation studies

In the univariate literature it is well known that outliers can affect the estimation of parameters and volatility in the context of GARCH models. However, there are still few studies devoted to analyse the effects of these observations on the estimated correlations using multivariate GARCH models. In this section we contribute in this line by implementing simulation studies for Gaussian and Student-t_7 distributed errors.

Experimental conditions All simulation studies involve single, multiple and patches of additive level outliers and additive volatility outliers included in the models described in section 2.3.

The frequency of the simulations is daily and the number of simulated series is $N = 2$. Outliers are placed randomly across the simulated series, but in the same position for each pair of series. We consider that the outlier affects each pair of series at the same instant of time. Each scenario involves 1000 replications and series are simulated from CCC, DCC and D-BEKK(1,1,1) models with either normal or Student-t_7 distributed errors (see the Appendix). The number of replications is selected to provide robust results.

Given a model, we analysed 24 scenarios, that are defined from the type and number of outliers (one isolated ALO, multiple ALOs, patches of three ALOs, one isolated AVO), the size of the outlier ($\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}, 10\boldsymbol{\sigma}_{\mathbf{y}}$ for ALOs and $\boldsymbol{\omega} = 25\boldsymbol{\sigma}_{\mathbf{y}}, 50\boldsymbol{\sigma}_{\mathbf{y}}$ for AVOs) and the sample size of the simulated series ($n = 1000, 3000, 5000$).

GAUSSIAN ERRORS Parameter values were chosen by fitting the models to real time series of financial returns including commodities such as oil. In particular, for the D-BEKK model: $\{\text{vec}(\mathbf{C}) = (0.053, 0.042, 0, 0.020)^{\intercal}, \text{diag}(\mathbf{A}) = (0.161, 0.164)^{\intercal}, \text{diag}(\mathbf{B}) = (0.983, 0.981)^{\intercal}\}$; for the CCC: $\{\boldsymbol{\alpha}_0 = (0.010, 0.013), \boldsymbol{\alpha}_1 = (0.049, 0.067), \boldsymbol{\beta}_1 = (0.940, 0.926), \rho_{12} = -0.606)\}$; and for the DCC model: $\{\boldsymbol{\alpha}_0 = (0.010, 0.013), \boldsymbol{\alpha}_1 = (0.049, 0.067), \boldsymbol{\beta}_1 = (0.940, 0.926), \alpha = 0.015, \beta = 0.981\}$. Symbols $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}_1$ stand for vectors of parameters of the univariate GARCH(1,1) models (see Grané and Veiga, 2010, for the details).

Simulation results are robust to the choice of parameter values and N. Therefore, we use $N = 2$ since it allows presenting results graphically without losing generality.

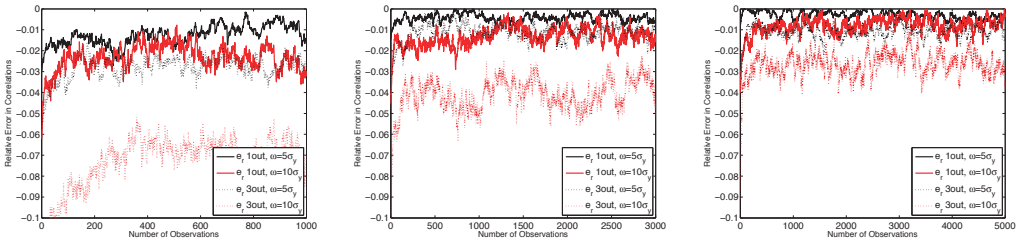
The results of this simulation study are reported in Table 1 (16 top rows) and Figures 1–3.[1]

Table 1: *Relative bias in the estimated correlations obtained from a CCC model from 1000 simulated series of size n that include outliers of different magnitudes.*

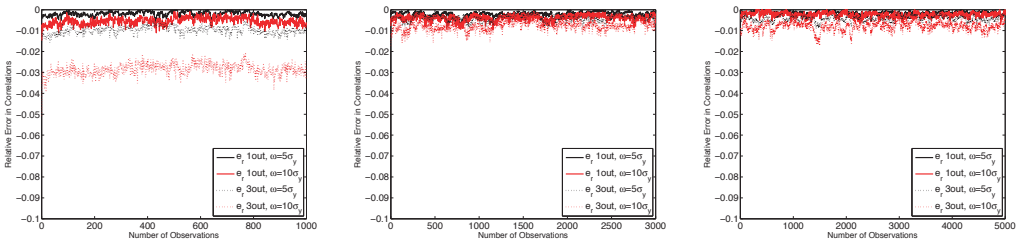
CCC Model with Gaussian errors

	n	Estimated Correlation	Relative Bias		n	Estimated Correlation	Relative Bias
1 ALO	1000	-0.5987	-0.013	3 ALOSs	1000	-0.5892	-0.028
$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.6042	-0.004	$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.6007	-0.010
	5000	-0.6051	-0.002		5000	-0.6017	-0.008
1 ALO	1000	-0.5872	-0.032	3 ALOs	1000	-0.5545	-0.086
$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.5970	-0.016	$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.5810	-0.042
	5000	-0.6012	-0.009		5000	-0.5902	-0.027
Patch of	1000	-0.5972	-0.015	1 AVO	1000	-0.5614	-0.074
3 ALOs	3000	-0.6031	-0.006	$\boldsymbol{\omega} = 25\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.5805	-0.043
$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	5000	-0.6041	-0.004		5000	-0.5847	-0.036
Patch of	1000	-0.5839	-0.037	1 AVO	1000	-0.5318	-0.123
3 ALOs	3000	-0.5959	-0.017	$\boldsymbol{\omega} = 50\boldsymbol{\sigma}_{\mathbf{y}}$	3000	-0.5627	-0.072
$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	5000	-0.5999	-0.011		5000	-0.5642	-0.070
No outliers	1000	-0.6064					
	3000	-0.6065					
	5000	-0.6064					

1. Preliminary results concerning isolated ALOs were presented in an invited conference in the 7th International Workshop on Statistical Simulation (Rimini, 2012) and published in the Proceedings of the workshop (see Grané, Veiga and Martín-Barragán, 2014).

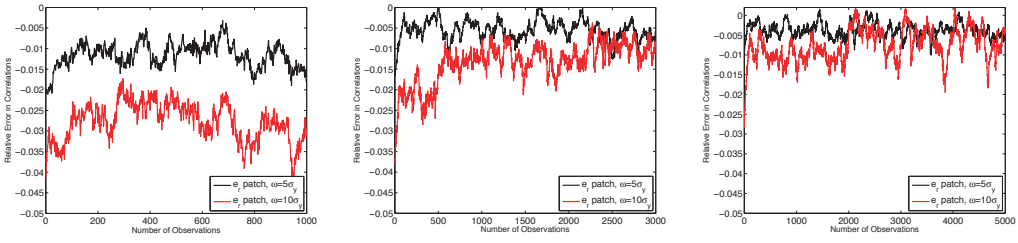


(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

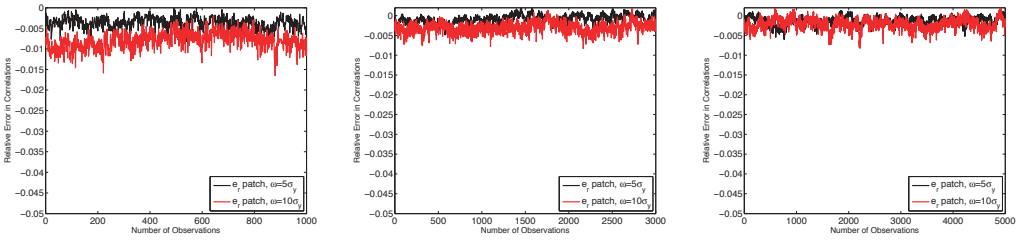


(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure 1: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following normal distributions from 1000 simulated series of size n that include ALOs of different magnitudes.*

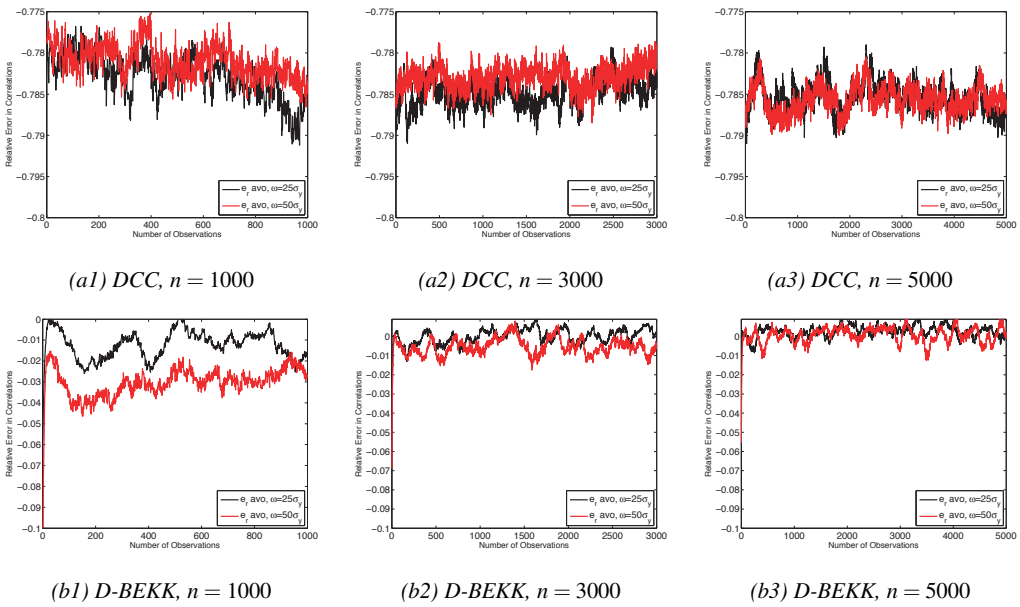


(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*



(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure 2: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following normal distributions from 1000 simulated series of size n that include patches of different magnitudes.*



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure 3: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following normal distributions from 1000 simulated series of size n that include 1 AVO of different magnitudes.*

Table 1 shows the estimated correlations (each reported value is the sample mean computed on 1000 values) for the CCC model and the relative biases (or relative errors) with respect to the estimated correlations in the absence of outliers. Lines 13-15 correspond to the estimated correlations in the absence of outliers. Figures 1–3 contain the relative errors of DCC and D-BEKK models for different sample sizes. In particular, in Figure 1 we plot the relative bias obtained in the estimation of the correlations using DCC and D-BEKK models for the case of isolated ALOs, whereas Figures 2–3 correspond, respectively, to patches of ALOs and 1 isolated AVO. For each time t (going from 1 to n), the plotted value is the sample mean computed from 1000 replications.

From Table 1 and Figures 1–3 we can observe that the estimated correlations are affected by the presence of outliers and the relative errors are higher the higher is the magnitude of the outlier, the larger the number of outliers included in the simulated series and the smaller the sample sizes of the simulated time series. Moreover, the biases in the correlations are higher for the DCC model in comparison to the CCC and D-BEKK models. In particular, the latter seems to be more robust to the presence of outliers since the correlations present small relative errors over the sample size. Finally, another conclusion is that additive outliers (level or volatility) bias the estimated correlations towards zero for the three considered models when the errors are Gaussian.

STUDENT-*t* ERRORS Next we perform simulations assuming that the errors follow a Student-t distribution (see Appendix A.1). From the results of Table A we can conclude

that, for the CCC model and $N = 2$, the biases are of less magnitude (in absolute value) when considering a Student-t distribution. This was expected since the Student-t distribution is more robust to outliers. With respect to DCC and D-BEKK, we also observe that the biases are slightly smaller in absolute value when considering level outliers, isolated or patches. However, when we consider volatility outliers we observe that the impact on the correlations is larger in absolute value.

MORE SERIES Finally, we conduct a third simulation study considering $N = 4$ series (see Appendix A.2). The main conclusions are: Firstly, the impact of the outliers on the correlations tend to decrease for all level outliers; And secondly, regarding the volatility outliers, passing from two to four series leads to a decrease of the impact of outliers for the D-BEKK model, whereas for the DCC model the impact of volatility outliers on the correlations remains almost the same.

4 Wavelet-based detection procedure

Grané and Veiga (2010) proposed a general outlier detection method based on wavelets for the univariate case. This procedure was evaluated through an intensive Monte Carlo study and compared to other existing competitors. The method was proven to be very effective in detecting isolated level outliers, patches of level outliers and volatility outliers in large univariate financial time series. Additionally, the results showed the reliability of the method (in front of other competitors), since it detected a significantly small number of false outliers. More recently, Kamranfar, Chinipardaz and Mansouri (2017) has extended the procedure of Franses and Ghijsels (1999) to allow for level change and temporary change outliers.

The purpose of this work is to extend the method by Grané and Veiga (2010) to a multivariate setting, preserving as much as possible the good properties already proven in the univariate case, in particular, effectiveness and reliability, and also a feasible implementation in large data sets. This will be achieved by applying the random projection method (Cuesta-Albertos, Fraiman and Ransford, 2006, Cuesta-Albertos et al., 2007), that allows to translate a multivariate problem into a univariate context. These authors intuitively describe the random projection method in the following way. Imagine we have to deal with a problem related to d-dimensional objects. The random projection method consists of choosing, at random, a subspace of dimension k (where k is low compared to d), solve the problem in the k-dimensional subspace and translate the solution to the original (d-dimensional) space. In practice, $k = 1, 2$, which is exactly contrary to the Projection Pursuit paradigm, avoiding implementation problems due to high-dimensionality. Random projections have been successfully applied as a simple method for dimensionality reduction in high-dimensional problems, in fields such as computer science, data mining, image processing, etc. (see, for instance, Bingham and Mannila, 2001, Vempala, 2004).

Our procedure is to be applied to the residuals of multivariate GARCH models or other multivariate volatility models such as the stochastic volatility models, although in this paper we focus our attention on the former class. To our knowledge, our procedure is the first to detect outliers in multivariate volatility models.

Next in Section 4.1 we describe the proposed method to detect outliers and evaluate its performance in Section 4.2.

4.1 The procedure

The method we propose is an extension of the procedure by Grané and Veiga (2010), that was based on the detail wavelet coefficients resulting from the discrete wavelet transform (DWT) of a univariate series of (standardized) residuals.

The method requires a preliminary step that consists in fitting a multivariate GARCH model and obtaining the series of multivariate residuals. Note that our proposal is model-dependent, but general enough to cope with a wide variety of models.

In the first step, the multivariate series of residuals is transformed into a univariate series to which DWT will be applied. At this point two possibilities are considered whether the conditional covariance matrix of the fitted model fulfills the decomposition property. In case this property is satisfied, it is enough to consider only the univariate marginals. This will be the case for conditional correlation models, such as CCC and DCC. On the other hand, if the decomposition property is not fulfilled, like happens in the D-BEKK model, in addition to the marginals, we consider one randomly chosen projection (see Cuesta-Albertos et al., 2006). Therefore, we end the first step with a vector containing the univariate marginals or either the univariate marginals plus an extra series containing the result of the random projection. Note that only one projection is needed regardless the number of series considered. In Section 4.3 we show that for $N = 10$ series the procedure is still effective and no advantage is achieved by increasing the number of projections.

In the second step we apply the DWT to each of the univariate series under consideration and in further steps the procedure proceeds by identifying outliers as those observations in the original series whose detail coefficients are greater (in absolute value) than a certain threshold (see more details below). This threshold is a percentile of the distribution of a certain test statistic. The underlying idea for the threshold relies in the fact that in the context of financial return time series it is common to assume an underlying model for the data. Therefore, if the fitted model has captured the structure of the data, the residuals are supposed to be independent and identically distributed random variables following a specified distribution. In particular, the threshold is associated to the following test statistic: the maximum of the detail wavelet coefficients (in absolute value) resulting from the DTW of a univariate series of (standardized) residuals. When the univariate series under consideration is a marginal of the multivariate one, the thresholds given in Grané and Veiga (2010) for the univariate case are still valid. In case the

univariate series is the result of a random projection, the distribution of the test statistic is obtained via Monte Carlo and the threshold is derived analogously.

Table 2: *Threshold values: Percentiles of the distribution of the test statistics (with Bonferroni correction).*

		Gaussian distributed errors				Student's t distributed errors			
		only marginals		marginals and random projection		only marginals		marginals and random projection	
	n	1st level	2nd level	1st level	2nd level	1st level	2nd level	1st level	2nd level
$\alpha = 0.05$	1000	4.0595	3.8827	4.1386	3.9731	5.2583	4.6062	5.2390	4.6399
	3000	4.2995	4.1437	4.3885	4.2383	5.7469	5.0101	5.7214	5.0092
	5000	4.4062	4.2664	4.5027	4.3503	6.0131	5.1269	5.9162	5.1953
$\alpha = 0.10$	1000	3.7216	3.5280	3.9944	3.8207	4.9470	4.3384	4.9215	4.4039
	3000	3.8965	3.7114	4.2319	4.0873	5.4087	4.7086	5.3850	4.7845
	5000	4.2620	4.0992	4.3607	4.2012	5.6332	4.9015	5.6013	4.9383

In practice, we find that in order to detect isolated ALOs it suffices to work with the first level detail wavelet coefficients. However, if there are patches of ALOs or isolated AVOs, it is necessary to use both first level and second level detail wavelet coefficients. From the simulation study (see section 4.2) we believe that a reasonable threshold to use in the detection of isolated ALOs is the 95-th percentile, whereas for the detection of patches of ALOs and isolated AVOs the 90-th percentile is more useful. An analogous situation occurred in the univariate case.

Since in the multivariate case we are considering more than one series, the thresholds proposed in Grané and Veiga (2010) for the univariate case are not directly applicable and the union-intersection principle (Roy, 1953) with Bonferroni correction is applied. As a reference, Table 2 contains the values of the thresholds after applying the Bonferroni correction for bivariate series. The third, fourth, seventh and eighth columns correspond to the thresholds for the case in which only the marginals are considered and the fifth, sixth, ninth and tenth columns, to the case in which both the marginals and a random projection are considered. The thresholds are shown for two different significance levels $\alpha = 0.05$ and 0.10, three different sample sizes $n = 1000, 3000$, and 5000 and two different error distributions.

A brief description of the algorithm

Next we give a brief description of the procedure for detecting ALOs. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be the multivariate series of residuals of size n obtained after fitting a CCC, DCC or a D-BEKK(1,1,1) model with normal distributed errors.

Step 1 In case the fitted model is D-BEKK(1,1,1) (or any other model in which the decomposition property does not apply) consider a random vector $h = (h_1, \ldots, h_N)^{\intercal}$ such that $\|h\| = 1$ and obtain the random projection of the multivariate series of residuals on that direction, that is, $X_{N+1} = h \cdot \mathbf{X}$. Let $\mathbf{X}^* = (X_1, \ldots, X_N, X_{N+1})$.

Step 2 Apply the DWT to each marginal of $\mathbf{X}$ (or alternatively $\mathbf{X}^*$) to obtain the first level wavelet detail coefficients $D_j = (d_{i,j})$, $i = 1,\ldots,n/2$, $j = 1,\ldots,N$ (or alternatively $j = 1,\ldots,N+1$).

Step 3 Set the threshold k^α equal to some percentile of the distribution of the maximum of the first level wavelet detail coefficients (in absolute value) resulting from the DWT of n iid random variables following a standard normal distribution, considering the Bonferroni correction. See Table 2 for some examples and other probability distributions.

Step 4 Find $S_j = \{i : |d_{i,j}| > k^\alpha\}$, for $j = 1,\ldots,N$ (or alternatively $j = 1,\ldots,N+1$) and consider $S = \cup_{j\geq 1} S_j$ the set formed by the union of all the elements in the S_j's.

Step 5 Use S to locate the exact positions of the ALOs in any of the X_j's. Let s be a generic element in S. Let $\overline{x}_{n-2}$ be the sample mean of X_j without observations at locations $2s$ and $2s-1$. Then, set the position of the ALO equal to $2s$ if $|X_{2s,j} - \overline{x}_{n-2}| > |X_{2s-1,j} - \overline{x}_{n-2}|$, or equal to $2s-1$, otherwise.

The algorithms that respectively search for patches of ALOs and for AVOs differ from the previous one in the sense that two level wavelet coefficients are computed, and consequently there are two thresholds, one for each set of detail wavelet coefficients $\mathbf{D}^{(1)} = \cup_{j\geq 1} D_j^{(1)}$ and $\mathbf{D}^{(2)} = \cup_{j\geq 1} D_j^{(2)}$, for $j = 1,\ldots N$ (or alternatively $j = 1,\ldots N+1$). However, the main idea remains unchanged. These algorithms have been implemented in Matlab and are available from the authors upon request.

4.2 Performance of the procedure

In this section we present the results of an intensive simulation study to assess the performance of our detection proposal. In this study, we simulate the contaminated and no–contaminated multivariate series following the experimental conditions described in Section 3. We also consider D-BEKK, CCC and DCC models with Student-t distributed errors.[2]

We apply the detection method described in Section 4.1 where the assumed model is the true model used to generate the series. The results are shown in Table 3.[3] The measures used in the performance study are the percentage of times that the localization of the outliers is correctly detected and the percentage of false outliers. The threshold values used in the study are contained in Table 2.

The detection rate is larger for models with Gaussian errors. From Table 3 we can see that when the magnitude of the outlier is $\boldsymbol{\omega} = 10\boldsymbol{\sigma}_\mathbf{y}$, the procedure detects more than

2. Parameter values used for models with Student-t_7 distributed errors:
$\{\text{vec}(\mathbf{C}) = (0.106, 0.110, 0, 0.0371)', \text{diag}(\mathbf{A}) = (0.0571, 0.050)', \text{diag}(\mathbf{B}) = (0.983, 0.985)'\}$ for the D-BEKK model, $\{\boldsymbol{\alpha}_0 = (0.010, 0.013), \boldsymbol{\alpha}_1 = (0.049, 0.067), \boldsymbol{\beta}_1 = (0.740, 0.759), \rho_{12} = 0.506\}$ for the CCC model and $\{\boldsymbol{\alpha}_0 = (0.106, 0.110, 0.0371), \boldsymbol{\alpha}_1 = (0.0571, 0.050), \boldsymbol{\beta}_1 = (0.740, 0.759), \alpha = 0.015, \beta = 0.781\}$ for the DCC model.

3. The Matlab codes are available from the authors upon request.

96% of the isolated outliers, reaching the 100% in two cases. When the magnitude of the outlier is relatively small, $\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$, the detection rate goes from 36% to 43% for the D-BEKK model and from 68% and 77% for the CCC and DCC models. Regarding patches and volatility outliers, the detection rate also increases with the size of the outlier and it ranges from 24.1% (AVO and D-BEKK) to 99.8% (AVO and DCC). Finally, the percentage of false positives is at most 0.001% in 80% of the cases and under 0.007% in the rest (the only exception is the DCC model for $\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}, n = 1000$). Concerning models with Student-t distributed errors, we observe that, for example, when the magnitude of the outlier is $\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$, the procedure detects from 70.9% to 99.2% of isolated ALOs. As expected, the detection rate is low when $\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$, since it is difficult to distinguish small size outliers from the thick tail of Student-t distribution. The percentage of false positives is still very small, being at most 0.006% in more than 77% of the cases. These results lead us to conclude that the method is very reliable.

Table 3: *Percentage of correct detection of outliers and percentage of false outliers in 1000 replications of size n for multivariate GARCH models with either normal or Student-t_7 distributed errors.*

		Gaussian errors						Student-t_7 distributed errors					
		% of correct detections			% of false outliers			% of correct detections			% of false outliers		
	n	D-BEKK	CCC	DCC	D-BEKK	CCC	DCC	D-BEKK	CCC	DCC	D-BEKK	CCC	DCC
1 ALO	1000	43.8	77.1	77.2	0.004	0.005	0.005	12.5	7.7	7.8	0.0130	0.0082	0.0083
$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	3000	38.7	76.2	75.3	0.001	0.001	0.001	11.5	3.2	3.2	0.0082	0.0055	0.0056
	5000	36.1	69.8	70.8	0.001	0.001	0.001	11.2	2.1	2.1	0.0067	0.0050	0.0049
1 ALO	1000	99.1	100.0	100.0	0.004	0.004	0.048	92.2	98.8	99.0	0.0127	0.1080	0.0067
$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	3000	99.3	99.9	99.9	0.001	0.001	0.001	81.7	98.4	98.4	0.0081	0.0050	0.0050
	5000	99.3	99.9	99.8	0.001	0.001	0.001	73.3	99.2	99.1	0.0068	0.0047	0.0047
3 ALOs	1000	36.7	69.6	68.9	0.003	0.004	0.004	12.3	5.6	5.6	0.0113	0.0568	0.0566
$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	3000	36.5	71.2	71.1	0.001	0.001	0.001	10.3	2.5	2.5	0.0077	0.0050	0.0050
	5000	36.1	71.5	71.4	0.001	0.001	0.001	11.4	2.2	2.2	0.0066	0.0047	0.0048
3 ALOs	1000	96.5	97.8	97.8	0.002	0.005	0.005	88.2	93.0	93.3	0.0103	0.0062	0.0059
$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	3000	97.8	98.9	98.8	0.001	0.001	0.001	78.8	97.9	97.8	0.0075	0.0542	0.0042
	5000	97.8	99.1	98.9	0.001	0.001	0.001	70.9	98.1	98.2	0.0065	0.0043	0.0043
Patch of	1000	26.4	20.5	20.4	0.0001	0	0	24.0	1.7	1.9	0	0	0.0001
3 ALOs	3000	30.5	18.8	18.8	0	0	0	26.9	0.7	0.7	0.0001	0	0
$\boldsymbol{\omega} = 5\boldsymbol{\sigma}_{\mathbf{y}}$	5000	33.1	17.7	18.9	0.00002	0	0	26.6	0.1	0.1	0.0001	0.00004	0.00002
Patch of	1000	73.2	89.2	88.5	0	0	0	52.3	77.8	77.8	0	0	0
3 ALOs	3000	70.4	88.1	87.6	0	0	0	44.1	71.1	71.1	0.0001	0	0
$\boldsymbol{\omega} = 10\boldsymbol{\sigma}_{\mathbf{y}}$	5000	70.5	86.0	85.3	0.00002	0	0	41.1	63.2	63.3	0.0001	0.00002	0.00002
1 AVO	1000	24.2	66.4	95.6	0.001	0.0001	0	3.2	46.3	70.1	0.0001	0	0
$\boldsymbol{\omega} = 25\boldsymbol{\sigma}_{\mathbf{y}}$	3000	24.1	66.8	94.8	0.0003	0	0	3.3	47.1	66.6	0.0001	0.0001	0
	5000	24.4	66.5	95.6	0.0002	0	0	2.8	44.4	66.3	0.0002	0	0
1 AVO	1000	52.2	87.0	99.6	0.004	0	0	15.0	75.1	94.5	0.0002	0	0
$\boldsymbol{\omega} = 50\boldsymbol{\sigma}_{\mathbf{y}}$	3000	55.6	88.0	99.3	0.002	0	0	17.7	75.5	94.0	0.0002	0	0
	5000	53.9	88.8	99.8	0.001	0	0	17.3	75.0	94.2	0.0002	0	0
No	1000				0.004	0.006	0.005				0.0142	0.0092	0.0094
outliers	3000				0.001	0.001	0.001				0.0082	0.0057	0.0058
	5000				0.001	0.001	0.001				0.0068	0.0050	0.0050

In general, the percentage of correctly detected outliers is smaller for the D-BEKK model than for CCC and DCC and this is confirmed by the results presented in Section 3, where we observed that the effect of outliers in the estimation of the correlations was lower for D-BEKK model than for the CCC or DCC models.

These results show the reliability of the detection method for bivariate series. A natural question arises for higher-dimensional cases: is one random projection enough for detecting outliers or should the number of random projections be increased with N? As shown in what follows, our results suggest that there is no need to increase the number of random projections considered in the algorithm.

4.3 One or more random projections?

We focus now on analysing the performance of our procedure for $N = 10$ series and the D-BEKK model.[4] In particular, we are interested in studying whether increasing the number of random projections may increase the percentage of correctly detected outliers. In this case, threshold values are computed as suggested by Benjamini and Yekutieli (2001), instead of Bonferroni correction, which is too conservative. Results are contained in Table 4. As before, the measures used are the percentage of times that the localization of the outliers is correctly detected and the percentage of false outliers. The number of random projections is shown in the first column. We analyse here the case of 1 ALO of sizes $\omega = 5\sigma_{\mathbf{y}}, 10\sigma_{\mathbf{y}}$. For $\omega = 10\sigma_{\mathbf{y}}$ the proportion of correct detections stays constant when the number of random projections is increased, whereas for $\omega = 5\sigma_{\mathbf{y}}$ the increase is very low (0.2 percentage points). In contrast, the percentage of false outliers and the computational burden increase with a raise of the number of random projections, suggesting that it is not worth to use more than one random projection for large values of N; Similar conclusions where found by Cuesta-Albertos et al. (2006).

Table 4: *Percentage of correct detection of outliers and percentage of false outliers in 1000 replications of size $n = 1000$ for a D-BEKK model with Gaussian distributed errors. Series contaminated with one ALO of two different sizes.*

	1 ALO $\omega = 5\sigma_{\mathbf{y}}$		1 ALO $\omega = 10\sigma_{\mathbf{y}}$	
num. of random projections	% of correct detections	% of false outliers	% of correct detections	% of false outliers
1	21.4	0.0065	97.7	0.0241
2	21.4	0.0064	97.7	0.0178
5	21.4	0.0079	97.7	0.0212
10	21.6	0.0085	97.7	0.0280
20	21.6	0.0115	97.7	0.0377
50	21.6	0.0182	97.7	0.0672

4. Regarding the computational burden of this simulation study, we want to remark that estimating 1000 times the D-BEKK model for 10 series took approximately one week in an ordinary computer.

5 Empirical application

In this section, we analyse nine time series of returns to illustrate the performance of our method on real data. The data set is composed of the most important indices of the U.S. stock market, such as Nasdaq and S&P500 and company stocks such as Marathon Oil Corporation (MRO), International Business Machines Corporation (IBM), Coca-Cola Corporation (Cola), Colgate-Palmolive Corporation (Colgate), British Petroleum (BP), Microsoft Corporation (Micro) and American Express Company (AE). The data was collected from Yahoo Finance website (http://finance.yahoo.com) and spans the period from January 2, 1990 to January 30, 2015.

Figure 4 depicts the nine return series, $y_t = (\log p_t - \log p_{t-1}) \cdot 100$, where p_t is the value at time t of the corresponding index and Table 5 reports some summary statistics and the results of the Kiefer and Salmon test for normality in the context of conditional heteroscedastic series.

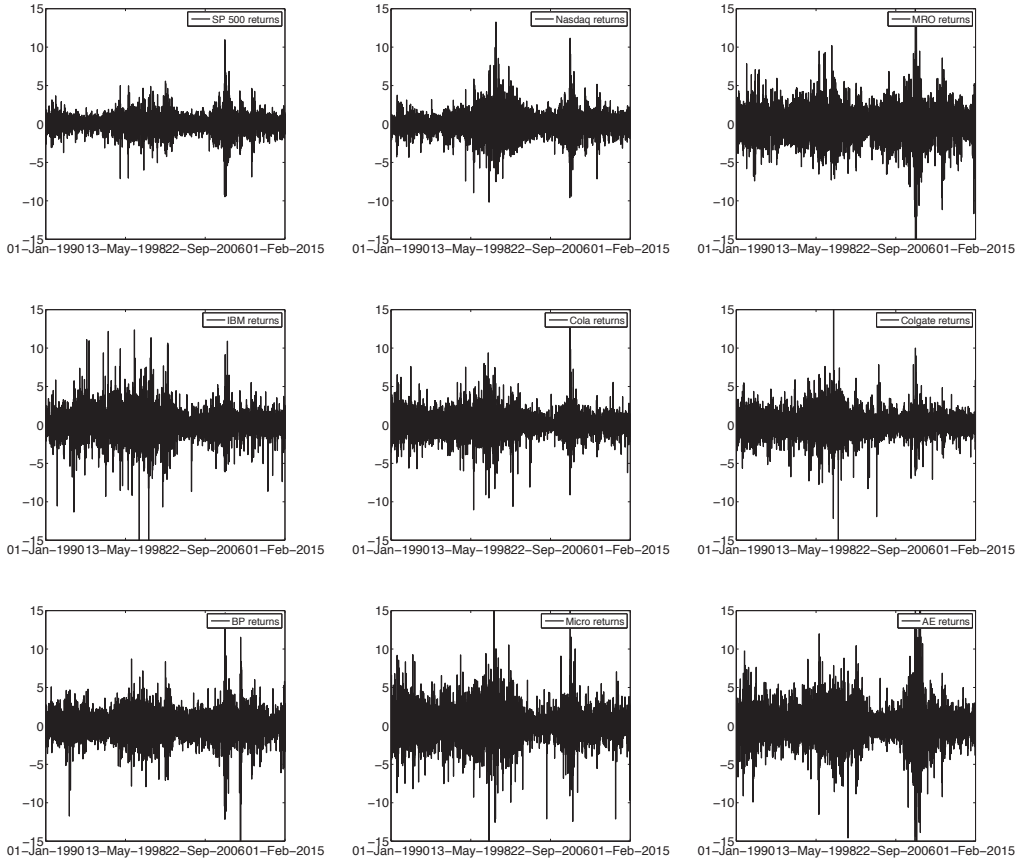


Figure 4: *Returns in percentage for the nine time series considered.*

Table 5: *Descriptive statistics for daily returns.*

Summary statistics of daily returns

Returns	Nasdaq	S&P 500	MRO	IBM	Cola	Colgate	BP	Micro	AE
Mean	0.0366	0.0271	0.0296	0.0360	0.0418	0.0531	0.0297	0.0715	0.0419
Variance	2.2333	1.3031	4.4819	3.1945	2.1428	2.2649	2.8337	4.3696	5.1985
Skewness	-0.0858	-0.2392	-0.2547	-0.0210	0.0353	-0.0230	-0.3780	-0.0101	0.0294
Kurtosis	9.1796	11.7389	10.9489	10.3444	8.6336	13.1178	12.3986	8.5061	10.5041

Results of the Kiefer and Salmon test

KS_S	-2.7843*	-7.7615*	-8.2652*	-0.6809	1.1449	-0.7462	-12.2656*	-0.3285	0.9542
KS_K	100.2478*	141.7666*	128.9508*	119.1434*	91.3907*	164.1346*	152.4680*	89.3229*	121.7340*

$*$ Means that we reject at 1%, 5% and 10% the nulls of skewness and kurtosis similar to those of a variable that follows a normal distribution.

From Table 5, we observe that the nine return series are in majority negatively skewed (except Cola and AE returns) and have significant kurtosis, ranging from 8.5061 for Micro to 13.1178 for Colgate, which suggests the existence of some outliers. It is known that this type of observations in time series leads to fat tail distributions, and some outlier detection methods, specially in the multivariate context, are based on this information (see for example Peña and Prieto, 2001, Galeano, Peña and Tsay, 2006). Table 5 also contains the results of the Kiefer and Salmon (1983) test (*KS*), which is a formal test of normality in the context of conditional heteroscedastic series.[5] The results of the test confirm the non Gaussianity of the nine return series.

Next, we estimate the three multivariate GARCH models considered in this work: the D-BEKK, the CCC and the DCC with Gaussian and Student-*t* distributed errors, and we proceed by applying our method to detect outliers. The degrees of freedom of the Student-*t* distributions are considered endogenous (they are included in the likelihood) and therefore estimated. Results are shown in Table 6.

Some dates are often detected as outliers or belong to a patch of outliers. Note that, the dates identified as outliers in the conditional correlation models are also identified as outliers in the D-BEKK model. Most of the outliers can be related to specific events. In particular, 19-Mar-91, 20/21-Oct-99, 21-Sep-04, 27/28-Apr-06, 19-Apr-13 and 18-Jul-13. On March 19, 1991 IBM announced that its returns were expected to decreased by half which led to an immediate plunge of its shares. On October 21, 1999 IBM stocks tumbled pulling the rest of the market with it. On April 28, 2006 Microsoft announces lower-than-expected earnings due to research expenses that would hurt future results. On September 21, 2004 CNN announced that the optimism about technologic stocks lifted the U.S. stock market at the open on this day. On April 19, 2013 it was announced that IBM shares posted their biggest one-day percentage drop in eight years. Finally, on

5. The Kiefer and Salmon (1983) test is given by $KS_N = (KS_S)^2 + (KS_K)^2$ (test of normality), where $KS_S = \sqrt{\frac{T}{6}}\left[\frac{1}{T}\sum_{t=1}^{T} y_t^{*3} - \frac{3}{T}\sum_{t=1}^{T} y_t^{*}\right]$ (test of skewness),$KS_K = \sqrt{\frac{T}{24}}\left[\frac{1}{T}\sum_{t=1}^{T} y_t^{*4} - \frac{6}{T}\sum_{t=1}^{T} y_t^{*2} + 3\right]$ (test of kurtosis) and y_t^* are the standardized returns. If the distribution of y_t^* is conditional $N(0,1)$, then KS_S and KS_K are asymptotically $N(0,1)$ and KS_N is asymptotically $\chi^2(2)$.

Table 6: *Observations identified as possible outliers in nine time series of returns.*

D-BEKK				CCC				DCC			
Gaussian		Student's *t*		Gaussian		Student's *t*		Gaussian		Student's *t*	
ALOs	Patches	ALOs	Patches	ALOs	Patches	ALOs	Patches	ALOs	Patches	ALOs	Patches
18-Mar-91	21-Oct-99	19-Mar-91	21-Sep-04	19-Mar-91	21-Sep-04	21-Sep-04	21-Sep-04	19-Mar-91	21-Sep-04	21-Sep-04	21-Sep-04
25-Jun-92	15-Mar-00	8-Apr-02	28-Apr-06	15-Nov-91	27-Apr-06	27-Apr-06		15-Nov-91	27-Apr-06	27-Apr-06	
15-Dec-92	21-Sep-04	23-Oct-03	19-Apr-13	25-Jun-92	18-Jul-13			25-Jun-92	18-Jul-13		
22-Jul-94	28-Apr-06	21-Sep-04	18-Jul-13	20-Oct-99				20-Oct-99			
27-Oct-97	19-Apr-13	15-Apr-05		15-Sep-00				15-Sep-00			
21-Oct-99	18-Jul-13	28-Apr-06		17-Sep-01				17-Sep-01			
15-Mar-00	21-Oct-14	29-Oct-07		22-Oct-03				22-Oct-03			
14-Sep-00		14-Oct-08		23-Jul-04				23-Jul-04			
19-Oct-00		29-Apr-10		21-Sep-04				21-Sep-04			
17-Sep-01		1-Jun-10		15-Apr-05				15-Apr-05			
8-Apr-02		9-Jun-10		27-Apr-06				27-Apr-06			
23-Oct-03		19-Apr-13		29-Jul-08				29-Jul-08			
23-Jul-04		18-Jul-13		8-Aug-11				8-Aug-11			
21-Sep-04		20-Oct-14		15-Apr-13				15-Apr-13			
18-Apr-05		21-Oct-14		19-Apr-13				19-Apr-13			
28-Apr-06				18-Jul-13				18-Jul-13			
29-Oct-07				17-Oct-14				17-Oct-14			
29-Apr-08											
14-Oct-08											
29-Apr-10											
1-Jun-10											
9-Jun-10											
29-Jul-10											
8-Aug-11											
15-Apr-13											
19-Apr-13											
18-Jul-13											
4-Sep-14											
20-Oct-14											
21-Oct-14											
30-Jan-15											

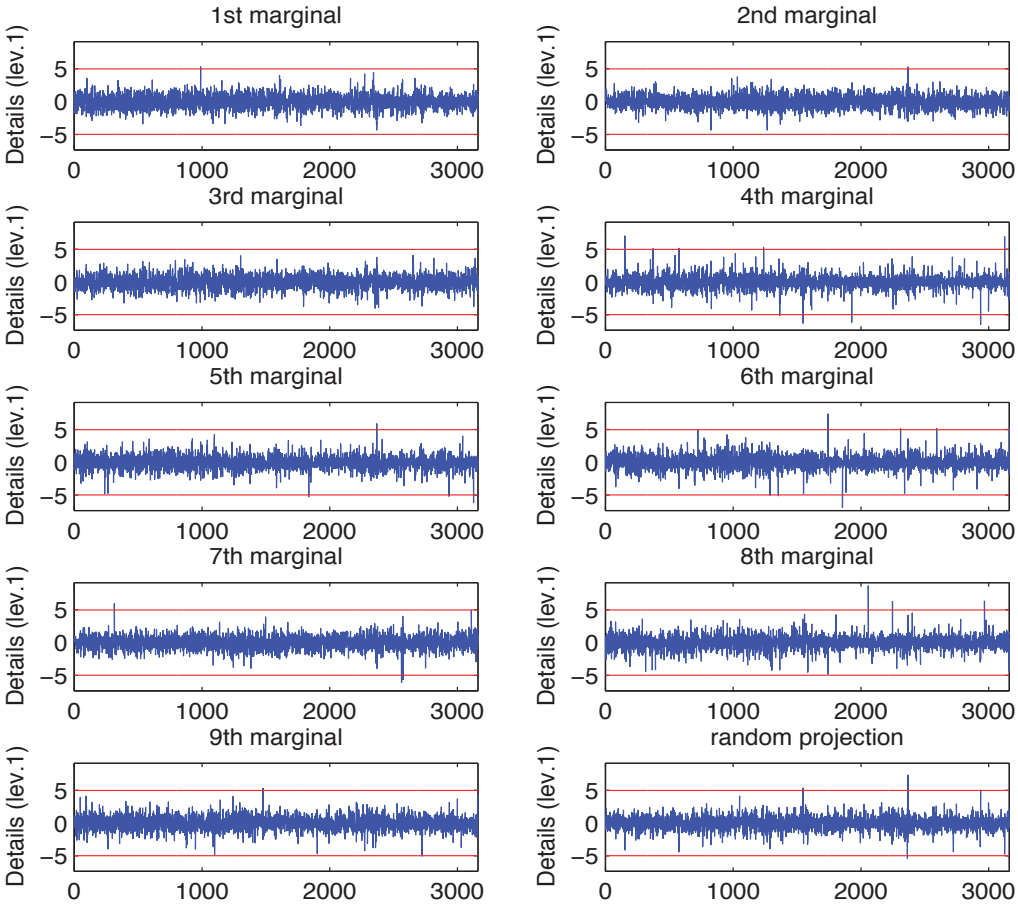


Figure 5: *Graphical output of the wavelet-based procedure for the returns of nine time series estimated with a D-BEKK model with Gaussian errors.*

July 18, 2013 news about the European Union plans to limit fees for using credit and debit cards pushed down the American Express company's shares. Furthermore, other dates like 15-Nov-91, 15-Dec-92, 22-Jul-94 and 29-Jul-08 can be related to oil shocks; In 1990 Irak invades Kuwait and Kuwait cut crude exports until 1994; Oil prices dropped from historic highs of $144.29 in July 2008, to $33.87 five months later.

Figure 5 shows a graphical output of the Matlab program, which corresponds to the analysis of the multivariate residuals obtained after fitting a D-BEKK model with Gaussian errors to nine series of returns.

All these observations correspond to important financial crashes or oil shocks that our procedure detected successfully.

6 Conclusion

The main contribution of this paper is the proposal of a general detection algorithm based on wavelets that can be applied to a large class of multivariate volatility models. The effectiveness of our method is evaluated both with simulated and real data. The simulations report evidence that our proposal is both effective and reliable since it detects very few false outlier.

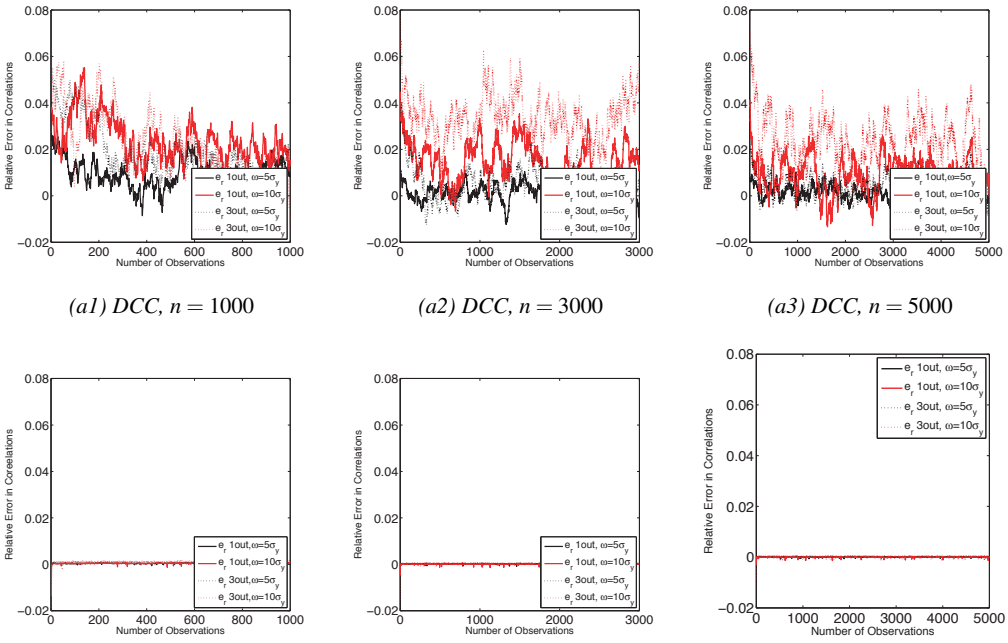
We also study the impact of outliers (isolated level outliers, patches of level outliers and volatility outliers) on the estimation of correlations when fitting well known multivariate GARCH models via several simulation studies. The results of the Monte Carlo experiments show that correlations are considerably affected by the presence of outliers. The impact on the correlations is stronger the higher is the magnitude of the outlier, the larger the number of outliers included in the simulated series and the smaller the sample sizes of the simulated time series. In the simulation, we consider scenarios that try to mimic portfolios that include asset returns and commodity returns such as oil, where the correlation is negative and quite negative when turmoils in the oil market propagate to stock markets. Therefore, the implications of these results are important for investments in oil commodities, as we identify several sources of impacts that are useful for controlling international risks of investments.

Appendix

A.1 Student-t distributed errors

The simulations are conducted following the experimental conditions explained in the beginning of section 3. In this case, the parameter values used are: for the D-BEKK model, $\{\text{vec}(\mathbf{C}) = (0.053, 0.042, 0, 0.020)^{\intercal}, \text{diag}(\mathbf{A}) = (0.161, 0.164)^{\intercal}, \text{diag}(\mathbf{B}) = (0.983, 0.981)^{\intercal}\}$; $\{\boldsymbol{\alpha}_0 = (0.010, 0.013), \boldsymbol{\alpha}_1 = (0.019, 0.027), \boldsymbol{\beta}_1 = (0.940, 0.826), \rho_{12} = -0.306)\}$ for the CCC model and $\{\boldsymbol{\alpha}_0 = (0.010, 0.013), \boldsymbol{\alpha}_1 = (0.019, 0.027), \boldsymbol{\beta}_1 = (0.940, 0.826), \alpha = 0.015, \beta = 0.981\}$ for the DCC model.

The idea is similar to the experiments showed in Section 3. We simulate with Student-t_7 errors and estimate the CCC, DCC and D-BEKK considering the degree of freedom of the Student-t distribution endogenous. Table A and Figures A–C contain the results of this second simulation study.



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

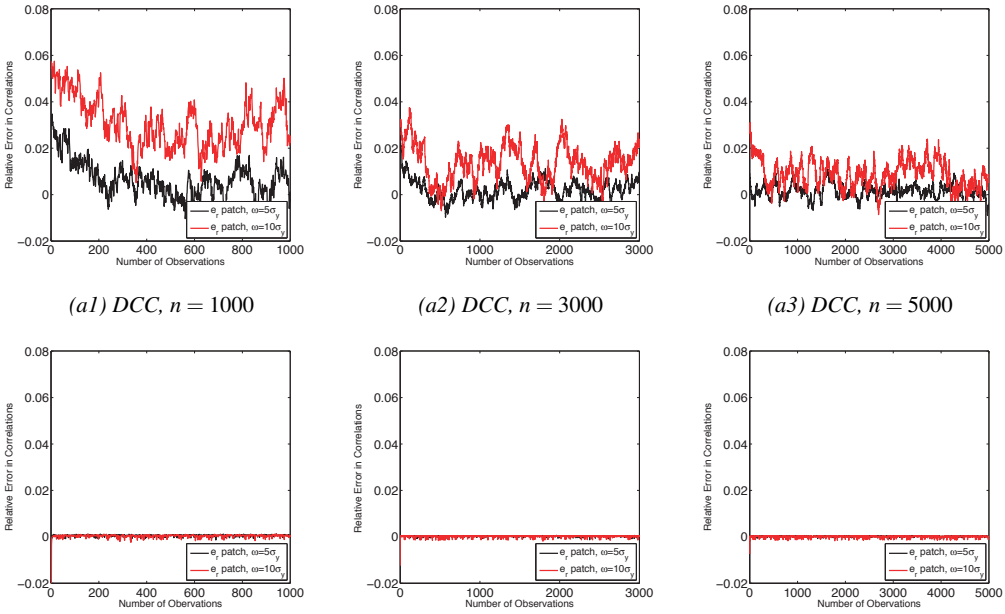
(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure A: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with* Student-t_7 *distributed errors from 1000 simulated series of size n that include ALOs of different magnitudes.*

Table A: *Relative bias in the estimated correlations obtained from a CCC model from 1000 simulated series of size n that include outliers of different magnitudes.*

CCC Model with Student-t_7 distributed errors

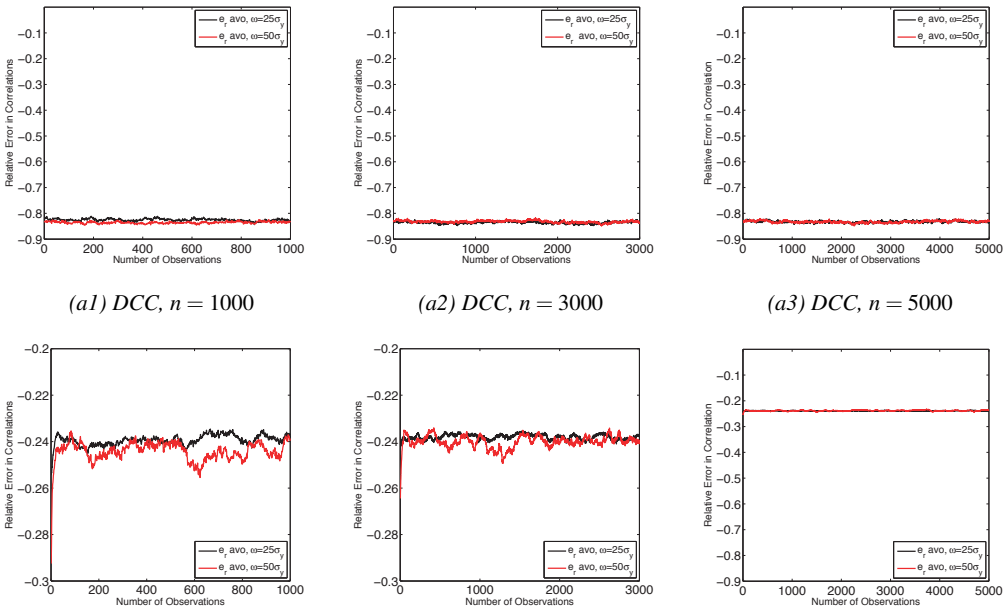
	n	Estimated Correlation	Relative Bias		n	Estimated Correlation	Relative Bias
1 ALO	1000	-0.3073	0.0062	3 ALOSs	1000	-0.3106	0.0170
$\omega = 5\sigma_y$	3000	-0.3071	0.0029	$\omega = 5\sigma_y$	3000	-0.3086	0.0078
	5000	-0.3070	0.0020		5000	-0.3079	0.0049
1 ALO	1000	-0.3093	0.0128	3 ALOs	1000	-0.3141	0.0285
$\omega = 10\sigma_y$	3000	-0.3100	0.0124	$\omega = 10\sigma_y$	3000	-0.3145	0.0271
	5000	-0.3089	0.0082		5000	-0.3128	0.0209
Patch of	1000	-0.3075	0.0069	1 AVO	1000	-0.2918	-0.0445
3 ALOs	3000	-0.3072	0.0033	$\omega = 25\sigma_y$	3000	-0.3006	-0.0183
$\omega = 5\sigma_y$	5000	-0.3072	0.0026		5000	-0.3025	-0.0127
Patch of	1000	-0.3137	0.0272	1 AVO	1000	-0.2745	-0.1012
3 ALOs	3000	-0.3099	0.0121	$\omega = 50\sigma_y$	3000	-0.2906	-0.0509
$\omega = 10\sigma_y$	5000	-0.3089	0.0082		5000	-0.2955	-0.0356
No outliers	1000	-0.3054					
	3000	-0.3062					
	5000	-0.3064					



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure B: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with* Student-t_7 *distributed errors from 1000 simulated series of size n that include patches of different magnitudes.*



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure C: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with* Student-t_7 *distributed errors from 1000 simulated series of size n that include 1 AVO of different magnitudes.*

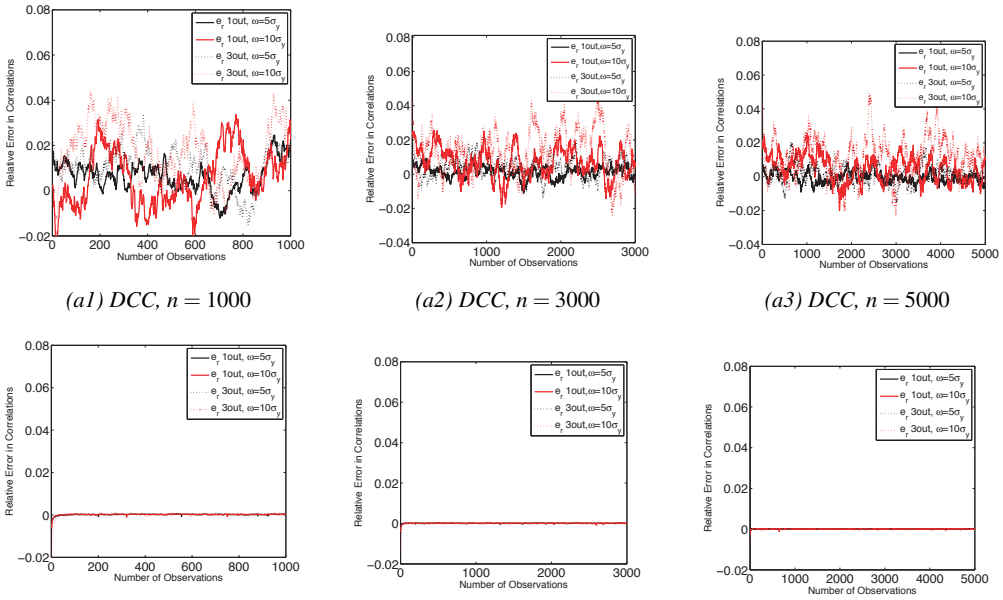
A.2 $N = 4$ and Student-t_7 distributed errors

The simulations are conducted following the experimental conditions explained in section 3, with $N = 4$. For better comparison with results of Appendix A.1, we choose the same parameter values for the first two series.

Once again, the outliers are placed randomly across the simulated series, but in the same position for each pair of series. We consider that the outlier affects each pair of series at the same instant. Each scenario involves 1000 replications and series are simulated from CCC, DCC and D-BEKK(1,1,1) models with Student-t_7 errors. The number of replications is selected to provide robust results. Given a model, we analysed 24 scenarios, that are defined from the type and number of outliers (one isolated ALO, multiple ALOs, patches of three ALOs, one isolated AVO), the size of the outlier ($\omega = 5\sigma_y, 10\sigma_y$ for ALOs and $\omega = 25\sigma_y, 50\sigma_y$ for AVOs) and the sample size of the simulated series ($n = 1000, 3000, 5000$). Table B and Figures D–F contain the results of this simulation study. In order to simplify the presentation, we only report the results for the correlation between the first two simulated series.

Table B: *Relative bias in the estimated correlations (first and second series) obtained from a CCC model with errors following* Student-t_7 *distributions from 1000 simulated series of size n that include outliers of different magnitudes.*

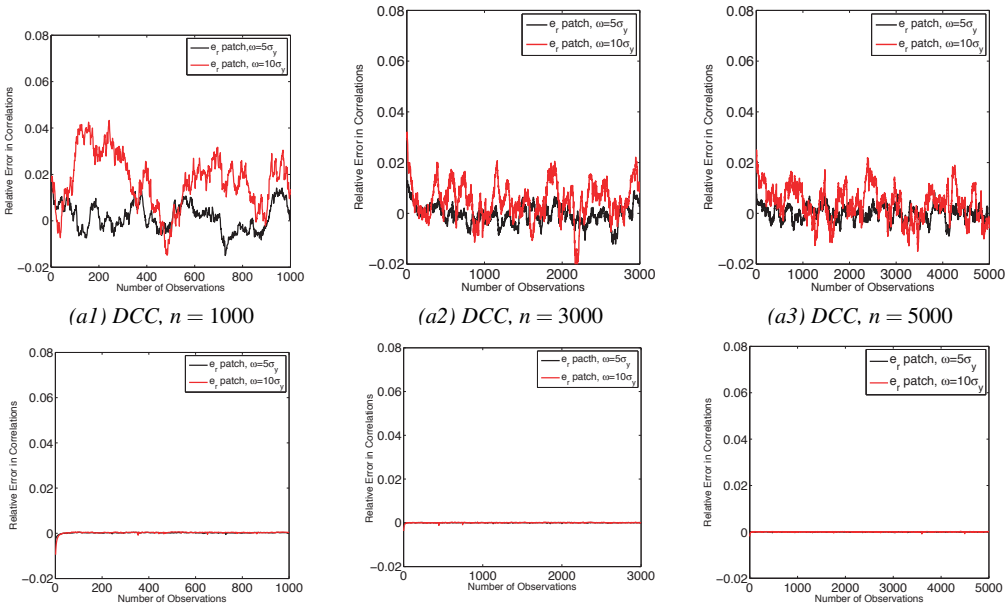
	n	Estimated Correlation	Relative Bias		n	Estimated Correlation	Relative Bias
1 ALO	1000	-0.3076	0.0052	3 ALOSs	1000	-0.3105	0.0147
$\omega = 5\sigma_y$	3000	-0.3069	0.0029	$\omega = 5\sigma_y$	3000	-0.3082	0.0072
	5000	-0.3061	0.0013		5000	-0.3069	0.0039
1 ALO	1000	-0.3093	0.0108	3 ALOs	1000	-0.3123	0.0206
$\omega = 10\sigma_y$	3000	-0.3066	0.0020	$\omega = 10\sigma_y$	3000	-0.3153	0.0304
	5000	-0.3078	0.0069		5000	-0.3121	0.0210
Patch of	1000	-0.3078	0.0058	1 AVO	1000	-0.2934	-0.0411
3 ALOs	3000	-0.3065	0.0018	$\omega = 25\sigma_y$	3000	-0.3003	-0.0186
$\omega = 5\sigma_y$	5000	-0.3061	0.0012		5000	-0.3013	-0.0144
Patch of	1000	-0.3156	0.0314	1 AVO	1000	-0.2713	-0.1134
3 ALOs	3000	-0.3096	0.0117	$\omega = 50\sigma_y$	3000	-0.2922	-0.0451
$\omega = 10\sigma_y$	5000	-0.3079	0.0071		5000	-0.2962	-0.0311
No outliers	1000	-0.3060					
	3000	-0.3060					
	5000	-0.3057					



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

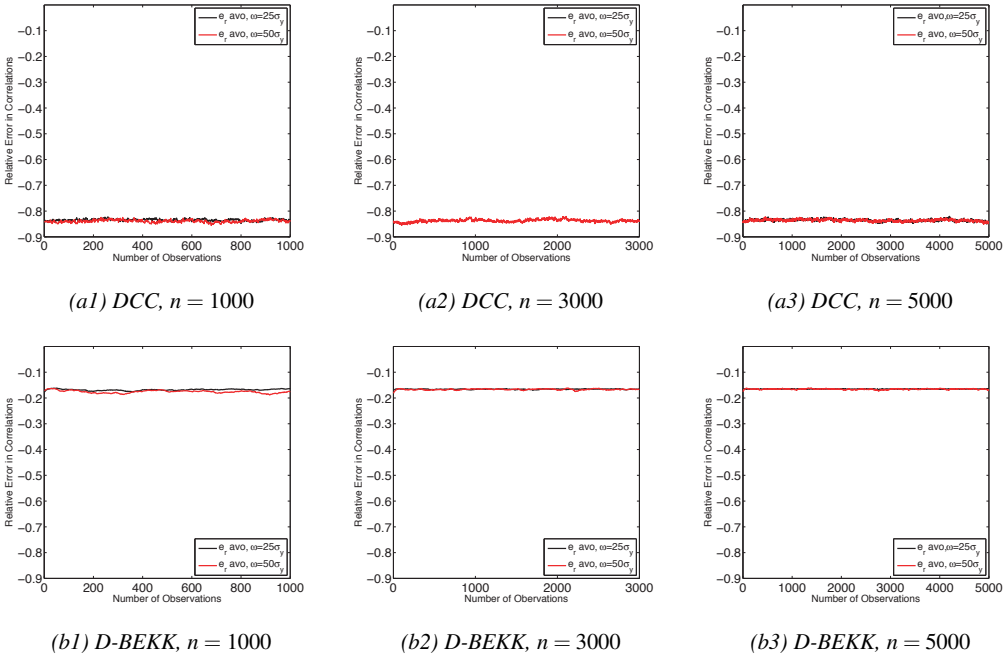
Figure D: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following* Student-t_7 *distributions from 1000 simulated series of size n that include ALOs of different magnitudes.*



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure E: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following* Student-t_7 *distributions from 1000 simulated series of size n that include patches of different magnitudes.*



(a1) DCC, $n = 1000$ *(a2) DCC, $n = 3000$* *(a3) DCC, $n = 5000$*

(b1) D-BEKK, $n = 1000$ *(b2) D-BEKK, $n = 3000$* *(b3) D-BEKK, $n = 5000$*

Figure F: *Relative bias in the estimated correlations obtained from a (a) DCC model and a (b) D-BEKK model with errors following* Student-t_7 *distributions from 1000 simulated series of size n that include 1 AVO of different magnitudes.*

References

Baillie, R. and Bollerslev, T. (1989). The message in daily exchange rates: A conditional variance tale. *Journal of Business and Economic Statistics*, 7, 297–309.

Bali, R. and Guirguis, H. (2007). Extreme observations and non-normality in ARCH and GARCH. *International Review of Economics and Finance*, 16, 332–346.

Bauwens, L., Laurent, S. and Rombout, J. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21, 79–109.

Behmiri, N. and Manera, M. (2015). The role of outliers and oil price shocks on volatility of metal prices. *Resources Policy*, 46, 139–150.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165–1188.

Bingham, E. and Mannila, H. (2001). *Proceedings of the seventh ACM SIGKDD international conference on knowledge and data mining*, Random projection in dimensionality reduction: Applications to image and text data. 245–250. ACM New York.

Bollerslev, T. (1990). Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics*, 42, 498–505.

Boudt, K. and Croux, C. (2010). Robust m-estimation of multivariate GARCH models. *Computational Statistics and Data Analysis*, 54, 2459–2469.

Boudt, K., Daníelsson, J. and Laurent, S. (2013). Robust forecasting of dynamic conditional correlation GARCH models. *International Journal of Forecasting*, 29, 244–257.

Carnero, M., Peña, D. and Ruiz, E. (2007). Effects of outliers on the identification and estimation of GARCH models. *Journal of Time Series Analysis*, 28, 471–497.

Carnero, M., Peña, D. and Ruiz, E. (2012). Estimating GARCH volatility in the presence of outliers. *Economic Letters*, 114, 86–90.

Charles, A. and Darné, O. (2014). Volatility persistence in crude oil markets. *Energy Policy*, 65, 729–742.

Chen, C. and Liu, L. (1993). Joint estimation of model parameters and outlier effects. *Journal of American Statistical Association*, 88, 284–297.

Cuesta-Albertos, J., del Barrio, E., Fraiman, R. and Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Computational Statistics and Data Analysis*, 51, 4814–4831.

Cuesta-Albertos, J., Fraiman, R. and Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, 37, 1–25.

Duan, J.-C., Gauthier, G., Simonato, J.-G. and Sasseville, C. (2006). Approximating the GJR-GARCH and EGARCH option pricing models analytically. *Journal of Computational Finance*, 9, 41–69.

Engle, R. (2002). Dynamic conditional correlation? A simple class of multivariate GARCH models. *Journal of Business and Economic Statistics*, 20, 339–350.

Engle, R. and Kroner, K. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11, 122–150.

Fox, A. (1972). Outliers in time series. *Journal of Royal Statistical Society B*, 34, 350–363.

Franses, P. and Ghijsels, H. (1999). Additive outliers, GARCH and forecasting volatility. *International Journal of Forecasting*, 15, 1–9.

Galeano, P. and Peña, D. (2013). *Robustness and complex data structures*, Finding outliers in linear and nonlinear time series, pp. 243–262. Springer-Verlag.

Galeano, P., Peña, D. and Tsay, R. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, 101, 654–669.

Grané, A. and Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Computational Statistics and Data Analysis*, 54, 2580–2593.

Grané, A. and Veiga, H. (2014). Outliers, GARCH-type models and risk measures: A comparison of several approaches. *Journal of Empirical Finance*, 26, 26–40.

Grané, A., Veiga, H. and Martín-Barragán, B. (2014). Additive Level Outliers in Multivariate GARCH Models. In V. Melas, S. Mignani, P. Monari, and L. Salmaso (Eds.), *Topics in Statistical Simulation*, Volume 114 of *Springer Proceedings in Mathematics & Statistics*, pp. 247–255. Springer.

Grossi, L. and Laurini, F. (2009). A robust forward weighted lagrange multiplier test for conditional heteroscedasticity. *Computational Statistics and Data Analysis*, 53, 2251–2263.

Hotta, L. and Tsay, R. (2012). Outliers in GARCH processes. In W. Bell, S. Hollan, and T. McElroy (Eds.), *Economic Time Series: Modeling and Seasonality*, pp. 337–358. CRC Press, Boca Raton, FL.

Hotta, L. K. and Trucíos, C. (2018). *Advances in Mathematics and Applications.*, Inference in (M)GARCH Models in the Presence of Additive Outliers: Specification, Estimation, and Prediction, pp. 179–202. Springer.

Kamranfar, H., Chinipardaz, R. and Mansouri, B. (2017). Detecting outliers in garch(p,q) models. *Communications in Statistics - Simulation and Computation 46*(10), 7844–7854.

Kiefer, N. and Salmon, M. (1983). Testing normality in econometric models. *Economics Letters*, 11, 123–127.

Ledolter, J. (1989). The effect of additive outliers on the forecasts from ARIMA models. *International Journal of Forecasting*, 5, 231–240.

Muler, N. and Yohai, V. J. (2008). Robust estimates for GARCH models. *Journal of Statistical Planning and Inference*, 138, 2918 – 2940.

Peña, D. and Prieto, F. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43, 286–310.

Ramos, S., Martín-Barragán, B. and Veiga, H. (2015). Correlations between oil and stock markets: A wavelet-based approach. *Economic Modelling*, 50, 212–227.

Roy, S. (1953). On a Heuristic Method of Test Construction and its use in Multivariate Analysis. *The Annals of Mathematical Statistics*, 24, 220–238.

Silvennoinen, A. and Teräsvirta, T. (2009). *Handbook of Financial Time Series*, Multivariate GARCH models, pp. 201–226. Springer.

Teräsvirta, T. (1996). Two stylized facts and the GARCH(1,1) model. *Working Paper 96, Stockholm School of Economics*.

Van Dijk, D., Franses, P. and Lucas, A. (1999). Testing for ARCH in the presence of additive outliers. *Journal of Applied Econometrics*, 14, 539–562.

Vempala, S. (2004). *The Random Projection Method*. Providence, RI: American Mathematical Society.

Verhoeven, P. and McAleer, M. (2000). Modelling outliers and extreme observations for ARMA-GARCH processes. *Working Paper, University of Western Australia*.

SORT 43 (2) July-December 2019, 317-336 DOI: 10.2436/20.8080.02.90

A class of goodness-of-fit tests for circular distributions based on trigonometric moments

S. Rao Jammalamadaka[1], M. Dolores Jiménez-Gamero[2] and
Simos G. Meintanis[3,4]

Abstract

We propose a class of goodness–of–fit test procedures for arbitrary parametric families of circular distributions with unknown parameters. The tests make use of the specific form of the characteristic function of the family being tested, and are shown to be consistent. We derive the asymptotic null distribution and suggest that the new method be implemented using a bootstrap resampling technique that approximates this distribution consistently. As an illustration, we then specialize this method to testing whether a given data set is from the von Mises distribution, a model that is commonly used and for which considerable theory has been developed. An extensive Monte Carlo study is carried out to compare the new tests with other existing omnibus tests for this model. An application involving five real data sets is provided in order to illustrate the new procedure.

MSC: 62H15, 62G20.

Keywords: Goodness-of-fit, Circular data, Empirical characteristic function, Maximum likelihood estimation, von Mises distribution.

1 Introduction

Let Θ be an arbitrary circular random variable with cumulative distribution function (CDF) F. Then on the basis of independent and identically distributed (i.i.d.) copies $\vartheta_1,\ldots,\vartheta_n$ of Θ we are interested in testing goodness–of–fit (GOF) of the composite null hypothesis,

$$\mathscr{H}_0 : F \in \mathscr{F}_{\boldsymbol{\beta}} \tag{1}$$

against general alternatives, where $\mathscr{F}_{\boldsymbol{\beta}} = \{F(\cdot;\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathscr{B}\}$ denotes a parametric family of CDFs indexed by the parameter $\boldsymbol{\beta} \in \mathscr{B} \subset \mathbb{R}^p$.

[1] Department of Statistics and Applied Probability, University of California Santa Barbara, USA, sreenivas@ucsb.edu

[2] Department of Statistics and Operations Research, University of Sevilla, Spain, dolores@us.es

[3] Department of Economics, National and Kapodistrian University of Athens, Greece, simosmei@econ.uoa.gr

[4] Unit for Business Mathematics and Informatics, North–West University, South Africa.

Received: April 2019
Accepted: October 2019

A well-known class of GOF tests that have been discussed in the literature, is obtained by comparing a nonparametric estimator of the CDF of Θ with the corresponding parametric estimator of the same quantity reflecting the null hypothesis. To this end, denote by $\widehat{\boldsymbol{\beta}}$ a consistent estimator of the parameter $\boldsymbol{\beta}$, and write $F(\cdot;\widehat{\boldsymbol{\beta}})$ for the CDF corresponding to (1) with estimated parameter. Also let

$$F_n(x) = \frac{\#\{j : \vartheta_j' s \leq x\}}{n},$$

be the empirical CDF. Then, based on a distance function Δ, the CDF–based test statistics may be formulated as

$$\Delta_n := \Delta(F_n(\cdot), F(\cdot;\widehat{\boldsymbol{\beta}})), \tag{2}$$

and rejects the null hypothesis $\mathscr{H}_0$ stated in (1) for large values of Δ_n. The specific type of distance Δ_n adopted in (2) leads to different GOF methods, chief among these are the Kuiper (1960) and the Watson (1961) tests, which are a variation of the Kolmogorov–Smirnov and the Cramér–von Mises tests, respectively. Note that both tests are appropriately adapted from the case of testing a distribution on the real line to the case of testing for circular distributions; see e.g. Jammalamadaka and SenGupta (2001) §7.2.1.

In this paper we suggest a new class of GOF tests which is based on the characteristic function (CF) of circular distributions. Such CF-based GOF tests for distributions on the real-line have proved to be more convenient, and compete well with corresponding methods based on the CDF; see for instance the normality test proposed by Epps and Pulley (1983), the test for the Cauchy distribution of Gürtler and Henze (2000), and the tests for the stable distribution suggested by Matsui and Takemura (2008), and Meintanis (2005).

The remainder of the paper is organized as follows. In Section 2 we introduce the new GOF procedure for circular distributions and prove consistency of the corresponding test criteria. In Section 3 we derive the limit distribution of the test statistic under the null hypothesis. Given the highly non–trivial structure of this distribution, we investigate in Section 4 the consistency of an appropriate resampling version of our method. In Section 5 the particular case of testing for the von Mises distribution is studied in detail. The finite–sample properties of the test are illustrated by means of a Monte Carlo study in Section 6, while Section 7 provides an application. Section 8 includes a brief summary and discussion. The paper contains a Supplement that includes the necessary R scripts for the benefit of potential users. Technical assumptions and proofs are deferred to the Appendix.

2 Tests based on the characteristic function

In a somewhat similar spirit with the Kuiper and Watson tests that use a distance between CDFs, we propose to use a distance between CFs instead of the CDFs. To this end, write

$\varphi(r) = \mathbb{E}(e^{\mathrm{i}r\Theta}),\ r \in \mathbb{R}$, for the CF of Θ and define the empirical CF corresponding to $\vartheta_1, \ldots, \vartheta_n$, as

$$\varphi_n(r) = \frac{1}{n}\sum_{j=1}^{n} e^{\mathrm{i}r\vartheta_j},\ \ (\mathrm{i} = \sqrt{-1}). \tag{3}$$

Also write $\varphi(\cdot;\boldsymbol{\beta}) := \Re\varphi(r;\boldsymbol{\beta}) + \mathrm{i}\Im\varphi(r;\boldsymbol{\beta})$ for the CF under the null hypothesis, where $\Re(z)$ (resp. $\Im(z)$) denotes the real (resp. imaginary) part of a complex number. In this paper we consider CF–based test statistics in the form $\Delta\ (\varphi_n(\cdot),\ \varphi(\cdot;\widehat{\boldsymbol{\beta}}))$. As before, rejection is for large values of the test statistic.

Specifically we consider a Cramér–von Mises type distance. However, since for circular distributions the CF needs to be evaluated only at integer values (Jammalamadaka and SenGupta, 2001, §2.2), and taking into account further the symmetry property of the CF and the empirical CF, our test statistic can be formulated as

$$C_{n,p} = n\sum_{r=0}^{\infty}\left|\varphi_n(r) - \varphi(r;\widehat{\boldsymbol{\beta}})\right|^2 p(r), \tag{4}$$

where $p(\cdot)$ denotes a probability function over the non–negative integers.

By straightforward algebra we have from (4)

$$C_{n,p} = n\sum_{r=0}^{\infty}\left\{R_n(r;\widehat{\boldsymbol{\beta}}) + I_n(r;\widehat{\boldsymbol{\beta}})\right\} p(r),$$

with

$$R_n(r;\widehat{\boldsymbol{\beta}}) = \left\{\frac{1}{n}\sum_{j=1}^{n}\cos(r\vartheta_j) - \Re\varphi(r;\widehat{\boldsymbol{\beta}})\right\}^2$$

and

$$I_n(r;\widehat{\boldsymbol{\beta}}) = \left\{\frac{1}{n}\sum_{j=1}^{n}\sin(r\vartheta_j) - \Im\varphi(r;\widehat{\boldsymbol{\beta}})\right\}^2.$$

Because of the one–to–one correspondence between CFs and CDFs, it readily follows that the test based on $C_{n,p}$ is consistent against any fixed alternative to $\mathscr{H}_0$ provided that

$$p(r) > 0, \quad \forall \quad r \geq 0. \tag{5}$$

To see this, assume that the estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ has a strong probability limit, say $\boldsymbol{\beta}^0$, even under alternatives, and that $\varphi(r;\boldsymbol{\beta})$ is continuous as a function of $\boldsymbol{\beta}$. Then since $\left|\varphi_n(r) - \varphi(r;\widehat{\boldsymbol{\beta}})\right|^2 \leq 4$, we have from (4),

$$\frac{C_{n,p}}{n} \longrightarrow \sum_{r=0}^{\infty} \left|\varphi(r) - \varphi(r;\boldsymbol{\beta}^0)\right|^2 p(r) \quad a.s. \quad \text{as } n \to \infty, \tag{6}$$

due to the strong consistency of the empirical CF (see Csörgő, 1981 and Marcus, 1981), and by invoking Lebesgue's dominated convergence theorem. In view of the uniqueness of the CF, the right–hand side of (6) is positive, unless $F(\cdot) = F(\cdot;\boldsymbol{\beta}^0)$, which shows the strong consistency of the test that rejects the null hypothesis $\mathscr{H}_0$ for large values of $C_{n,p}$ since, from Theorem 1 in next Section, $C_{n,p}$ is bounded in probability.

In the next section we investigate the large–sample behavior of $C_{n,p}$ under the null hypothesis. From now on, it will be assumed that (5) holds.

3 The limit null distribution of the CF test statistic

Let ℓ_p^2 denote the (separable) Hilbert space of all infinite sequences $z = (z_0, z_1, \ldots)$ of complex numbers such that $\sum_{r\geq 0} |z_r|^2 p(r) < \infty$, with the inner product defined as

$$\langle z, w\rangle_{\ell_p^2} = \sum_{r\geq 0} z_r \bar{w}_r p(r),$$

for $z = (z_0, z_1, \ldots)$, $w = (w_0, w_1, \ldots) \in \ell_p^2$, where for any complex number $x = a + \mathrm{i}b$, $\bar{x} = a - \mathrm{i}b$ stands for its complex conjugate. Let also $\|\cdot\|_{\ell_p^2}$ denote the norm in this space. With this notation our test statistic may be written as,

$$C_{n,p} = \|Z_n\|^2_{\ell_p^2}, \tag{7}$$

where $Z_n(r) = \sqrt{n}\{\varphi_n(r) - \varphi(r;\widehat{\boldsymbol{\beta}})\}$.

Also let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and write

$$\nabla \Re\varphi(r;\boldsymbol{\beta}) = \left(\frac{\partial}{\partial\beta_1}\Re\varphi(r;\boldsymbol{\beta}), \ldots, \frac{\partial}{\partial\beta_p}\Re\varphi(r;\boldsymbol{\beta})\right)^\top,$$

$$\nabla \Im\varphi(r;\boldsymbol{\beta}) = \left(\frac{\partial}{\partial\beta_1}\Im\varphi(r;\boldsymbol{\beta}), \ldots, \frac{\partial}{\partial\beta_p}\Im\varphi(r;\boldsymbol{\beta})\right)^\top.$$

Next theorem shows convergence in distribution of $Z_n(\cdot)$ under Assumptions A, B and C stated in the Appendix.

Theorem 1 *Assume that $\vartheta_1, \ldots, \vartheta_n$, are i.i.d. copies of Θ and that Assumptions A, B and C are fulfilled. Then, under the null hypothesis $\mathscr{H}_0$, there is a centred Gaussian random element $Z(\cdot)$ of ℓ_p^2 having covariance kernel*

$$K(r,s) = \mathbb{E}\{\Upsilon(r,\Theta;\boldsymbol{\beta})\overline{\Upsilon}(s,\Theta;\boldsymbol{\beta})\},$$

such that

$$Z_n \xrightarrow{\mathscr{L}} Z, \quad as \;\; n \to \infty,$$

where

$$\Upsilon(r,\Theta;\boldsymbol{\beta}) = \cos(r\Theta) - \Re\varphi(r;\boldsymbol{\beta}) - \nabla\Re\varphi(r;\boldsymbol{\beta})^\top L(\Theta;\boldsymbol{\beta})$$
$$+ \mathrm{i}\left\{\sin(r\Theta) - \Im\varphi(r;\boldsymbol{\beta}) - \nabla\Im\varphi(r;\boldsymbol{\beta})^\top L(\Theta;\boldsymbol{\beta})\right\},$$

with $L(\Theta;\boldsymbol{\beta})$ defined in Assumption A.

In view of (7), the asymptotic null distribution of $C_{n,p}$ stated in next corollary is an immediate consequence of Theorem 1 and the Continuous Mapping Theorem.

Corollary 1 *Suppose that assumptions in Theorem 1 hold, then*

$$C_{n,p} \xrightarrow{\mathscr{L}} \|Z\|^2_{\ell^2_p},$$

where $Z(\cdot)$ is the Gaussian random element appearing in Theorem 1.

Remark 1 *The distribution of $\|Z\|^2_{\ell^2_p}$ is the same as that of $\sum_{j=1}^{\infty}\lambda_j N_j^2$, where $\lambda_1,\lambda_2,\ldots$ are the positive eigenvalues of the integral operator $f \mapsto Af$ on ℓ^2_p associated with the kernel $K(\cdot,\cdot)$ given in Theorem 1, i.e., $(Af)(r) = \sum_{s\geq 0} K(r,s)f(s)p(s)$, and $N_1,N_2,\ldots$ are i.i.d. standard normal random variables. In general, the calculation of those eigenvalues is a very difficult task.*

Remark 2 *Assumptions A, B and C in Theorem 1 are quite standard in the context of GOF testing. Specifically Assumption A refers to an asymptotic (Bahadur) representation of a given estimator of the parameter $\boldsymbol{\beta}$ and is satisfied by common estimators such as maximum likelihood and moment estimators. Assumptions B and C imply smoothness of the CF as a function of $\boldsymbol{\beta}$.*

Since our assumptions are relatively weak, our CF approach is quite general and may be applied for testing GOF for a wide spectrum of circular distributions. In Section 5 we will specialize to a CF–based GOF test for the von–Mises distribution, which is as popular for circular data as the Gaussian distribution is for linear data.

4 The parametric bootstrap

As pointed out in Remark 1, the asymptotic null distribution of the test statistic $C_{n,p}$ is complicated and depends on several unknown quantities in a highly complicated manner. There exists no feasible approximation of the distribution in Theorem 1 which will allow us to actually carry out the test. We study here a resampling method labelled

"parametric bootstrap", which is a computer–assisted automatic procedure for performing this task. The parametric bootstrap estimates the null distribution of the test statistic $C_{n,p}$ by means of its conditional distribution, given the data, when the data come from $F(\cdot;\widehat{\boldsymbol{\beta}})$. Although the exact bootstrap estimator is still difficult to derive, it can be approximated as outlined below within the (fairly general) setting considered in Section 3. Specifically, write for simplicity $C^{\mathrm{o}}_{n,p} := C_{n,p}(\vartheta_1,\ldots,\vartheta_n;\widehat{\boldsymbol{\beta}})$ for the test statistic based on the original observations. Then parametric bootstrap critical points are calculated in practice as follows:

(i) Generate i.i.d. observations, $\{\vartheta^*_j, 1 \le j \le n\}$ from $F(\cdot;\widehat{\boldsymbol{\beta}})$.
(ii) Using the bootstrap observations $\{\vartheta^*_j, 1 \le j \le n\}$, obtain the bootstrap estimate $\widehat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}$.
(iii) Calculate the bootstrap test statistic, say $C^*_{n,p} := C_{n,p}(\vartheta^*_1,\ldots,\vartheta^*_n;\widehat{\boldsymbol{\beta}}^*)$.
(iv) Repeat steps (i) to (iii) a number of times, say B, and obtain $\{C^{*\mathrm{b}}_{n,p}\}_{b=1}^{\mathrm{B}}$.
(v) Calculate the critical point of a test of size α as the order $(1-\alpha)$ empirical quantile $C_{1-\alpha}$ of $\{C^{*\mathrm{b}}_{n,p}\}_{b=1}^{\mathrm{B}}$.

In next theorem we show that, under Assumptions A*, B* and C stated in the Appendix, this procedure provides a consistent estimator of the null distribution of the test statistic. With this aim, as in Section 2, we will assume that the estimator of $\widehat{\boldsymbol{\beta}}$ has a strong probability limit, say $\boldsymbol{\beta}^0$, even under alternatives. Let P_β denote the probability by assuming that the data come from $F(\cdot;\boldsymbol{\beta})$ and let $P_\star$ denote the bootstrap probability.

Theorem 2 *Assume that $\vartheta_1,\ldots,\vartheta_n$, are i.i.d. copies of Θ and that Assumptions A*, B* and C are fulfilled. Then,*

$$\sup_x \left| P_*(C^*_{n,p} \le x) - P_{\boldsymbol{\beta}^0}(C_{n,p} \le x) \right| \to 0 \quad a.s., \qquad as \;\; n \to \infty.$$

Theorem 2 holds whether the null hypothesis is true or not. In particular, if $\mathscr{H}_0$ is true, then it states that the bootstrap distribution and the null distribution of $C_{n,p}$ are close. Thus the test Ψ^*, which rejects the null when $C^o_{n,p} > C_{1-\alpha}$, is asymptotically correct in the sense that $\lim_{n\to\infty} P(\Psi^* = 1) = \alpha$, when the null hypothesis is true. Also an immediate consequence of (6) and Theorem 2 is that the test Ψ^* is consistent, that is $P(\Psi^* = 1) \to 1$, as $n \to \infty$, whenever $F \notin \mathscr{F}_{\boldsymbol{\beta}}$.

5 Tests for the von Mises distribution

5.1 Goodness-of-fit tests

For data distributed over the unit circle, the von Mises distribution (vMD), also called the Circular Normal distribution, is the pre-eminent model in circular data analysis when one has reason to believe the data might be symmetric and unimodal, much as the Normal distribution is on the real line. Sampling theory and inferential methods have been developed for this model, and as such it is a natural choice for our consideration. The density of the vMD with parameter vector $\boldsymbol{\beta} := (\mu, \kappa)$ is given by

$$f(\vartheta; \mu, \kappa) = \frac{1}{2\pi \mathscr{I}_0(\kappa)} e^{\kappa \cos(\vartheta - \mu)},\ 0 \le \vartheta < 2\pi, \tag{8}$$

where $\mathscr{I}_r(\cdot)$ denotes the modified Bessel function of the first kind of order r, and $0 \le \mu < 2\pi$ and $\kappa \ge 0$ are location and concentration parameters, respectively.

Our CF–based test utilizes the CF corresponding to (8) which is given by

$$\varphi(r; \mu, \kappa) = e^{\mathrm{i}r\mu} A_r(\kappa),\ r \in \mathbb{Z}, \tag{9}$$

where $A_r(\kappa) = \mathscr{I}_r(\kappa)/\mathscr{I}_0(\kappa)$.

Specifically the test statistic figuring in (4) may readily be written as

$$C_{n,p} = n \sum_{r=0}^{\infty} |\widehat{\varphi}_n(r) - \varphi(r; 0, \hat{\kappa})|^2 \, p(r) = S_1 + S_2 - 2S_3, \tag{10}$$

with $\widehat{\varphi}_n(r)$ the empirical CF of $\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_n$,

$$S_1 = \frac{1}{n} \sum_{j,k=1}^{n} \mathscr{E}_1(\widehat{\vartheta}_j - \widehat{\vartheta}_k), \tag{11}$$

$$S_2 = n \mathscr{E}_2(\widehat{\kappa}), \tag{12}$$

and

$$S_3 = \sum_{j=1}^{n} \mathscr{E}_3(\widehat{\vartheta}_j; \widehat{\kappa}), \tag{13}$$

where $(\widehat{\mu}, \widehat{\kappa})$ is a consistent estimator of the parameter (μ, κ), and $\widehat{\vartheta}_j = \vartheta_j - \widehat{\mu},\ j = 1, \ldots, n$. The series appearing in (11)-(13) are defined as

$$\mathscr{E}_1(\theta) = \sum_{r=0}^{\infty} \cos(\theta r) p(r),$$

$$\mathscr{E}_2(\kappa) = \sum_{r=0}^{\infty} A_r^2(\kappa) p(r),$$

and

$$\mathscr{E}_3(\theta;\kappa) = \sum_{r=0}^{\infty} \cos(\theta r) A_r(\kappa) p(r).$$

To proceed further note that all three series above may be viewed as expectations of corresponding quantities taken with respect to the law $p(r)$, and while these expectations are generally hard to obtain, they may be approximated by Monte Carlo by means of simulating i.i.d. variates from the law $p(r)$. In fact certain choices of $p(r)$ lead to closed form expressions, at least for the expectation in (11). Specifically if we let $p(r)$ be the Poisson law with parameter λ, we have

$$\mathscr{E}_1(\theta) = \cos(\lambda \sin\theta) e^{\lambda(\cos\theta - 1)}.$$

As for the calculation of S_2 and S_3 and since the corresponding series appearing in (12)-(13) converge rapidly, instead of Monte Carlo, we decided to approximate them by direct numerical computation of only a few terms. We have observed through simulations that summing up to $r = 100$ gives very accurate results. Strictly speaking this cut–off test is not universally consistent, but the practical effect on the power is negligible.

5.2 Estimation of parameters and a limit statistic

As for estimating parameters, we suggest the use of the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}} := (\widehat{\mu}, \widehat{\kappa})$ which is given by the following equations:

$$\frac{1}{n}\sum_{j=1}^{n} \sin(\vartheta_j - \widehat{\mu}) = 0, \;\; \frac{1}{n}\sum_{j=1}^{n} \cos(\vartheta_j - \widehat{\mu}) = A_1(\widehat{\kappa}). \tag{14}$$

It is well known that the MLE $\widehat{\mu}$ of μ satisfies $\widehat{\mu}(\vartheta_1 + a, \ldots, \vartheta_n + a) = \widehat{\mu}(\vartheta_1, \ldots, \vartheta_n) + a$, while the MLE $\widehat{\kappa}$ of κ satisfies $\widehat{\kappa}(\vartheta_1 + a, \ldots, \vartheta_n + a) = \widehat{\kappa}(\vartheta_1, \ldots, \vartheta_n)$, for each a, where the operations of addition in these equations are to be treated mod(2π) for circular data. Thus if one uses, instead of the original data $\vartheta_1, \ldots, \vartheta_n$, the centered data $\widehat{\vartheta}_j = \vartheta_j - \widehat{\mu},\ j = 1, \ldots, n$, then the distribution of any test statistic that depends on $\widehat{\mu}$ via $\widehat{\vartheta}_j,\ j = 1, \ldots, n$, will not depend on the specific parameter–value of μ, and hence without loss of generality we can set $\mu = 0$. On the other hand, since the concentration parameter κ is a shape parameter, it cannot be standardized out. Consequently the distribution of such a test always depends on the value of this parameter. One way out is to use the limit null distribution for fixed κ along with a look–up table with a sufficiently dense grid on κ. This approach is suggested in Lockhart and Stephens (1985), and is fairly accurate for most of the parameter space if based on the MLE of κ, but as already mentioned in

Section 4 we will instead use the parametric bootstrap which consistently estimates the limit null distribution of any given test uniformly over κ.

We close this section with an interesting limit statistic resulting from $C_{n,p}$ appearing in (10). To this end notice that since $\varphi_n(0) = \varphi(0) = 1$, the first term in $C_{n,p}$ vanishes regardless of the distribution being tested, while the second term also vanishes on account of (14) since we employ the MLEs as estimators of μ and κ. Now write $C_{n,\lambda}$ for the criterion in (10) with $p(r)$ being the Poisson probability function, with parameter λ. Then we have

$$C_{n,\lambda} = e^{-\lambda}\left(|\widehat{\varphi}_n(2) - A_2(\widehat{\kappa})|^2\,\frac{\lambda^2}{2} + \mathrm{o}(\lambda^2)\right),\ \lambda \to 0,$$

so that

$$\lim_{\lambda\to 0}\frac{2C_{n,\lambda}}{\lambda^2} = |\widehat{\varphi}_n(2) - A_2(\widehat{\kappa})|^2 := C_{n,0}. \tag{15}$$

Notice that the limit statistic $C_{n,0}$ only uses information on the CF of the underlying law as this is information is reflected on the corresponding empirical trigonometric moment of order $r = 2$.

On the other hand the test statistic $C_{n,\lambda}$ (and more generally $C_{n,p}$) uses an infinite weighted sum in which the empirical trigonometric moments of all integer orders $r \geq 0$ are accounted for. Thus the probability function $p(r)$ plays the role of a weight function that typically downweights the higher order terms which are known to be more prone to the periodic behavior intrinsically present in the empirical CF. A natural related question is whether there is some optimal choice for the probability function $p(\cdot)$. As asserted by Bugni et al. (2009) in a related context, the weight function cannot be selected empirically as this would require knowing how the true data-generation process differs from the parametric model. In this connection, and using the analogy with the choice of kernel in density estimation, prior experience has shown that the specific functional form of $p(\cdot)$ is not all that important. Carrying this analogy further, one suspects that the value of λ might have some sway over the results. Proper choice of λ however translates to a highly non–trivial analytic problem for which there are only a few results available in the literature; see Tenreiro (2009) and Meynaoui et al. (2019). This option is empirically investigated in the next section.

6 Finite-sample comparisons and simulations

This section summarizes the results of a simulation study, designed to evaluate the proposed GOF test for the vMD, and compare its performance with other existing tests. As competitors we include the Kuiper test and the Watson test for which there exist computationally convenient formulae; see for instance Section 7.2.1 of Jammalamadaka and SenGupta (2001). Specifically let $U_j = F(\vartheta_j; \widehat{\mu}, \widehat{\kappa})$ and write $U_{(j)},\ j = 1, \ldots, n$, for the

corresponding order statistics. Then we have

$$\mathrm{K} = \max_{1\le j\le n}\left\{U_{(j)} - \frac{j-1}{n}\right\} + \max_{1\le j\le n}\left\{\frac{j}{n} - U_{(j)}\right\}.$$

$$\mathrm{W} = \frac{1}{12n} + \sum_{j=1}^{n}\left(\left(U_{(j)} - \frac{2j-1}{2n}\right) - \left(\overline{U} - \frac{1}{2}\right)\right)^2,$$

where $\overline{U} = n^{-1}\sum_{j=1}^{n} U_j$.

We also include a test statistic based on the characterization of maximum entropy of the vMD suggested by Lund and Jammalamadaka (2000), denoted by E. These three criteria will be included in our Monte Carlo study. For our test statistic we took as $p(r)$ the probability function of a Poisson law with mean λ. This test is indexed by λ, and will be denoted by C_λ. We note that there exist alternative tests such as the conditional tests suggested by Lockhart (2012) (Lockhart, O'Reilly and Stephens, 2007, 2009), which we do not consider in our simulation study.

The simulated distributions are (i) the vMD, $vM(0,\kappa)$, (ii) mixtures of vMDs, $(1-\epsilon)vM(\mu_1,\kappa_1)+\epsilon\, vM(\mu_2,\kappa_2)$, $\epsilon\in(0,1)$, (iii) the generalized vMD, $GvM(\mu_1,\mu_2,\kappa_1,\kappa_2)$, with probability density function given by

$$f(\theta;\mu_1,\mu_2,\kappa_1,\kappa_2) = \frac{1}{2\pi G_0(\mu_1-\mu_2,\kappa_1,\kappa_2)}\exp\{\kappa_1\cos(\theta-\mu_1)+\kappa_2\cos(\theta-\mu_2)\},$$

where $G_0(\delta,\kappa_1,\kappa_2) = (1/2\pi)\int_0^{2\pi}\exp\{\kappa_1\cos(\theta)+\kappa_2\cos(\theta+\delta)\}d\theta$, (see Gatto and Jammalamadaka, 2007) and (iv) the wrapped normal distribution, $wn(\mu,\rho)$, with probability density function given by

$$f(\theta;\mu,\rho) = \frac{1}{2\pi}\left(1+2\sum_{m=-\infty}^{\infty}\rho^{p^2}\cos\{p(\theta-\mu)\}\right),$$

(Jammalamadaka and SenGupta, 2001, Ch. 2). Table 1 displays the specific alternatives (ii) and (iii), while the densities of such alternatives jointly with the density of the closer vMD (in the sense that the parameters are chosen so that they minimize the Kullback-Leibler distance), are depicted in Figure 1. These alternatives exhibit either bimodality and/or asymmetry and/or heavier tails than the vMD. We also considered several instances of the family of wrapped normal distributions, which are known to possess densities that are quite close to those of the vMD. This fact can be graphically appreciated by looking at Figure 2, which displays the probability density function of a $wn(0,\rho)$ law for $\rho = 0.1(0.1)0.9$, together with the density of the closer vMD distribution (in the sense explained before). Looking at this figure it becomes evident that it is rather hard to discriminate between these distributions and the vMD, particularly for small and large values of ρ.

Table 1: *Alternatives (ii) and (iii).*

Alternative 1	$0.9vM(\pi,5)+0.1vM(\pi/2,5)$
Alternative 2	$0.8vM(\pi,5)+0.2vM(\pi/2,5)$
Alternative 3	$0.65vM(\pi,5)+0.35vM(\pi/2,5)$
Alternative 4	$0.5vM(\pi,5)+0.5vM(\pi/2,5)$
Alternative 5	$(2/3)vM(\pi,3)+(1/3)vM(0.62\pi,3)$
Alternative 6	$(1/3)vM(\pi,8)+(2/3)vM(\pi,0.1)$
Alternative 7	$GvM(0,0.5,1,0.6)$
Alternative 8	$GvM(0,0.5,1,0.2)$

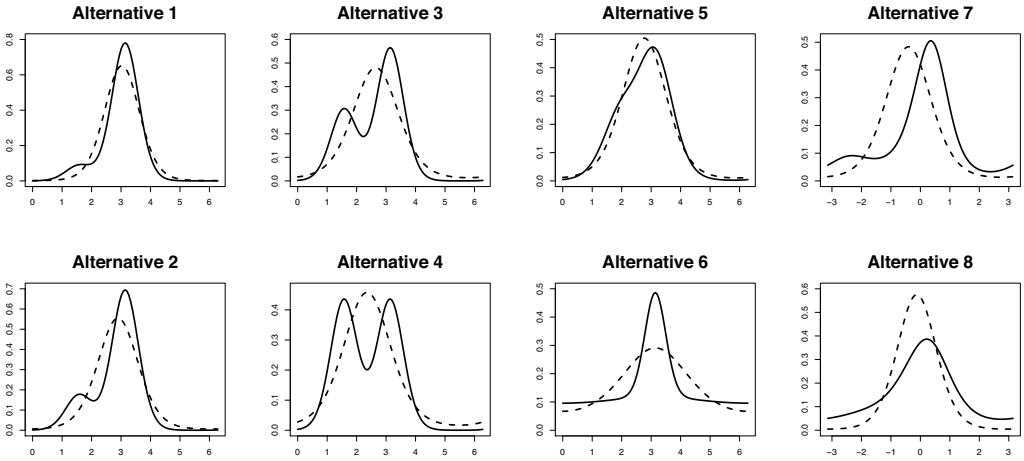


Figure 1: *Probability density function of alternatives in Table 1 (solid) and the probability density function of the closer vMD (dashed).*

All computations were performed using programs written in the R language. Specifically, we used the package CircStats for generating data from a vMD, and from mixtures of vMDs, and in order to calculate the MLEs of the parameters. Data from the generalized vMD were generated by the acceptance-rejection algorithm of von Neumann suggested in Gatto (2008). In all cases the p–values were approximated by using the parametric bootstrap algorithm given in Section 4 with B $= 1000$. For the benefit of potential users, we include the R codes necessary for calculating the new test statistics, in a Supplement.

We tried a wide range of values for λ and observed that the power of the proposed test depends on the value of λ. Tables 2 and 3 report the results for those values of λ giving the greater, or closer to the greater power, in all tried alternatives. Table 2 displays the observed proportion of rejections in 1,000 Monte Carlo samples of size $n = 25$ under the null hypothesis and for the set of alternatives in Table 1. We also tried $n = 50$ and $n = 100$ yielding a quite similar picture (in the sense of comparison between tests, but

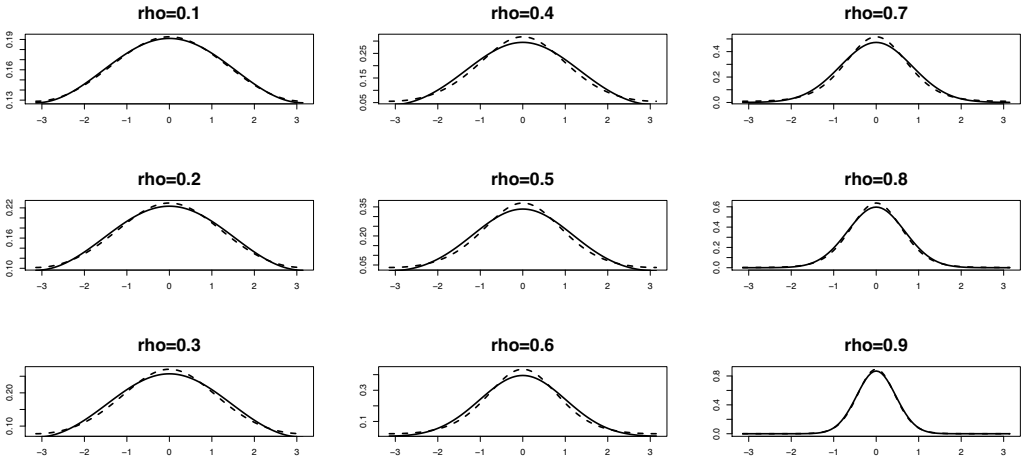


Figure 2: *Probability density function of a $wn(0,\rho)$ law for $\rho = 0.1(0.1)0.9$ (solid), and the probability density function of the closer vMD (dashed).*

Table 2: *Observed proportion of rejection in 1,000 Monte Carlo samples of size $n = 25$.*

Law	α	E	K	W	$C_{0.3}$	$C_{0.5}$	$C_{0.7}$	$C_{0.9}$	C_1
$vM(0,1)$	0.05	0.053	0.047	0.044	0.062	0.059	0.059	0.061	0.059
	0.10	0.105	0.093	0.089	0.115	0.117	0.117	0.122	0.122
$vM(0,5)$	0.05	0.054	0.048	0.047	0.033	0.032	0.036	0.040	0.043
	0.10	0.106	0.099	0.095	0.086	0.090	0.092	0.091	0.092
$vM(0,10)$	0.05	0.051	0.046	0.046	0.039	0.042	0.042	0.042	0.043
	0.10	0.103	0.090	0.092	0.093	0.096	0.095	0.095	0.098
Alt. 1	0.05	0.171	0.150	0.166	0.311	0.310	0.304	0.309	0.307
	0.10	0.267	0.235	0.272	0.450	0.451	0.445	0.443	0.437
Alt. 2	0.05	0.114	0.255	0.337	0.459	0.478	0.482	0.487	0.487
	0.10	0.197	0.422	0.470	0.631	0.634	0.645	0.635	0.627
Alt. 3	0.05	0.048	0.411	0.477	0.550	0.570	0.589	0.596	0.600
	0.10	0.097	0.547	0.620	0.720	0.737	0.747	0.749	0.742
Alt. 4	0.05	0.036	0.500	0.541	0.559	0.583	0.604	0.617	0.623
	0.10	0.059	0.627	0.688	0.719	0.739	0.741	0.750	0.751
Alt. 5	0.05	0.019	0.092	0.090	0.079	0.084	0.090	0.094	0.097
	0.10	0.056	0.151	0.163	0.176	0.184	0.195	0.209	0.211
Alt. 6	0.05	0.139	0.244	0.259	0.249	0.252	0.262	0.274	0.279
	0.10	0.243	0.358	0.397	0.379	0.390	0.397	0.410	0.409
Alt. 7	0.05	0.059	0.253	0.318	0.646	0.631	0.608	0.594	0.581
	0.10	0.102	0.381	0.465	0.774	0.757	0.737	0.721	0.713
Alt. 8	0.05	0.003	0.131	0.154	0.130	0.153	0.176	0.192	0.198
	0.10	0.007	0.212	0.244	0.267	0.305	0.320	0.329	0.337

Table 3: *Observed proportion of rejection in 1,000 Monte Carlo samples of size n from a $wn(0,\rho)$ law.*

n	ρ	α	E	K	W	$C_{0.3}$	$C_{0.5}$	$C_{0.7}$	$C_{0.9}$	C_1
50	0.3	0.05	0.060	0.052	0.059	0.081	0.081	0.078	0.072	0.072
		0.10	0.116	0.116	0.113	0.131	0.131	0.131	0.129	0.132
	0.4	0.05	0.053	0.053	0.053	0.072	0.072	0.073	0.070	0.069
		0.10	0.096	0.103	0.103	0.140	0.139	0.136	0.133	0.131
	0.5	0.05	0.041	0.072	0.072	0.099	0.096	0.096	0.097	0.095
		0.10	0.084	0.139	0.130	0.182	0.179	0.174	0.174	0.172
	0.6	0.05	0.035	0.069	0.072	0.089	0.091	0.087	0.090	0.088
		0.10	0.062	0.142	0.149	0.184	0.182	0.182	0.184	0.183
	0.7	0.05	0.019	0.079	0.092	0.098	0.098	0.098	0.103	0.103
		0.10	0.046	0.139	0.157	0.182	0.187	0.192	0.195	0.191
100	0.3	0.05	0.048	0.057	0.055	0.074	0.072	0.071	0.070	0.067
		0.10	0.092	0.114	0.109	0.144	0.143	0.140	0.138	0.139
	0.4	0.05	0.052	0.097	0.092	0.125	0.123	0.123	0.123	0.123
		0.10	0.102	0.149	0.175	0.212	0.210	0.211	0.208	0.203
	0.5	0.05	0.031	0.095	0.107	0.171	0.168	0.162	0.159	0.158
		0.10	0.067	0.162	0.194	0.272	0.269	0.264	0.262	0.261
	0.6	0.05	0.030	0.106	0.122	0.203	0.196	0.185	0.176	0.173
		0.10	0.049	0.185	0.195	0.316	0.310	0.302	0.283	0.279
	0.7	0.05	0.021	0.117	0.108	0.162	0.159	0.157	0.153	0.153
		0.10	0.040	0.190	0.193	0.285	0.284	0.275	0.262	0.254

with greater powers as the sample size increases), and therefore we omit those results. By contrast, and since the power for $n = 25$ is quite low we opted to present results for larger sample size for wrapped normal alternatives. Specifically Table 3 presents the results for wrapped normal alternatives for sample size $n = 50$ and $n = 100$, and $\rho = 0.3(0.1)0.7$.

Regarding level, we conclude that the observed empirical rejection rates are reasonably close to the nominal values. In fact, for larger sample sizes (not displayed), we observed greater closeness. As for power, we observe that the power of the proposed test is comparable and most often greater than that of the tests based on the empirical CDF. On the other hand, the test based on the characterization of maximum entropy presents the poorest performance under the considered alternatives.

A natural question is which value of λ should be used in practical applications. Although the powers exhibited in the tables are quite close for the values of λ selected, it seems that $C_{0.5}$ has an intermediate behaviour in all tried cases, so we recommend $\lambda = 0.5$ as a compromise choice.

Another possibility is to choose λ by using some data-dependent method (see Cuparić, Milosević and Obradović, 2019, for a related approach). In this sense, Tenreiro (2019) has proposed a method for choosing the tuning parameter λ so that the power is maximized. It works as follows. Let $C_{n,\lambda}(\alpha)$ denote the upper α percentile of the null distribution of $C_{n,\lambda} = C_{n,\lambda}(\vartheta_1,\ldots,\vartheta_n)$. Assume that $\lambda \in \Lambda$, with Λ having a finite number of points. Then, reject H_0 if

$$\max_{\lambda\in\Lambda}\{C_{n,\lambda} - C_{n,\lambda}(u)\} > 0,$$

where u is chosen so that the test has level α. The key point is the way to determine u. In the context discussed in Tenreiro (2019), it is assumed that the exact null distribution of the test statistic can be calculated (or at least it can be approximated by simulation). Since this is not our case, we have adapted his procedure to calculate u to our setting as follows:

1. First, we must approximate the critical points $C_{n,\lambda}(u)$, $u \in (0,1)$, $\lambda \in \Lambda$. With this aim, we generate B_1 bootstrap samples and estimate $C_{n,\lambda}(u)$ by means of their bootstrap analogues, $C^*_{1,n,\lambda}(u)$, for $u \in \{1/B_1, 2/B_1, \ldots, (B_1-1)/B_1\} := U_{B_1}$, $\lambda \in \Lambda$.
2. Then, we must calibrate u so that the test has level α. For this purpose, we generate B_2 bootstrap samples, independently of those generated in the first step, and determine $u^* \in U_{B_2}$ such that

$$P_*\left(\max_{\lambda\in\Lambda}\{C^*_{n,\lambda} - C^*_{1,n,\lambda}(u^*)\} > 0,\right) \leq \alpha.$$

3. Finally,

$$\text{reject } H_0 \text{ if} \quad \max_{\lambda\in\Lambda}\{C_{n,\lambda} - C^*_{1,n,\lambda}(u^*)\} > 0. \tag{16}$$

In addition to the determination of u, another delicate issue is the choice of the set Λ, which has a strong effect on the power of the resulting test. In order to study the practical behaviour of test (16), we repeated the experiment in Table 2 for $\Lambda = \Lambda_1$ and $\Lambda = \Lambda_2$, with $\Lambda_1 = \{0.1, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5, 7, 10\}$ and $\Lambda_2 = \{0.3, 0.5, 0.7, 0.9, 1, 2\}$, and $B_1 = B_2 = 1000$. Table 4 display the results obtained. Comparing the powers in that table with those in Table 2 we conclude that as Λ increases, the power of the test (16) decreases. This fact was also observed in the simulations in Tenreiro (2019). The power for $\Lambda = \Lambda_2$ is in most cases smaller than that obtained for $\lambda = 0.5$.

Table 4: *Observed proportion of rejection in 1,000 Monte Carlo samples of size $n = 25$, for $\alpha = 0.05$.*

	Alt1	Alt2	Alt3	Alt4	Alt5	Alt6	Alt7	Alt8
Λ_1	0.200	0.327	0.426	0.442	0.050	0.213	0.458	0.103
Λ_2	0.280	0.439	0.549	0.563	0.077	0.280	0.562	0.156

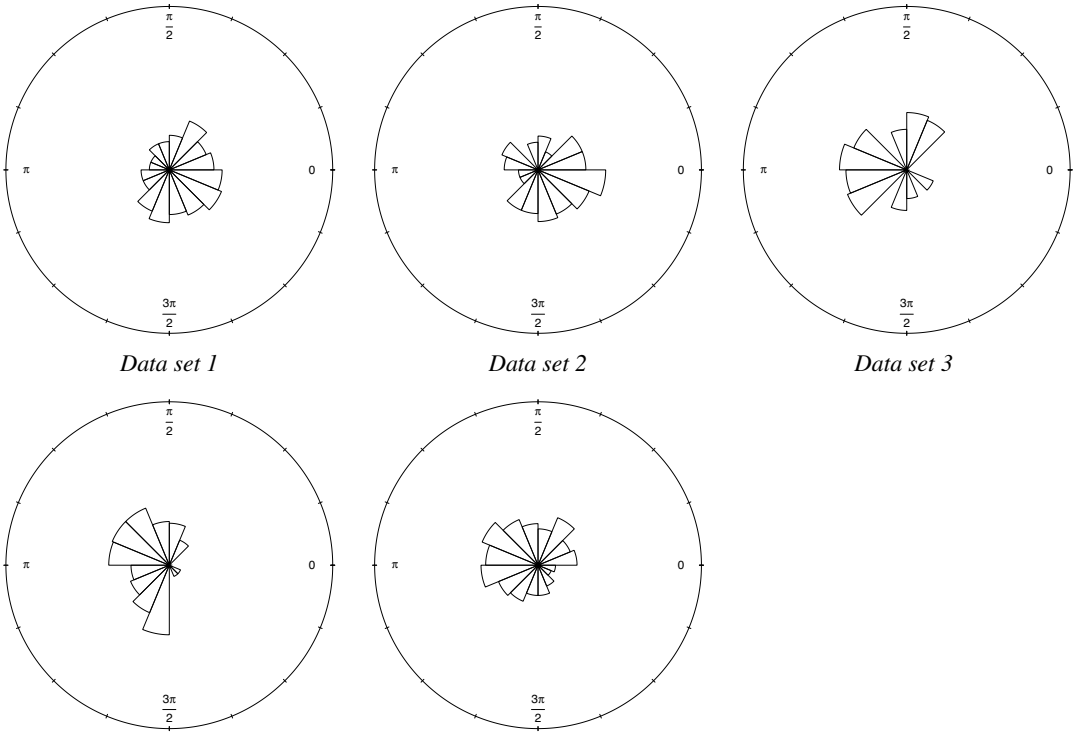


Figure 3: *Rose diagrams for the five real data sets.*

7 Real-data application

This section illustrates the proposed test on five real data sets. They come from a study by Taylor and Burns (2016) on the radial orientation of 2 species of mistletoes and 3 species of epiphytes, which the ecologists believe orient towards the direction of the availability of light and humidity. Specifically, Data Set 1 consists of $n = 67$ observations on *Peraxilla colensoi*, Data Set 2 consists of $n = 70$ observations on *Peraxilla tetrapetala*, Data Set 3 consists of $n = 65$ observations on *Asplenium flaccidum*, Data Set 4 consists of $n = 182$ observations on *Hymenophyllum multifidum*, and Data Set 5 consists of $n = 263$ observations on *Notogrammitis billardierei*. Taylor and Burns (2016) tested for uniformity in the five data sets and in all cases such hypothesis was rejected, indicating that the distribution of each of the studied species have certain orientation, as can be easily appreciated by looking at Figure 3, which displays the rose diagrams for each data set. So, it would be interesting to check if the data follow some distribution, such as the vMD. In fact, Taylor and Burns (2016) calculated certain confidence intervals based on the vMD. Table 5 reports the values of the maximum likelihood estimates by assuming a vMD, as well as the p-values for testing goodness-of-fit to that distri-

Table 5: *Maximum likelihood estimators of the parameters and p-values for the real data sets.*

	$\hat{\mu}$	$\hat{\kappa}$	K	W	$C_{0.5}$
1	2.5551	0.7700	0.5335	0.6410	0.4710
2	5.7677	0.8447	0.1505	0.2815	0.7555
3	2.8226	1.1120	0.0080	0.0050	0.0265
4	3.0454	1.2589	0.0080	0.0050	0.0220
5	2.5551	0.7699	0.8310	0.0060	0.0050

bution that resulted by applying the tests K, W and $C_{0.5}$. These three test criteria lean towards the null hypothesis for Data Set 1 and Data Set 2, and all of them suggest that the vMD is not a good model for Data Set 3 and Data Set 4. For Data Set 5, the tests W and $C_{0.5}$ reject that the vMD provides an adequate description of the data, while test K concludes in the opposite direction. From the power results in our simulations, we deduce that the vMD does not provide a satisfactory fit to Data Set 5.

8 Discussion

We suggest here a general class of GOF tests for circular distributions. The proposed test statistic may conveniently be expressed as a weighted L2–type distance between the empirical trigonometric moments and the corresponding theoretical quantities, and is shown to compete well with classical tests based on the CDF. Our method imposes minimal technical conditions is widely applicable for arbitrary distributions under test. Here however we focus specifically on GOF testing for the vMD because it is one of the most commonly used distributions in practice, and one would like to verify if this model fits a given data set before utilizing the various parametric tools that have been developed for this particular model.

A Appendix

All limits are understood to be taken as $n \to \infty$.

A.1 Technical assumptions

ASSUMPTION A. Under $\mathscr{H}_0$, if $\boldsymbol{\beta} \in \mathscr{B}$ denotes the true parameter value, then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L(\vartheta_j; \boldsymbol{\beta}) + o_P(1),$$

with $\mathbb{E}\{L(\Theta; \boldsymbol{\beta})\} = 0$ and $J(\boldsymbol{\beta}) = \mathbb{E}\{L(\Theta; \boldsymbol{\beta}) L(\Theta; \boldsymbol{\beta})^\top\} < \infty$.

ASSUMPTION B. $\frac{\partial}{\partial \beta_k}\Re\varphi(r;\boldsymbol{\beta})$ and $\frac{\partial}{\partial \beta_k}\Im\varphi(r;\boldsymbol{\beta})$, exist $\forall r \in \mathbb{N}_0$ and $1 \le k \le p$, and satisfy

$$\sum_{r\ge 0}\frac{\partial}{\partial \beta_k}\Re\varphi(r;\boldsymbol{\beta})^2 p(r) < \infty,$$
$$\sum_{r\ge 0}\frac{\partial}{\partial \beta_k}\Im\varphi(r;\boldsymbol{\beta})^2 p(r) < \infty.$$

Let $\|\cdot\|$ stand for the Euclidean norm.

ASSUMPTION C. For any $\varepsilon > 0$ there is a bounded neighborhood $\mathscr{N}_\varepsilon \subseteq \mathbb{R}^p$ of $\boldsymbol{\beta}$, such that if $\boldsymbol{\gamma} \in \mathscr{N}_\varepsilon$ then $\nabla\Re\varphi(r;\boldsymbol{\gamma})$ and $\nabla\Im\varphi(r;\boldsymbol{\gamma})$ exist and satisfy

$$\|\nabla\Re\varphi(r;\boldsymbol{\gamma}) - \nabla\Re\varphi(r;\boldsymbol{\beta})\| \le \rho_\Re(r),\ \forall r \in \mathbb{N}_0, \quad \text{with} \quad \sum_{r\ge 0}\rho_\Re^2(r)p(r) < \varepsilon,$$
$$\|\nabla\Im\varphi(r;\boldsymbol{\gamma}) - \nabla\Im\varphi(r;\boldsymbol{\beta})\| \le \rho_\Im(r),\ \forall r \in \mathbb{N}_0, \quad \text{with} \quad \sum_{r\ge 0}\rho_\Im^2(r)p(r) < \varepsilon.$$

Assumptions A* and B* below are a bit stronger than Assumptions A and B, respectively. They are required for the consistency of the parametric bootstrap null distribution estimator.

ASSUMPTION A*. (a) There is a $\boldsymbol{\beta}^0 \in \mathscr{B}$ so that $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^0$, a.s., $\boldsymbol{\beta}^0$ being the true parameter value if $\mathscr{H}_0$ is true,
(b)

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L(\vartheta_j^*;\widehat{\boldsymbol{\beta}}) + o_{P_*}(1),$$

with $\mathbb{E}_*\{L(\Theta^*;\widehat{\boldsymbol{\beta}})\} = 0$, $J(\widehat{\boldsymbol{\beta}}) = \mathbb{E}_*\{L(\Theta^*;\widehat{\boldsymbol{\beta}})L(\Theta^*;\widehat{\boldsymbol{\beta}})^\top\} \to J(\boldsymbol{\beta}^0) < \infty$, a.s.
(c) $\sup_{\boldsymbol{\beta}\in\mathscr{N}_0}\mathbb{E}_{\boldsymbol{\beta}}\left[\|L(\Theta;\boldsymbol{\beta})\|_{\ell_p^2}^2 I\left\{\|L(\Theta;\boldsymbol{\beta})\|_{\ell_p^2} > \epsilon\sqrt{n}\right\}\right] \longrightarrow 0$, $\forall \epsilon > 0$, where $\mathscr{N}_0 \subseteq \mathscr{B}$ is an open neighborhood of $\boldsymbol{\beta}_0$, where $\mathbb{E}_{\boldsymbol{\beta}}$ stands for the expectation when data have CDF $F(x;\beta)$.

ASSUMPTION B*. Assumption B holds true $\forall\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}^0$, where $\boldsymbol{\beta}^0$ is as defined in Assumption A*.

A.2 Proofs

Proof of Theorem 1

By Taylor expansion,

$$\Re\varphi(r;\widehat{\boldsymbol{\beta}}) = \Re\varphi(r;\boldsymbol{\beta}) + \nabla\Re\varphi(r;\boldsymbol{\beta})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + g_{1n}(r).$$

From Assumptions A and C, it follows that

$$\|\sqrt{n}g_{1n}\|_{\ell_p^2}^2 = o_P(1).$$

From Assumptions A and B, it follows that

$$\nabla\Re\varphi(r;\boldsymbol{\beta})^\top(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \nabla\Re\varphi(r;\boldsymbol{\beta})^\top\frac{1}{n}\sum_{j=1}^{n}L(\vartheta_j;\boldsymbol{\beta}) + g_{2n}(r)$$

with

$$\|\sqrt{n}g_{2n}\|_{\ell_p^2}^2 = o_P(1).$$

Analogous expansions hold for $\Im\varphi(r;\widehat{\boldsymbol{\beta}})$, so that if we let

$$Z_{0,n}(r) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Upsilon(r,\vartheta_j;\boldsymbol{\beta}),$$

these expansions imply that

$$Z_n(r) = Z_{0,n}(r) + g_{3n}(r), \tag{17}$$

with

$$\|g_{3n}\|_{\ell_p^2}^2 = o_P(1). \tag{18}$$

From Assumptions A and B, it follows that $\mathbb{E}_{\boldsymbol{\beta}}\left\{\|\Upsilon(\cdot,\Theta;\boldsymbol{\beta})\|_{\ell_p^2}^2\right\} < \infty$. Therefore, by applying the Central Limit Theorem in Hilbert spaces (van der Vaart and Wellner, 1996, p. 50), we get that

$$Z_{0,n} \xrightarrow{\mathscr{L}} Z, \tag{19}$$

and then the result follows from (17)–(19). ■

Proof of Theorem 2

Let $Z_n^*(r) = \sqrt{n}\{\varphi_n^*(r) - \varphi(r;\widehat{\boldsymbol{\beta}}^*)\}$, with $\widehat{\varphi}_n^*(r) = n^{-1}\sum_{j=1}^{n}e^{\mathrm{i}r\vartheta_j^*}$. Proceeding as in the proof of Theorem 1, we have that

$$Z_n^*(r) = Z_{0,n}^*(r) + g_n^*(r),$$

with $Z_{0,n}^*(r) = n^{-1/2}\sum_{j=1}^{n}\Upsilon(r,\vartheta_j^*;\widehat{\boldsymbol{\beta}})$,

$$\|g_n^*\|_{\ell_p^2}^2 = o_{P_*}(1), \quad a.s.$$

To prove the result we derive the asymptotic distribution of $Z^*_{0,n}(r)$, showing that it coincides with the asymptotic distribution of $C_{n,p}$ when the data come from $F(\cdot;\boldsymbol{\beta}^0)$. Notice that, for each n, the elements in the set $\{\Upsilon(\cdot,\vartheta^*_1;\widehat{\boldsymbol{\beta}}),\ldots,\Upsilon(\cdot,\vartheta^*_n;\widehat{\boldsymbol{\beta}})\}$ are independent and identically distributed random elements taking values in the separable Hilbert space ℓ^2_p, but their common distribution may vary with n. Because of this reason, in order to derive the asymptotic distribution of $Z^*_{0,n}(r)$, we apply Theorem 1.1 in Kundu, Majumdar and Mukherjee (2000). So we will prove that conditions (i)–(iii) in that theorem hold. For $k\geq 0$, let $e_k(j)=I(k=j)/\sqrt{p(k)}$. $\{e_k\}_{k\geq 0}$ is an orthonormal basis of ℓ^2_p.

Let $\mathscr{C}_n$ and $\mathscr{K}_n$ denote the covariance operator and the covariance kernel of $Z^*_{0,n}$, respectively. Let $\mathscr{C}_0$ and $\mathscr{K}_0$ denote the covariance operator and the covariance kernel of Z_0, respectively, where Z_0 stands for the random element figuring in Theorem 1 with $\boldsymbol{\beta}=\boldsymbol{\beta}^0$. Assumptions A* and C imply that

$$\langle \mathscr{C}_n e_k, e_r\rangle_{\ell^2_p} = \sqrt{p(k)p(r)}\mathscr{K}_n(k,r) \to \sqrt{p(k)p(r)}\mathscr{K}_0(k,r) = \langle \mathscr{C}_0 e_k, e_r\rangle_{\ell^2_p}, \quad a.s.,$$

Setting $a_{kr}=\langle \mathscr{C}_0 e_k, e_r\rangle_{\ell^2_p}$ in the aforementioned Theorem 1.1, this proves that condition (i) holds.

Assumptions A*, B* and C imply that

$$\sum_{k\geq 0}\langle \mathscr{C}_n e_k, e_k\rangle_{\ell^2_p} = \sum_{k\geq 0}\mathscr{K}_n(k,k)p(k) \to \sum_{k\geq 0}\mathscr{K}_0(k,k)p(k) = \mathbb{E}\left\{\|Z_0\|^2_{\ell^2_p}\right\} < \infty, \quad a.s.,$$

and thus condition (ii) holds. Finally, condition (iii) readily follows from Assumption A*. ■

Acknowledgements

The authors thank the anonymous referees and the editor for their constructive comments and suggestions which helped to improve the presentation. MD Jiménez-Gamero has been partially supported by grant MTM2017-89422-P of the Spanish Ministry of Economy, Industry and Competitiveness, the State Agency of Investigation, and the European Regional Development Fund.

References

Bugni, F.A., Hall, P., Horowitz, J.L. and Neumann, G.R. (2009). Goodness-of-fit tests for functional data. *Econometrics Journal*, 12, S1–S18.

Cuparić, M., Milosević, B. and Obradović, M. (2019) New L2-type exponentiality tests. *SORT*, 43, 25–50.

Epps, T.W. and Pulley, L.B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70, 723–726.

Gatto, R. (2008). Some computational aspects of the generalized von Mises distribution. *Statistics and Computing*, 18, 321–331.

Gatto, R. and Jammalamadaka, S.R. (2007). The generalized von Mises distribution. *Statistical Methodology*, 4, 341–353.

Gürtler, N. and Henze, N. (2000). Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of the Institute of Statistical Mathematics*, 52, 267–286.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Kuiper, N.H. (1960). Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A,* 63, 38–47.

Kundu, S., Majumdar, S. and Mukherjee, K. (2000). Central limit theorems revisited. *Statistics & Probability Letters*, 47, 265–275.

Lockhart, R.A. (2012). Conditional limit laws for goodness-of-fit tests. *Bernoulli*, 18, 857–882.

Lockhart, R.A., O'Reilly, F. and Stephens, M.A. (2007). The use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika*, 94, 992–998.

Lockhart, R.A., O'Reilly, F. and Stephens, M.A. (2009). Exact conditional tests and approximate bootstrap tests for the von Mises distribution. *Journal of Statistical Theory and Practice*, 3, 543–554.

Lockhart, R.A. and Stephens, M.A. (1985). Tests of fit for the von Mises distribution. *Biometrika*, 72, 647–652.

Lund, U. and Jammalamadaka, S.R. (2000). An entropy–based test for goodness–of–fit of the von Mises distribution. *Journal of Statistical Computation and Simulation*, 67, 319–332.

Matsui, M. and Takemura, A. (2008). Goodness-of-fit tests for symmetric stable distributions–Empirical characteristic function approach. *Test*, 17, 546–566.

Meintanis, S.G. (2005) Consistent tests for symmetric stability with finite mean based on the empirical characteristic function. *Journal of Statistical Planning and Inference*, 128, 373–380.

Meynaoui, A., Mélisande, A., Laurent-Bonneau, B. and Marrel, A. (2019) Adaptive tests of independence based on HSIC measures. https://arxiv.org/abs/1902.06441

Taylor, A. and Burns, K. (2016). Radial distributions of air plants: a comparison between epiphytes and mistletoes. *Ecology*, 97, 819–825.

Tenreiro, C. (2009). On the choice of the smoothing parameter for the BHEP goodness-of-fit test. *Computational Statistics & Data Analysis*, 53, 1038–1053.

Tenreiro, C. (2019). On the automatic selection of the tuning parameter appearing in certain families of goodness-of-fit tests. *Journal of Statistical Computation and Simulation*, 89, 1780–1797.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Watson, G.S. (1961). Goodness-of-fit tests on the circle. *Biometrika*, 48, 109–114.

SORT 43 (2) July-December 2019, 337-354 DOI: 10.2436/20.8080.02.91

Data envelopment analysis efficiency of public services: bootstrap simultaneous confidence region

Jesús A. Tapia[1], Bonifacio Salvador[1] and Jesús M. Rodríguez[2]

Abstract

Public services, such as higher education, medical services, libraries or public administration offices, provide services to their customers. To obtain opinion-satisfaction indices of customers, it would be necessary to survey all the customers of the service (census), which is impossible. What is possible is to estimate the indices by surveying a random customer sample. The efficiency obtained with the classic data envelopment analysis models, considering the opinion indices of the customers of the public service as output data estimated with a user sample, will be an estimation of the obtained efficiency if the census is available. This paper proposes a bootstrap methodology to build a confidence region to simultaneously estimate the population data envelopment analysis efficiency score vector of a set of public service-producing units, with a fixed confidence level and using deterministic input data and estimated customer opinion indices as output data. The usefulness of the result is illustrated by describing a case study comparing the efficiency of libraries.

MSC: 62D05, 90C99.

Keywords: Data envelopment analysis, sampling survey research, public sector, bootstrap, simultaneous confidence region.

1 Introduction

Data envelopment analysis (DEA) is clearly of enormous potential in measuring public sector efficiency, particularly in areas where there exists a large number of agencies to compare (see Smith and Mayston, 1987 or chapter 15 of Cooper, Seiford and Zhu, 2011). In this context, DEA efficiency is usually evaluated using determinist input/output data. However, the quality of the service delivered by a provider can therefore have important implications and available results. Bayraktar et al. (2012), Witte and Geys (2013), Mayston (2015, 2017), Santín and Sicilia (2017) and Førsund (2017) analyse the efficiency of any individual public services producer, where output

[1] Department of Statistic and Operative Research, University of Valladolid, Campus Miguel Delibes, Paseo Belén nº 7,47011 Valladolid, Spain.

[2] Statistical & Budgetary General Management of the Regional Government of Castile & Leon, Regional Tax Office, José Cantalapiedra nº 2, 47014 Valladolid, Spain

Received: July 2018

Accepted: October 2019

variables can include the quality of the service produced and the said quality is measured by a consumer satisfaction survey. Tapia, Salvador and Rodríguez (2018) study the relationship between customer sample size and accuracy in estimating the efficiency of public services. Of course, a public service is more efficient when, with its resources, it is able to achieve the highest opinion-satisfaction of its users. Customer opinion-satisfaction surveys are widely used tools to measure the perception of the quality of the service (Parasuraman, Berry and Zeithaml, 1993), and to obtain outputs and the DEA efficiency scores (Lee and Kim, 2014). In this paper, using confidence regions, we estimate the DEA efficiency of a fixed (not random) set of public service-producing units, i.e., our decision-making units (DMUs). We do so using indices of the service quality obtained as the mean of the answers given by the sample of surveyed people in the opinion-satisfaction survey as the data output, and the resources of the services measured in a deterministic way as the data input. For example, when comparing the DEA efficiency of all the cinemas in a city, user opinion on the quality of each cinema is measurable with opinion indices estimated using a survey of the cinemagoers to know their satisfaction with the location, the staff, the state of the cinema, etc. The number of seats, screens in the cinema, daily movies on show, or monthly premiering movies are the resources of the service.

The studies where the DEA efficiency is evaluated in the presence of sampling information have had two approaches until now. In the first one, the set of DMUs from which input/output information is known is considered as a sample of a population of DMUs and the randomness comes from the DMU sample. In this approach, Banker (1993), Simar and Wilson (1998, 2000, 2007, 2011, 2013, 2015), Kneip, Park and Simar (1998) and Kneip, Simar and Wilson (2008), have proved statistical properties of the nonparametric estimators used to estimate the productivity efficiency of DMUs, derived the asymptotic distribution of DEA estimators and tested hypotheses about the structure of the underlying nonparametric model.

In the second approach, samples are used to estimate input and/or output data. The efficiency is evaluated using linear programming (LP) problems subject to constraints defined in terms of probability, or chance-constrained problems. A great number of papers have reported a wide range of uses of chance-constrained programming, including: Charnes and Cooper (1959, 1963), Land, Lovell and Thore (1993), Olesen and Petersen (1995), Cooper, Huang and Li (1996), Cooper et al. (2002), Huang and Li (2001), Wu and Olson (2008), Khodabakhshi and Asgharian (2009), Khodabakhshi (2010), Wu and Lee (2010), Wu (2010) and Tavana, Shiraz and Hatami-Marbini (2014). Charles and Kumar (2014) introduced a chance-constrained model to measure the stochastic efficiency of the service quality.

In this paper, we assume a fixed set of homogeneous DMUs in terms of the nature of the operations they perform, the measures of their efficiency, and the conditions under which they operate, as in the classic DEA models (Charnes, Cooper and Rhodes, 1978). The randomness comes solely from the customer sample in each DMU with which we estimate the output data. However, to evaluate the DEA efficiency with a

bootstrap confidence region, we do not use chance-constrained programming, only the classic DEA models with constant (CCR) and variable returns-to-scale (BCC), i.e., LP problems subject to deterministic constraints.

Using estimated output data with a sample of customers instead of population output data causes an estimation error to be transferred to the evaluation of the DEA efficiency (Ceyhan and Benneyan, 2014). The vector of DEA efficiency scores obtained is, therefore, an estimation of the vector of the population DEA efficiency scores that would be obtained if we had the customer population data, i.e., if the output data were obtained with a customer census in each DMU. In our study, we solve the problem of determining how many customers need to be surveyed in order to estimate the output data in each DMU, with a previously fixed estimation error, when estimating the vector of population DEA efficiency scores with a bootstrap simultaneous confidence region. With the same assumptions as in this study, Tapia et al. (2018) obtained the customer sample size needed in each DMU to estimate the population DEA efficiency with a fixed accuracy in each DMU; while in this paper, the customer sample size necessary in each public service-producing unit is determined so that the maximum efficiency estimation error in the service-producing units will be smaller than a previously fixed value.

Using bootstrap, smooth bootstrap or double-smooth bootstrap methodologies to evaluate the efficiency of the public sector with confidence intervals is not new (Simar and Wilson, 1998, 2000, Simar and Zelenyuk, 2006, Kneip, Simar and Wilson, 2011). For instance, these methodologies have been used to measure the efficiency in health care (Tsekouras et al., 2010, Chowdhury and Zelenyuk, 2016), universities and research institutes (Barra and Zotti, 2016), government (Benito, Solana and Moreno, 2014), public libraries (Liu and Chuang, 2009), schools (Essid, Ouellette and Vigeant, 2014, Alexander, Haug and Jaforullah, 2010), tourism (Assaf and Agbola, 2011), banks (Casu and Molyneux, 2003) or public transport services (Assaf, 2010, Gil, Turias and Cerbán, 2019). In all these references, the different bootstrap resampling techniques are used considering the observed DMUs to be a sample taken from a population of DMUs and the resampling is done over the estimated efficiencies. In our study, we consider a fixed (not random) set of services, a customer sample in each service to estimate the client opinion indices (outputs) and a bootstrap resampling on the customer sample. As far as we know, the bootstrap efficiency simultaneous confidence region introduced in this paper has not been attempted in the literature. Our confidence region is the product of intervals and these intervals allow efficiency rankings, dominance relations and efficiency bounds to be determined as in Salo and Punkka (2011).

The rest of the paper is organized as follows. The problem is introduced in Section 2. Section 3 studies the determination of the customer sample size in each public service, in order to achieve a fixed accuracy in the simultaneous DEA efficiency estimation. In Section 4, a bootstrap simultaneous confidence region to estimate the population DEA efficiency in a fixed set of public services is determined. Section 5 contains an application of the proposed approach using real inputs and opinion indices estimated with a user sample (output data) of 15 libraries. Finally, the main conclusions are given.

2 Preliminaries

Consider a fixed set of M service-producing units, our DMUs, m resources of the services as known inputs $\mathbf{X}_j = (x_{1j},\ldots,x_{mj})$; $j = 1,\ldots,M$ and s customer opinion indices as unknown outputs. We distinguish between the population and sampling contexts. As for the population context, we consider $\mathbf{U}_j = \left(\mathbf{U}_{1j},\ldots,\mathbf{U}_{N_j j}\right); j = 1,\ldots,M$ the opinion of all of the N_j customers of the jth DMU (DMU$_j$ for short). Each $\mathbf{U}_{kj} = (U_{k1j},\ldots,U_{ksj})$ is the quantitative answer of the kth customer $(k = 1,\ldots,N_j)$ of the DMU$_j$ $(j = 1,\ldots,M)$ to the s opinion items. The output data $\mathbf{Y}_j$ in the DMU$_j$ is $g(\mathbf{U}_j)$ where $g : \Re^{N_j \times s} \longrightarrow \Re^s$. In this paper we consider g as the sample mean $\mathbf{Y}_j = \left(\frac{\sum_{k=1}^{N_j} U_{k1j}}{N_j},\ldots,\frac{\sum_{k=1}^{N_j} U_{ksj}}{N_j}\right)$. The LP model CCR or BCC with the output orientation of Table 1 (CCR-O or BCC-O), taking data $\{(\mathbf{X}_j,\mathbf{Y}_j)\}_{j=1,\ldots,M}$, determines the population DEA efficiency scores $\{\varphi_j\}_{j=1,\ldots,M}$. The output orientation is selected because the interest is to know which services, with their resources, can improve the opinion indices of their customers. Keeping in mind the impossibility of getting the opinion of all the population of N_j customers of the DMU$_j$, the outputs $\mathbf{Y}_j$ and the efficiencies φ_j , $j = 1,\ldots,M$, are unknown.

Table 1: *DEA models with constant (CCR) and variable (BCC) returns-to-scale; output orientation.*

$$max \quad \varphi + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$$

s.t.

$$\sum_{j=1}^{M} \lambda_j y_{rj} - s_r^+ = \varphi y_{ro}, \;\; r = 1,\ldots,s$$

$$\sum_{j=1}^{M} \lambda_j x_{ij} + s_i^- = x_{io}, \;\; i = 1,\ldots,m$$

CCR-O

$$\lambda_j \geq 0, s_i^- \geq 0, s_r^+ \geq 0; \;\; j = 1,\ldots,M; \; i = 1,\ldots,m; \;\; r = 1,\ldots,s \qquad (1)$$

BCC-O

$$\sum_{j=1}^{M} \lambda_j = 1 \qquad (2)$$

where s_i^- and s_r^+ are slack variables and $\varepsilon > 0$ is a non-Archimedian element.

In the sampling context, in the DMU$_j$, we take a random customer sample $\left(\mathbf{U}_{1j},\ldots,\mathbf{U}_{n_j j}\right) \subset \mathbf{U}_j$ of size n_j and we estimate the output data $\mathbf{Y}_j$ by $\widehat{\mathbf{Y}}_j = \left(\frac{\sum_{k=1}^{n_j} U_{k1j}}{n_j},\ldots,\frac{\sum_{k=1}^{n_j} U_{ksj}}{n_j}\right)$. We denote by $\widehat{\mathbf{y}}_j$ the observed value of this estimator. The LP model (1) or (2), taking $\left\{\left(\mathbf{X}_j,\widehat{\mathbf{Y}}_j\right)\right\}_{j=1,\ldots,M}$ as input-output data, determines the estimators $\{\widehat{\varphi}_j\}_{j=1,\ldots,M}$ of the population efficiency scores $\{\varphi_j\}_{j=1,\ldots,M}$, understanding that the model is maximized with the data $\{(\mathbf{X}_j,\widehat{\mathbf{y}}_j)\}_{j=1,\ldots,M}$ to obtain the estimates $\{\widehat{\omega}_j\}_{j=1,\ldots,M}$. Tapia et al. (2018)

prove that the estimator $\widehat{\varphi}_j$ is statistically consistent in the particular case of the CCR-O model with one known input and one estimated output.

Therefore, our statistical model (Ω, P) corresponds to independent, random samples in each DMU, that is, the sample space is $\Omega = \prod_{j=1}^{M} \Omega_j$, where $\Omega_j = \{\text{samples } u_j \text{ of sample size } n_j \text{ in the DMU}_j\}$, and the probability P depends on the sample design used.

The problem in this paper is to estimate the population efficiency scores vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_M)$ with a simultaneous confidence region. Formally, for any $\delta \in (0,1)$ and $\alpha \in (0,1)$, we then calculate the customer sample size n_j in the DMU$_j$, $j = 1, \ldots, M$, to guarantee

$$P\left(max_{j=1,\ldots,M} |\widehat{\varphi}_j - \varphi_j| \leq \delta\right) \geq 1 - \alpha, \tag{3}$$

that is,

$$\prod_{j=1}^{M} [\widehat{\varphi}_j \pm \delta] \tag{4}$$

defines a simultaneous region of confidence $1 - \alpha$ for $\boldsymbol{\varphi}$.

3 How many customers to interview?

We analytically solve the problem to determine the customer sample size proposed in (3) of Section 2, proving Theorem 2 under these assumptions:

C1 Fixed M DMUs
C2 One known input $\{X_j\}_{j=1,\ldots,M}$ and one unknown opinion index (output) $\{Y_j\}_{j=1,\ldots,M}$
C3 CCR-O model

Lemma 1 is the result used to prove Theorem 2, establishing the relation between sample size and accuracy, in order to simultaneously estimate the vector of DEA efficiencies.

Lemma 1 *Under assumptions C1, C2 and C3, for any* $0 < p < 1$*, we consider the sets of* Ω*:*

$$A_j = \left\{ u = (u_1 \times \cdots \times u_M) \in \Omega \ / \ \left| \widehat{Y}_j(u_j) - Y_j \right| \leq pY_j \right\};\ j = 1, \ldots, M \tag{5}$$

$$B_j = \left\{ u \in \Omega \ / \ \left| \widehat{\varphi}_j(u) - \varphi_j \right| \leq \frac{2p}{1+p} \right\};\ j = 1, \ldots, M \tag{6}$$

then

$$\bigcap_{j=1}^{M} A_j \subset \bigcap_{j=1}^{M} B_j.$$

Theorem 2 *Under assumptions C1, C2 and C3, for any* $0 < \delta < 1$ *and any* $0 < \alpha < 1$, *for every* $j = 1, \ldots, M$, *let* n_j *be the sample size in the DMU*$_j$ *such that*

$$P\left(\left|\widehat{Y}_j - Y_j\right| \leq pY_j\right) \geq \sqrt[M]{1-\alpha} \tag{7}$$

with $p = \frac{\delta}{2-\delta} \in (0,\ 1)$. *Then*

$$\prod_{j=1}^{M} [\widehat{\varphi}_j \pm \delta] \tag{8}$$

defines a simultaneous region of confidence $1 - \alpha$ *for the population efficiency scores vector.*

Remark 3 gives the explicit formulas to obtain the sample size under the usual simple random sample without replacement in a finite population.

Remark 3 If the customer sampling in each DMU is a simple random sample without replacement and the output is a population mean then

$$Y_j = \frac{\sum_{k=1}^{N_j} u_{kj}}{N_j};\ j = 1, \ldots, M$$

where u_{kj} is the answer (opinion) of the customer k in the DMU$_j$ and N_j its population size; then the sample size n_j that it verifies

$$P\left(\left|\widehat{Y}_j - Y_j\right| \leq pY_j\right) \geq \alpha_1$$

is (Särndal, Swensson and Wretman, 2003)

$$n_j \geq \frac{n_{oj}}{\left(\frac{n_{oj}}{N_j} + 1\right)} \tag{9}$$

with $n_{oj} = \frac{\tau^2_{1-\left(\frac{1-\alpha_1}{2}\right)}}{\left(pY_j\right)^2}{\sigma_j}^2$ and $\tau_{1-\left(\frac{1-\alpha_1}{2}\right)} = \phi^{-1}\left(1 - \left(\frac{1-\alpha_1}{2}\right)\right)$, where ${\sigma_j}^2$ is the population variance and ϕ the normal standard distribution function.

4 Bootstrap efficiency simultaneous confidence region

We carried out a simulation to check the confidence of the simultaneous region (8) in the case of two known inputs, two outputs estimated with a simple random sample without replacement of customers of size (9) and BCC-O model (2). This confidence is approximately one, so the region (8) is very conservative. We propose, as an alternative,

an algorithm to construct a simultaneous confidence region for the population DEA efficiency scores vector, using Theorem 2 to determine the sample size, and bootstrap resampling of the samples of the customers' answers to the opinion items to estimate the population efficiency.

The algorithm is, considering M DMUs, each one using $m \geq 1$ known inputs $\mathbf{X}_j = (X_{1j}, \ldots, X_{mj})$ and $s \geq 2$ unknown outputs $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{sj})$:

i. Taking $0 < \delta < 1$ and $0 < \alpha < 1$, we calculate the customer sample size n_j in the DMU_j as
$$n_j = max\{n_{1j}, \ldots, n_{sj}\} \tag{10}$$
where n_{rj} is the sample size to estimate the rth output in the DMU_j such that
$$P\left(\left|\widehat{Y}_{rj} - Y_{rj}\right| \leq pY_{rj}\right) \geq \sqrt[M]{1-\alpha};\ r = 1, \ldots, s \tag{11}$$
with $p = \frac{\delta}{2-\delta} \in (0,\ 1)$.

ii. In the DMU_j, $j = 1, \ldots, M$, we take the simple random sample without replacement of customers $\mathbf{u}_{kj} = (u_{k1j}, \ldots, u_{ksj})$ of size n_j, $k = 1, \ldots, n_j$, and we estimate the outputs $\widehat{\mathbf{y}}_j = (\widehat{y}_{1j}, \ldots, \widehat{y}_{sj})$ with the sample mean:
$$\widehat{y}_{rj} = \frac{\sum_{k=1}^{n_j} u_{krj}}{n_j};\ r = 1, \ldots, s,\ j = 1, \ldots, M. \tag{12}$$

iii. We take a bootstrap sample with replacement $\widehat{\mathbf{u}}^*_{kj}$ from $\mathbf{u}_{kj}$ of size n_j, $j = 1, \ldots, M$, with which we obtain the bootstrap version of the s output estimations $\widehat{\mathbf{y}}^*_j = \left(\widehat{y}^*_{1j}, \ldots, \widehat{y}^*_{sj}\right); j = 1, \ldots, M$.
With the data $\left\{\left(X_{1j}, \ldots, , X_{mj}, \widehat{y}^*_{1j}, \ldots, \widehat{y}^*_{sj}\right)\right\}_{j=1,\ldots,M}$, using a Table 1 model, we obtain the bootstrap version of the estimated efficiency scores, $\left\{\widehat{\omega}^*_j\right\}_{j=1,\ldots,M}$.

iv. We repeat step iii B times and the B bootstrap versions of the estimated efficiency scores for the DMU_j, $j = 1, \ldots, M$, should be $\left\{\widehat{\omega}^{*(b)}_j\right\}_{b=1,\ldots,\ B}$.
For any $0 < \alpha' < 1$, let $1-\alpha'$ be the level of coverage intention; then the observed bootstrap simultaneous confidence region of the population efficiency vector is
$$RC^* = \prod_{j=1}^{M}\left(\widehat{\omega}^{*\left(\frac{\alpha'}{2n}\right)}_j, \widehat{\omega}^{*\left(1-\frac{\alpha'}{2n}\right)}_j\right) \tag{13}$$
where $\widehat{\omega}^{*(\alpha)}_j$ is the α-percentile of the B values $\left\{\widehat{\omega}^{*(b)}_j\right\}_{b=1,\ldots,\ B}$.

This algorithm is ad hoc. No theory is given to suggest if it permits an estimate of the confidence region with asymptotically correct coverages, only the simulation study check the estimate quality.

4.1 Simulation study

We generate a simulated population model, as in Tapia et al. (2018), using the health centre data of Cooper, Seiford, and Tone (2006) (Table 2): in the j-th health centre, $j = 1,\ldots,12$, a finite patient population $\mathbf{U}_j = \{\mathbf{u}_{1j},\ldots,\mathbf{u}_{N_j j}\}$, of size N_j, $j = 1,\ldots,12$, is generated where N_j are independent random variables with uniform distribution in $[10000, 50000]$, and

$$\mathbf{u}_{kj} = (u_{k1j}, u_{k2j}) \to N_2\left(\begin{pmatrix} z_{1j} \\ z_{2j} \end{pmatrix}, \begin{pmatrix} z_{1j}^2/4 & 0 \\ 0 & z_{2j}^2/4 \end{pmatrix}\right);\ k = 1,\ldots,N_j;\ j = 1,\ldots,12$$

and (z_{1j}, z_{2j}) are the original value outputs of the jth health centre, columns 4 and 5 of Table 2.

Table 2: *Number of doctors, nurses, outpatients and inpatients in 12 health centres.*

DMU	Doctor X1	Nurse X2	Outpatient Z1	Inpatient Z2	CCR-O efficiency score	BCC-O efficiency score
1	2.0	15.1	10	9	1	1
2	1.9	13.1	15	5	1	1
3	2.5	16	16	5.5	0.883	0.925
4	2.7	16.8	18	7.2	1	1
5	2.2	15.8	9.4	6.6	0.763	0.767
6	5.5	25.5	23	9	0.835	0.955
7	3.3	23.5	22	8.8	0.902	1
8	3.1	20.6	15.2	8	0.796	0.826
9	3	24.4	19	10	0.960	0.990
10	5	26.8	25	10	0.871	1
11	5.3	30.6	26	14.7	0.955	1
12	3.8	28.4	25	12	0.958	1

Source: Table 1.5 Cooper et al. (2006)

Table 3 shows the simulated population model: the patient population size for each health centre (column 2), the known inputs (columns 3 and 4) and the simulated values of the two outputs (columns 5 and 6) obtained with the population means

$$(Y_{1j}, Y_{2j}) = \left(\frac{\sum_{k=1}^{N_j} u_{k1j}}{N_j}, \frac{\sum_{k=1}^{N_j} u_{k2j}}{N_j}\right);\ \ j = 1,\ldots,M \tag{14}$$

where u_{krj} is the answer of the kth patient of the jth health centre to the rth opinion question. Columns 6 and 7 show the population DEA efficiency CCR-O and BCC-O, respectively.

Table 3: *Simulated population model.*

DMU	Population size N_j	Doctor X1	Nurse X2	Y1	Y2	Population efficiency score φ_j	
						CCR-O	BCC-O
1	43341	2.0	15.1	9.98	9.01	1	1
2	24438	1.9	13.1	14.98	5.01	1	1
3	45606	2.5	16	15.99	5.50	0.883	0.926
4	12578	2.7	16.8	17.96	7.18	1	1
5	19314	2.2	15.8	9.40	6.58	0.763	0.766
6	21782	5.5	25.5	22.96	8.97	0.835	0.957
7	19024	3.3	23.5	21.99	8.77	0.901	0.998
8	36271	3.1	20.6	15.17	8.01	0.797	0.826
9	30691	3	24.4	19.02	10.04	0.963	0.991
10	28385	5	26.8	24.89	10.01	0.871	1
11	28005	5.3	30.6	26.11	14.69	0.958	1
12	49077	3.8	28.4	25.09	11.98	0.960	1

Supposing a simple random sample without replacement of patients in each health centre, having fixed an efficiency estimation error $\delta = 0.1$ and a probability $1-\alpha = 0.95$, Table 4 shows the customer sample size n_j, $j = 1,\ldots,12$, calculated using (10) and (11).

Table 4: *Patient sample size obtained for each DMU, fixed* $\delta = 0.1$ *and* $\alpha = 0.05$.

DMU	n_j
1	985
2	674
3	779
4	1078
5	900
6	1415
7	795
8	814
9	765
10	617
11	1194
12	682

With this sample size: we first repeat the bootstrapping methodology steps ii. - iv. 1000 times, obtaining 1000 observed bootstrap efficiency simultaneous confidence regions, $\left\{RC^{*(k)}\right\}_{k=1,\ldots,1000}$, as in (13). The confidence of the bootstrap efficiency simultaneous confidence region is approximated through

$$C^* = \frac{1}{1000}\sum_{k=1}^{1000} I_{(\varphi_1,\ldots,\varphi_{12})\in RC^{*(k)}}. \tag{15}$$

Having fixed a percentile bootstrap confidence $1-\alpha' = 0.9$, Table 5 shows the approximated confidence (15) of the bootstrap confidence region to simultaneously estimate the population efficiency score vector. This result confirms that, by determining the sample size in each DMU using (10), we get the fixed accuracy with the percentile bootstrap efficiency simultaneous confidence region.

Table 5: *Confidence approximation of the bootstrap efficiency simultaneous confidence region taking* $1-\alpha' = 0.9$ *and fixed* $\delta = 0.1$ *and* $\alpha = 0.05$*. CCR and BCC model with output orientation.*

	CCR-O	**BCC-O**
C^*	0.975	0.983

In order to justify the basic percentile method to obtain bootstrap confidence intervals, we analyse the bias of the bootstrap process. Using the 1000 data $\left\{\widehat{\omega}_j - \widehat{\omega}_j^{*(b)}\right\}_{b=1,\ldots,1000}$ for each DMU_j, we represent graphically the nonparametric density estimates with kernel $N(0,1)$ and smoothing parameter selected with rule-of-thumb (Silverman (1986)). The smooth density estimates obtained are approximately symmetrical with respect to 0, therefore the bias is negligible.

5 A Case Study

This section provides an empirical DEA efficiency analysis of libraries using real data input. Table 7 corresponds to the database from 15 libraries used in Tapia et al. (2018). The data input are the number of book loans, (X_1), the library's seating capacity and the number of computers for users, (X_2), and the data are scaled 0-10. Column 2 shows the distribution of the user population in each of the 15 libraries. In each library, we use only the output given by the users' mean, monthly time use of the library, in hours, measured for each user, on a scale of 0-10. In order to determine the user sample size n_j in the jth library, $j = 1,\ldots,15$, using Remark 3 in each library, we take a previous user simple random sample without replacement of 0.1% of the user population size and we estimate the output and the population variance, shown in columns 2, 3 and 4 of Table 7, respectively. Having fixed the efficiency estimation error $\delta = 0.1$ and a probability $1-\alpha = 0.9$, we determine the sample size n_j; $j = 1,\ldots,15$ with (10) and (11). We then take the user sample in each library and estimate the mean monthly time of permanence in the library $\{\widehat{y}_j\}_{j=1,\ldots,15}$, column 6 of Table 7.

Table 6: *Results of the previous simple random sample without replacement: User sample size, Estimation of monthly time use mean and of population variance.*

DMU	Sample size $n_j^{(0)}$	Estimation of the monthly time use library mean	Estimation of the population variance
1	89	6.61	14.34
2	79	5.78	13.76
3	64	3.87	12.93
4	59	5.06	13.26
5	57	6.12	14.48
6	51	2.88	8.67
7	42	5.41	17.08
8	37	4.31	13.79
9	35	6.06	12.79
10	32	4.10	10.24
11	24	4.57	15.16
12	21	5.27	16.97
13	19	3.09	13.10
14	17	4.37	14.55
15	13	7.81	10.63

Table 7: *Database for 15 libraries: User population size, book loans and user posts (inputs), user sample size to estimate the mean monthly library time use (output) with* $\delta = 0.1$ *and* $\alpha = 0.1$.

DMU	User population size N_j	Book loans X1	User posts X2	User sample size n_j	Estimated monthly time use library mean $\widehat{y}_j$
1	89300	7.04	7.82	855	6.03
2	78500	7.81	6.87	1065	5.33
3	64000	5.41	5.60	2190	3.26
4	59100	2.66	5.18	1326	5.33
5	56500	3.96	4.95	998	6.13
6	50700	3.28	4.44	2601	3.37
7	41600	4.36	3.64	1476	5.79
8	37000	6.29	3.24	1850	3.44
9	34600	5.82	3.03	891	5.14
10	32000	7.69	2.80	1523	3.48
11	23600	2.61	2.07	1761	5.28
12	21200	3.61	1.86	1492	5.87
13	18900	4.73	1.65	3028	5.33
14	17200	2.12	1.51	1791	3.40
15	12900	2.16	1.13	442	5.77

The estimated efficiency scores, $\{\widehat{\omega}_j\}_{j=1,\ldots,15}$ and the intervals whose product determines the bootstrap efficiency simultaneous confidence region (columns 2 and 3 in Table 8, respectively) are obtained using the data $\{(X_{1j},\, X_{2j},\, \widehat{y}_j)\}_{j=1,\ldots,15}$ and the BCC model with orientation output. The libraries $\{1, 5, 7, 14, 15\}$ can be considered efficient

because the corresponding intervals of the bootstrap efficiency simultaneous confidence region contains the value 1.

Table 8: *Estimated efficiency scores and intervals whose product determines the bootstrap efficiency simultaneous confidence region, taking* $1-\alpha' = 0.9$. *BCC model with output orientation.*

DMU	Estimated efficiency score	Intervals
1	0.983	[0.904, 1]
2	0.870	[0.797, 0.942]
3	0.531	[0.493, 0.567]
4	0.908	[0.841, 0.982]
5	1	[0.959, 1]
6	0.562	[0.524, 0.596]
7	0.962	[0.902, 1]
8	0.574	[0.529, 0.611]
9	0.862	[0.791, 0.925]
10	0.585	[0.539, 0.625]
11	0.900	[0.826, 0.981]
12	1	[0.913, 1]
13	0.914	[0.839, 0.959]
14	1	[1, 1]
15	1	[1, 1]

6 Conclusions

Over the last decade, the use of opinion-satisfaction surveys on customers of public services has been an essential tool in measuring the quality of the service given. Without a doubt, a public service will be more efficient when, with its resources, it is able to have the highest opinion-satisfaction of its customers. The questionnaire is a common tool to find out customer opinion-satisfaction with the service received. The mean of the opinion-satisfaction answers of the sample of customers are indices, indicators of the service quality, that can be considered as output data. If we add the deterministic information of the resources of the public service-producing unit as input data, we will have the necessary input and output data to calculate the DEA efficiency in the set of services.

We focus on this DEA efficiency problem as a statistical one, considering an unknown population efficiency vector that would be obtained if we had the opinion of the entire service customer population (census). We estimate this parametric vector with a confidence region using the outputs estimated with the opinion of the user sample, the known inputs and the classical DEA models (LP models subject to deterministic constraints). To our knowledge, this statistical view of the DEA is totally novel and the use of a simultaneous confidence region is a statistical concept that has not been used

in DEA efficiency analysis in the form that we propose. From a practical point of view, the application in library datasets shows the usefulness of the bootstrap region confidence efficiency methodology. This region, based on the product of confidence intervals in each library, detects whether the library is "efficient in any case" because the lower bound is equal to unity, or if the library is "efficient" because the upper bound is equal to unity and "inefficient" because the upper bound is less than unity. More public service examples where it may be interesting to apply the results of this paper are: leisure centres (where there are attractions for which it is necessary to maximize the demand, which can be evaluated by carrying out customer surveys), marketing or electoral polls (where the effect of the advertising or electoral campaign is evaluated with a survey), hospitals, airports, banks, universities, supermarkets, government services or schools (where existing resources can explain customer opinion-satisfaction).

In this paper, we obtain two types of confidence region for the population efficiency scores vector. Theorem 2 allows us to define the first simultaneous region, finding a relation between the accuracy of the simultaneous confidence region and the customer sample size needed to guarantee an output estimation error in each public service-producing unit. By simulation, we check that this confidence region is very conservative. As an alternative, we propose to determine the sample size of customers necessary using Theorem 2 and Remark 3 and to obtain an efficiency confidence region based on the basic percentile bootstrap method. The simulation shows that we are able to reach the confidence of the bootstrap efficiency simultaneous confidence region close to the desired level.

Other possible extensions currently under investigation by the authors also include considering stochastic inputs estimated with a provider sample or using other bootstrap methods, such as the adjusted percentile method or the ABC method, and comparing them with the method used in this paper.

7 Appendix section

7.1 Proof of Lemma 1

Let $u = (u_1 \times \cdots \times u_M) \in \bigcap_{j=1}^{M} A_j$.

Let us consider the DMU$_r$. Then, $\varphi_r = 1$ or $\varphi_r < 1$:

- If $\varphi_r = 1$, it is because $\frac{Y_r}{X_r} = max_j \frac{Y_j}{X_j}$.

  The most unfavourable situation, where B_r is verified, is that the output of the DMU$_r$ is as small as possible and the rest of the DMUs are as big as possible, that is to say

$$Y_j^* = (1-p)Y_j I_{(j=r)} + (1+p)Y_j I_{(j\neq r)}.$$

Let $\varphi_j^*,\ j=1,\ldots,M$ be the efficiencies of the DMUs obtained with the data $\left\{(X_j,Y_j^*)\right\}_{j=1,\ldots,M}$, therefore

$$\varphi_r^*=\begin{cases} 1 & \text{if } \frac{(1-p)Y_r}{X_r}\geq max_{j\neq r}\ \frac{(1+p)Y_j}{X_j} \quad (a)\\ \frac{(1-p)Y_r}{(1+p)\frac{Y_k}{X_k}X_r} & \text{if } \ (1+p)\frac{Y_k}{X_k}=max_{j\neq r}\ \frac{(1+p)Y_j}{X_j}>\frac{(1-p)Y_r}{X_r} \quad (b)\end{cases}$$

and

$$|\varphi_r-\widehat{\varphi}_r(u)|=1-\widehat{\varphi}_r(u)\leq 1-\varphi_r^*$$

where the first equality is obtained because $\varphi_r=1$ and the second inequality is verified because, as the efficiency of a DMU decreases when the output of this unit decreases, while the outputs of the rest of the DMUs also increase, $\widehat{\varphi}_r(u)\geq\varphi_r^*$, then

$$|\varphi_r-\widehat{\varphi}_r(u)|\leq\begin{cases}0 & \text{if } (a)\\ 1-\frac{(1-p)}{(1+p)}=\frac{2p}{1+p} & \text{if } (b)\end{cases}$$

and it is verified that $u\in B_r\ \forall r\ /\ \varphi_r=1$.

- If $\varphi_r<1$

Let $k\neq r\ /\ \frac{Y_k}{X_k}=max_j\frac{Y_j}{X_j}$ and $\varphi_r=\frac{\frac{Y_r}{X_r}}{\frac{Y_k}{X_k}}<1$.

There are two more favourable situations for B_r to be verified

Case (I) $Y_j^*=(1-p)Y_jI_{(j=r)}+(1+p)Y_jI_{(j\neq r)}$.

Case (II) $Y_j^{**}=(1+p)Y_jI_{(j=r)}+(1-p)Y_jI_{(j\neq r)}$.

Case (I): Let $\varphi_j^*,\ j=1,\ldots,M,$ be the DMU efficiencies obtained with the data $\left\{(X_j,Y_j^*)\right\}_{j=1,\ldots,M}$, then

$$\varphi_r^*=\frac{(1-p)\frac{Y_r}{X_r}}{(1+p)\frac{Y_k}{X_k}}=\frac{(1-p)}{(1+p)}\varphi_r<\varphi_r.$$

As the efficiency of a DMU decreases when the output of this unit decreases, while the outputs of the rest of the DMUs also increase,

$$|\varphi_r-\widehat{\varphi}_r(u)|\leq|\varphi_r-\varphi_r^*|$$

and therefore

$$|\varphi_r-\widehat{\varphi}_r(u)|\leq\varphi_r-\varphi_r^*=\varphi_r\left(\frac{2p}{1+p}\right)<\frac{2p}{1+p}.$$

Case (II): Let $\varphi_j^{**},\ j=1,\ldots,M,$ be the DMU efficiencies obtained with the data $\left\{(X_j,Y_j^{**})\right\}_{j=1,\ldots,M}$, then

- If $max\left(\frac{(1+p)Y_r}{X_r}, max_{j\neq r}\ \frac{(1-p)Y_j}{X_j}\right)=\frac{(1+p)Y_r}{X_r}$ then $\varphi_r^{**}=1$ and $\varphi_r\geq\frac{(1-p)}{(1+p)}$ and therefore

$$|\varphi_r-\widehat{\varphi}_r(u)|\leq|\varphi_r-\varphi_r^{**}|=1-\varphi_r\leq\frac{2p}{1+p}.$$

- If $max\left(\frac{(1+p)Y_r}{X_r}, max_{j\neq r}\ \frac{(1-p)Y_j}{X_j}\right)=\frac{(1-p)Y_k}{X_k}$ for $k\neq r$ then $\varphi_r<\frac{(1-p)}{(1+p)}$ and $\varphi_r^{**}=\frac{(1+p)\frac{Y_r}{Y_r}}{(1-p)\frac{Y_k}{X_k}}=\frac{(1+p)}{(1-p)}\varphi_r>\varphi_r$ and therefore

$$|\varphi_r-\widehat{\varphi}_r(u)|\leq\varphi_r^{**}-\varphi_r\leq\frac{2p}{1-p}\varphi_r<\frac{2p}{1+p}.$$

In consequence $u\in B_r\ \forall r\ /\ \varphi_r<1$.

7.2 Proof of Theorem 2

It is enough to prove that

$$P\left(max_{j=1,\ldots,M}\left|\widehat{\varphi}_j-\varphi_j\right|\leq\delta\right)\geq 1-\alpha. \tag{16}$$

Using the notation of (5) and (6), by Lemma 1, we know that

$$\bigcap_{j=1}^{M}A_j\subset\bigcap_{j=1}^{n}B_j$$

and, as the events $\{A_j\}_{j=1,\ldots,M}$ are independent, then

$$P\left(\bigcap_{j=1}^{M}B_j\right)\geq P\left(\bigcap_{j=1}^{M}A_j\right)=\prod_{j=1}^{M}P(A_j)\geq\left(\sqrt[M]{1-\alpha}\right)^M=1-\alpha.$$

Acknowledgements

The authors thank the Editor and two anonymous referees for their constructive comments and suggestions that have helped to improve the quality of the paper.

References

Alexander, W.R.J., Haug, A.A. and Jaforullah, M. (2010). A two-stage double-bootstrap data envelopment analysis of efficiency differences of New Zealand secondary schools. *Journal of Productivity Analysis*, 34, 99–110.

Assaf, A. (2010). Bootstrapped scale efficiency measures of UK airports. *Journal of Air Transport Management*, 16, 42–44.

Assaf, A.G. and Agbola, F.W. (2011). Modelling the performance of Australian hotels: A DEA double bootstrap approach. *Tourism Economics*, 17, 73–89.

Banker, R.D. (1993). Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. *Management Science*, 39, 1265–1273.

Barra, C. and Zotti, R. (2016). Measuring efficiency in higher education: An empirical study using a bootstrapped data envelopment analysis. *International Advances in Economic Research*, 22, 11–33.

Bayraktar, E., Tatoglu, E., Turkyilmaz, A., Delen, D. and Zaim, S. (2012). Measuring the efficiency of customer satisfaction and loyalty for mobile phone brands with DEA. *Expert Systems with Applications*, 39, 99–106.

Benito, B., Solana, J. and Moreno, M.R. (2014). Explaining efficiency in municipal services providers. *Journal of Productivity Analysis*, 42, 225–239.

Casu, B. and Molyneux, P. (2003). A comparative study of efficiency in European banking. *Applied Economics*, 35, 1865–1876.

Ceyhan, M.E. and Benneyan, J.C. (2014). Handling estimated proportions in public sector data envelopment analysis. *Annals of Operations Research*, 221, 107–132.

Charles, V. and Kumar, M. (2014). Satisficing data envelopment analysis: An application to servqual efficiency. *Measurement*, 51, 71–80.

Charnes, A. and Cooper, W.W. (1959). Chance-constrained programming. *Management Science*, 6, 73–79.

Charnes, A. and Cooper, W.W. (1963). Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, 11, 18–39.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2, 429–444.

Chowdhury, H. and Zelenyuk, V. (2016). Performance of hospital services in Ontario: DEA with truncated regression approach. *Omega*, 63, 111–122.

Cooper, W., Huang, Z. and Li, S.X. (1996). Satisficing DEA models under chance constraints. *Annals of Operations Research*, 66, 279–295.

Cooper, W.W., Deng, H., Huang, Z., Li, S.X. et al. (2002). Chance constrained programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis. *Journal of the Operational Research Society*, 53, 1347–1356.

Cooper, W.W., Seiford, L.M. and Tone, K. (2006). Introduction to data envelopment analysis and its uses: with DEA-solver software and references. *Springer Science & Business Media. New York, NY 10013, USA*.

Cooper, W.W., Seiford, L.M. and Zhu, J. (2011). Handbook on data envelopment analysis, volume 164. *Springer Science & Business Media. New York, NY 10013, USA*.

Essid, H., Ouellette, P. and Vigeant, S. (2014). Productivity, efficiency and technical change of Tunisian schools: a bootstrapped Malmquist approach with quasi-fixed inputs. *Omega*, 42, 88–97.

Førsund, F.R. (2017). Measuring effectiveness of production in the public sector. *Omega*, https://doi.org/10.1016/j.omega.2016.12.007.

Gil Ropero, A., Turias Dominguez, I. and Cerbán Jiménez, M.M. (2019). Bootstrapped operating efficiency in container ports: a case study in Spain and Portugal. *Industrial Management & Data Systems*, 119, 924–948.

Huang, Z. and Li, S.X. (2001). Stochastic DEA models with different types of input-output disturbances. *Journal of Productivity Analysis*, 15, 95–113.

Khodabakhshi, M. (2010). An output oriented super-efficiency measure in stochastic data envelopment analysis: Considering iranian electricity distribution companies. *Computers & Industrial Engineering*, 58, 663–671.

Khodabakhshi, M. and Asgharian, M. (2009). An input relaxation measure of efficiency in stochastic data envelopment analysis. *Applied Mathematical Modelling*, 33, 2010–2023.

Kneip, A., Park, B.U. and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14, 783–793.

Kneip, A., Simar, L. and Wilson, P.W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24, 1663–1697.

Kneip, A., Simar, L. and Wilson, P.W. (2011). A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators. *Computational Economics*, 38, 483–515.

Land, K.C., Lovell, C. and Thore, S. (1993). Chance-constrained data envelopment analysis. *Managerial and Decision Economics*, 14, 541–554.

Lee, H. and Kim, C. (2014). Benchmarking of service quality with data envelopment analysis. *Expert Systems with Applications*, 41, 3761–3768.

Liu, S.T. and Chuang, M. (2009). Fuzzy efficiency measures in fuzzy DEA/AR with application to university libraries. *Expert Systems with Applications*, 36, 1105–1113.

Mayston, D.J. (2015). Analysing the effectiveness of public service producers with endogenous resourcing. *Journal of Productivity Analysis*, 44, 115-126.

Mayston, D.J. (2017). Data envelopment analysis, endogeneity and the quality frontier for public services. *Annals of Operations Research*, 250, 185–203.

Olesen, O.B. and Petersen, N. (1995). Chance constrained efficiency evaluation. *Management Science*, 41, 442–457.

Parasuraman, A., Berry, L.L. and Zeithaml, V.A. (1993). More on improving service quality measurement. Journal of Retailing, 69, 140–147.

Salo, A. and Punkka, A. (2011). Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science*, 57, 200–214.

Santín, D. and Sicilia, G. (2017). Dealing with endogeneity in data envelopment analysis applications. *Expert Systems with Applications*, 68, 173–184.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). Model assisted survey sampling. *Springer Science & Business Media. New York, NY 10010, USA*.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. *Chapman and Hall. London; New York*.

Simar, L. and Wilson, P.W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, 44, 49–61.

Simar, L. and Wilson, P.W. (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, 27, 779–802.

Simar, L. and Wilson, P.W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics, 136, 31–64.

Simar, L. and Wilson, P.W. (2011). Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, 36, 33–53.

Simar, L. and Wilson, P.W. (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends in Econometrics*, 5, 183–337.

Simar, L. and Wilson, P.W. (2015). Statistical approaches for non-parametric frontier models: a guided tour. *International Statistical Review*, 83, 77–110.

Simar, L. and Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews*, 25, 497–522.

Smith, P. and Mayston, D. (1987). Measuring efficiency in the public sector. *Omega*, 15, 181–189.

Tapia, J.A., Salvador, B. and RodrÃguez, J.M. (2018). Data envelopment analysis in satisfaction survey research: sample size problem. *Journal of the Operational Research Society*, 69, 1096-1104.

Tavana, M., Shiraz, R.K. and Hatami-Marbini, A. (2014). A new chance-constrained DEA model with birandom input and output data. *Journal of the Operational Research Society*, 65, 1824–1839.

Tsekouras, K., Papathanassopoulos, F., Kounetas, K. and Pappous, G. (2010). Does the adoption of new technology boost productive efficiency in the public sector? the case of ICUs system. *International Journal of Production Economics*, 128, 427–433.

Witte, K.D. and Geys, B. (2013). Citizen coproduction and efficient public good provision: Theory and evidence from local public libraries. *European Journal of Operational Research*, 224, 592–602.

Wu, D. and Olson, D.L. (2008). A comparison of stochastic dominance and stochastic DEA for vendor evaluationand. *International Journal of Production Research*, 46, 2313–2327.

Wu, D.D. (2010). A systematic stochastic efficiency analysis model and application to international supplier performance evaluation. *Expert Systems with Applications*, 37, 6257–6264.

Wu, D.D. and Lee, C.-G. (2010). Stochastic DEA with ordinal data applied to a multi-attribute pricing problem. *European Journal of Operational Research*, 207, 1679–1688.

SORT 43 (2) July-December 2019, 355-384 DOI: 10.2436/20.8080.02.92

Forecasting with two generalized integer-valued autoregressive processes of order one in the mutual random environment

Predrag M. Popović[1], Petra N. Laketa[2] and Aleksandar S. Nastić[3]

Abstract

In this article, we consider two univariate random environment integer-valued autoregressive processes driven by the same hidden process. A model of this kind is capable of describing two correlated non-stationary counting time series using its marginal variable parameter values. The properties of the model are presented. Some parameter estimators are described and implemented on the simulated time series. The introduction of this bivariate integer-valued autoregressive model with a random environment is justified at the end of the paper, where its real-life data-fitting performance was checked and compared to some other appropriate models. The forecasting properties of the model are tested on a few data sets, and forecasting errors are discussed through the residual analysis of the components that comprise the model.

MSC: 62M10.

Keywords: INAR, negative binomial thinning, random states, time series of count, non-stationary process.

1 Introduction

After many scientific proposals of possible models of counting processes in the last decades of the 20th century, so far the best results have been obtained by the thinning-based integer-valued autoregressive models of order one (INAR(1)) introduced almost simultaneously by McKenzie (1985) and Al-Osh and Alzaid (1987). For the first time, they used an idea of defining the deterministic part of the counting process in a certain moment, designated by X_n, for the given $X_{n-1} = x_{n-1}$, using the random sum of x_{n-1} independent and identically distributed (i.i.d.) Bernoulli variables. Precisely,

$$X_n = \sum_{i=1}^{x_{n-1}} v_i + \varepsilon_n,$$

[1] popovicpredrag@yahoo.com, University of Niš, Faculty of Civil Engineering and Architecture, Serbia.

[2] petra.laketa@pmf.edu.rs, University of Niš, Faculty of Sciences and Mathematics, Serbia.

[3] anastic78@gmail.com, University of Niš, Faculty of Sciences and Mathematics, Serbia.

Received: April 2019

Accepted: November 2019

where $\{v_i\}$ is a counting sequence of i.i.d. Bernoulli random variables and ε_n are the innovation process. It is assumed that ε_n and X_{n-k} are independent for all $k > 0$. The INAR model of order one (INAR(1)) can be expressed such:

$$X_n = \alpha \circ x_{n-1} + \varepsilon_n.$$

where the operator $\alpha \circ X_{n-1} | X_{n-1} = x_{n-1}$ is equal to $\alpha \circ x_{t-1} = \sum_{i=1}^{x_{n-1}} v_i$, and it is called the binomial thinning or binomial subsampling operator. The first addend of the model above can be interpreted as a survival process. Therefore, these kinds of processes were ideal for modeling counts generated by limited surviving entities. During the adaptation of this INAR model to many different counting time series, many modifications and generalizations were done. Some researchers were focused on the thinning operator, and their innovative results can be found in Aly and Bouzar (1994), Latour (1998), Zheng, Basawa and Datta (2006, 2007), Zhu and Joe (2006), Ristić, Bakouch and Nastić (2009) and Zhu and Joe (2010). Even though Al-Osh and Aly (1992) as well as Alzaid and Al-Osh (1993) focused on the marginal distribution of the process, other authors preferred to concentrate on the distribution of the innovations, like Jazi, Jones and Lai (2012a, 2012b), Fernández-Fontelo, Fontdecaba and Puig (2017). Also, a certain modification of the innovation process was studied recently in Qi, Li and Zhu (2019).

Later, more attention was paid to the correlation characteristics of the observed processes, i.e. the additional assumptions about the dependence in the counting sequence were introduced. Initial results on the INAR models based on the thinning operator defined using dependent counting sequences were given by Ristić, Nastić and Miletić Ilić (2013). Also, the possibility of serially dependent innovations of the INAR model was studied and well-presented in Weiß (2015). In addition, Weiß, Homburg and Puig (2019) considered testing for zero inflation and overdispersion in INAR(1) models.

Parameter-driven models provided another approach to modeling counting processes. A good insight into these models can be found in Fokianos (2011) and some recent progress is presented in Chakraborty and Bhati (2016) (see also Chakraborty and Bhati, 2017) and Rydén (2017).

In addition to all the preceding models and given aspects of counting processes construction, there were many other approaches which resulted, especially in the last decade, with significant number of papers covering this area of time series research. Although, a great majority of them referred to the problems of modeling stationary processes, in the past years, some authors have been working to accommodate potential patterns of trend and seasonality in INAR models. Significant results in this area can be found in Moriña et al. (2011), Fernández-Fontelo et al. (2017).

Since non-stationarity may be noted in many real life situations, inspired by the work of Tang and Wang (2014), and in order to provide more efficient INAR modeling, a new random environment INAR process of order one (INAR(1) with variable marginal distribution was introduced in Nastić, Laketa and Ristić (2016). This model was non-stationary, which made it more applicable to counting processes in practice. The same

authors also presented a higher-order r-states random environment non-stationary INAR model, which can be found in Nastić, Laketa and Ristić (2019) and Laketa, Nastić and Ristić (2018). Fernández-Fontelo et al. (2016) gave an under-reported data analysis with INAR-hidden Markov chains. However, in the matter of modelling two correlated simultaneous integer-valued series, significant results were achieved by introducing bivariate INAR models which can be found in Pedeli and Karlis (2011), Ristić et al. (2012) and Popović, Ristić and Nastić (2016). The first model is based on the binomial thinning operator, and the dependence between time series was introduced through the innovation processes. The second model is based on the negative binomial thinning operator, considering geometric marginal distribution with the same mean parameters. The last model also has a geometric marginal distribution but assuming different mean parameters. Besides, while in the first model, the dependence between the series is considered in the innovation process, in the last two models, this dependence is considered in the survival components, i.e. the components defined through the thinning operator.

In this article, we focus on the bivariate random environment INAR model which is composed of the two univariate models discussed in Laketa et al. (2018). The two univariate series follow the same hidden process which determines the states of the observed processes. Thus, simultaneously with the observed process we have a Markov process $\{Z_n\}$, with a finite state space $E_r = \{1, 2, \ldots, r\}$, called the random state process. Its realized values z_n define marginal distribution parameter values. So, since each value from E_r corresponds to one state of the process environment, then the marginal distribution is directly dependent on the possible random states of the observed process environment. This can be found in nature every time we consider two random variables in the same circumstances. These variables do not have to be correlated directly, but only through their distributions which depend on the same conditions, i.e. random states. Also, considering such a bivariate model, we present its forecasting properties by conducting the residual analysis of its univariate components.

Like all random environment INAR models, the model proposed here is good for the data which are non-stationary (to be precise, they are part by part stationary), where we can suppose that the conditions in which they are measured can change and affect the measured values. So, this model is better than the other bivariate models for such data.

In the second section of this article, we give a short review of random environment INAR models. Then, in the following section, we introduce the corresponding bivariate model based on the realizations of the random environment process. Section 4 is mainly devoted to moment-based estimators. Also, a brief construction of the likelihood-based estimator is given. Section 5 deals with the residual analysis of the model. The quality of defined estimators is confirmed using simulated series of different sizes, presented in Section 6. The next section contains some real-life examples of the application of the introduced model to certain counting processes, where the model performance is compared to some other competitive bivariate INAR models. Also, the errors produced by one-step-ahead forecasting are analysed. Finally, all the proofs of the theorems are given in the Appendix.

2 A short review of random environment INAR models

The first random environment integer-valued autoregressive model was introduced in Nastić et al. (2016), and that is the random environment INAR(1) model. It is based on the random environment process, which represents the conditions of the environment in which the counting process is observed. Also, the corresponding process $\{Z_n\}$ is said to be an r-state random environment process if it is a Markov chain of order one and takes values from the set $E_r = \{1,2,\ldots,r\}$. The main assumption of the observed process is that the environment conditions have an effect on its marginal distribution. Thus, the r-state random environment INAR(1) process with the determined geometric marginal distribution, based on the negative binomial thinning operator (RrNGINAR(1)), is given by the equation

$$X_n(z_n) = \alpha * X_{n-1}(z_{n-1}) + \varepsilon_n(z_{n-1}, z_n),\ n \in \mathbb{N}, \tag{1}$$

where $\{z_n\}$ is a realization of the process $\{Z_n\}$. The notation $X_n(z_n)$ is used to emphasize the fact that the distribution of X_n depends on z_n. The value z_n determines the value μ_{z_n} from the supposed set of marginal parameter values $\{\mu_1, \mu_1, \ldots, \mu_r\}$ where, $X_n(z_n)$ has the geometric distribution with the expectation μ_{z_n}, since we supposed that its probability mass function (pmf) is defined as

$$P(X_n(z_n) = x) = \frac{\mu_{z_n}^x}{(1+\mu_{z_n})^{x+1}},\ x \in \mathbb{N}_0.$$

Here we gave an explanation on how the observable component X_n of the process depends on its latent component z_n. In addition, the denotation $\alpha*$ stands for the negative binomial thinning operator, which is defined by

$$\alpha * X = \sum_{i=1}^{X} U_i,$$

for an integer-valued random variable X, where $\alpha \in (0,1)$ and $\{U_i\}$, $i \in \mathbb{N}$, is a sequence of i.i.d. random variables with pmf given by

$$P(U_i = u) = \frac{\alpha^u}{(1+\alpha)^{u+1}},\ u \in \mathbb{N}_0.$$

In Laketa et al. (2018), this (RrNGINAR(1)) model is generalized, assuming that the realized random environment sequence $\{z_n\}$ determines not only the marginal distribution of the model, but also the order of the process and the thinning parameter value. In order to accurately present the models from Laketa et al. (2018), the following sets should be previously introduced: the set $\mathscr{M} = \{\mu_1, \mu_2, \ldots, \mu_r\}$ which consists of the possible mean values of the process in the corresponding states, the set $\mathscr{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ containing possible values of the thinning parameters corresponding to different states, and the set $\mathscr{P} = \{p_1, p_2, \ldots, p_r\}$ considering the order of the process. For example, when $z_n = i$,

the RrNGINAR(1) model is in its i-th state, and this means that the counting process is observed in the i-th environment state. Additionally, the model parameters in the i-th state are μ_i, α_i and p_i. In fact, in Laketa et al. (2018) two different RrNGINAR are introduced: RrNGINAR$_{max}(\mathscr{M},\mathscr{A},\mathscr{P})$ and RrNGINAR$_1(\mathscr{M},\mathscr{A},\mathscr{P})$. The set $\mathscr{P}$ contains actually the maximal orders for all states. The difference between these models (the one indexed by max, and other by 1) relies on the way of reaching these maximal orders. Let explain now this in more details, starting from the general form of these two models

$$X_n(z_n)=\begin{cases}\alpha_{z_n}*X_{n-1}(z_{n-1})+\varepsilon_n(z_{n-1},z_n), & \text{w.p. } \phi_{1,P_n}^{(z_n)},\\ \alpha_{z_n}*X_{n-2}(z_{n-2})+\varepsilon_n(z_{n-2},z_n), & \text{w.p. } \phi_{2,P_n}^{(z_n)},\\ \vdots & \vdots\\ \alpha_{z_n}*X_{n-P_n}(z_{n-P_n})+\varepsilon_n(z_{n-P_n},z_n), & \text{w.p. } \phi_{P_n,P_n}^{(z_n)},\end{cases}, \tag{2}$$

where $X_n(z_n)$ has geometric distribution with expectation μ_{z_n}. Since the distribution of the residuals would be complicated to obtain when $P_n=p_{z_n}$, P_n should be defined differently. Thus, for the RrNGINAR$_{max}(\mathscr{M},\mathscr{A},\mathscr{P})$ model, named INAR process with r-states, distribution parameters set $\mathscr{M}$, thinning parameters set $\mathscr{A}$ and maximal order set $\mathscr{P}$, it holds that

$$P_n=\min\{p_{z_n},p_n^*\},$$

$$p_n^*=\max\{i\in\{1,2,\ldots,n\}: z_{n-1}=z_{n-2}=\cdots=z_{n-i}\}.$$

From here, when the state change occurs, $z_n\neq z_{n-1}$, the process order becomes one, i.e. $P_n=1$, and afterwards it starts rising by 1 in every moment of the process, until it reaches its maximum value for that state, which equals p_{z_n}. Then it remains at maximum until the process state changes again. The alternative way, for the other type of the considered model (RrNGINAR$_1(\mathscr{M},\mathscr{A},\mathscr{P})$), is letting the value P_n equal 1 (instead of making it growing gradually), but still considering the value at the same moment at when the previously explained model RrNGINAR$_{max}(\mathscr{M},\mathscr{A},\mathscr{P})$ reaches the maximal order. Accordingly, for the RrNGINAR$_1(\mathscr{M},\mathscr{A},\mathscr{P})$ model, the only possible order values corresponding to the process state i are 1 and p_{z_n}

$$P_n=\begin{cases}p_{z_n}, & p_n^*\geq p_{z_n}\\ 1, & p_n^*<p_{z_n}\end{cases}$$

This model is named the random environment INAR process with r-states, distribution parameters set $\mathscr{M}$, thinning parameters set $\mathscr{A}$ and the order set $\mathscr{P}$.

If, as a special case, it holds that $p_1=p_2=\cdots=p_r=1$, then both models are the same and of order one. Also, the RrNGINAR(1) model is a special case of these two models, when $p_1=p_2=\cdots=p_r=1$ and $\alpha_1=\alpha_2=\cdots=\alpha_r$.

Explaining these two models from Laketa et al. (2018) further, let us now recall the Theorem 1 from that paper, which makes a point about residual distribution (see Appendix for details). Considering the models in Laketa et al. (2018), we should com-

prehend these random environment INAR models as an attempt of fitting counting processes in time-varying conditions, which directly affect to certain parameters of the observed process. As long as the conditions of the process environment do not change, the process itself has the same (and unchanged) value of its latent component z_n. However, when the environment eventually changes (e.g., social circumstances or economic factors), then the random environment INAR models introduced in this paper adapt to these changes. In fact, these models accommodate these environment changes by adequately modifying the values of specific parameters, including even the order of the process. Notice that these models are stationary while they are in the same state z_n, and their non-stationarity starts when changing this state. The latter is a consequence of changing the marginal distribution of the models, the thinning operator value, and the order of the process. When we observe the order, we notice that after the process state is changed from z_{n-1} to z_n, the process order is reduced to 1, which is necessary because of the definition of the model. The way it reaches its value of p_{z_n} depends on the model type (i.e., whether its type is "1" or "max", which was explained earlier). Finally, it should also be emphasized that we consider non-stationary processes, which are "part-by-part stationary", where each "part" corresponds to the period of a random process $\{Z_n\}$ remaining in the same state.

3 Considered models

Now, we focus on the model introduced in this article. Let $\{X_n(z_n)\}$ and $\{Y_n(z_n)\}$ be the RrNGINAR$_1(\mathscr{M}_1,\mathscr{A}_1,\mathscr{P}_1)$ and RrNGINAR$_1(\mathscr{M}_2,\mathscr{A}_2,\mathscr{P}_2)$ processes, respectively, where $\mathscr{M}_1=\{\mu_1,\mu_2,\ldots,\mu_r\}$, $\mathscr{M}_2=\{\nu_1,\nu_2,\ldots,\nu_r\}$, $\mathscr{A}_1=\{\alpha_1,\alpha_2,\ldots,\alpha_r\}$, $\mathscr{A}_2=\{\beta_1,\beta_2,\ldots,\beta_r\}$ and $\mathscr{P}_1=\mathscr{P}_2=\{1\}$. Then, they are defined with the following relations

$$X_n(z_n)=\alpha_{z_n}*X_{n-1}(z_{n-1})+\varepsilon_n(z_{n-1},z_n),\ n\in\mathbb{N}, \tag{3}$$

$$Y_n(z_n)=\beta_{z_n}*Y_{n-1}(z_{n-1})+\eta_n(z_{n-1},z_n),\ n\in\mathbb{N}. \tag{4}$$

In order to give a precise definition of the processes introduced in (3) and (4), we add some additional assumptions:

(C1) $\{\varepsilon_n(1,1)\},\{\varepsilon_n(1,2)\},\ldots,\{\varepsilon_n(r,r)\},\{\eta_n(1,1)\},\{\eta_n(1,2)\},\ldots,\{\eta_n(r,r)\}$ are mutually independent for all $n\in\mathbb{N}_0$,

(C2) $\varepsilon_m(i,j)$ and $\eta_m(i,j)$ are independent of $Y_n(k)$ and $X_n(k)$, respectively, for $n<m$ and for all $i,j,k\in E_r$,

(C3) the covariance between $X_n(z_n)$ and $Y_n(z_n)$ is the same as the covariance between $X_m(z_m)$ and $Y_m(z_m)$, when $z_n=z_m$.

Based on the Theorem 1 from Laketa et al. (2018), the distributions of the innovation series $\{\varepsilon_n\}$ and $\{\eta_n\}$ are given by the following relations:

$$\varepsilon_n(i,j) \stackrel{d}{=} \begin{cases} Geom\left(\frac{\mu_j}{1+\mu_j}\right), & \text{w.p. } 1-\frac{\alpha_j\mu_i}{\mu_j-\alpha_j}, \\ Geom\left(\frac{\alpha_j}{1+\alpha_j}\right), & \text{w.p. } \frac{\alpha_j\mu_i}{\mu_j-\alpha_j}. \end{cases} \tag{5}$$

$$\eta_n(i,j) \stackrel{d}{=} \begin{cases} Geom\left(\frac{\nu_j}{1+\nu_j}\right), & \text{w.p. } 1-\frac{\beta_j\nu_i}{\nu_j-\beta_j}, \\ Geom\left(\frac{\beta_j}{1+\beta_j}\right), & \text{w.p. } \frac{\beta_j\nu_i}{\nu_j-\beta_j}. \end{cases} \tag{6}$$

Now, we will present some new results of considered models. For the simplicity of notation, in the following text, we shall use X_n and Y_n instead of $X_n(z_n)$ and $Y_n(z_n)$, respectively. We will consider $\mathbf{X}_n = (X_n, Y_n)$ as a bivariate process named BRrNGINAR(1). Also, let us define vector $\boldsymbol{\mu}_n = \begin{bmatrix} \mu_{z_n} \\ \nu_{z_n} \end{bmatrix}$ and matrix $\mathbf{A}_n = \begin{bmatrix} \alpha_{z_n} & 0 \\ 0 & \beta_{z_n} \end{bmatrix}$.

The following theorem explains the process correlation structure.

Theorem 1 *(a) The covariance matrix of random variables $\mathbf{X}_n$ and $\mathbf{X}_{n-k}$, $k \in \{0,1,\ldots,n\}$, is given as*

$$Cov(\mathbf{X}_k, \mathbf{X}_0) = \mathbf{A}_1\mathbf{A}_2\ldots\mathbf{A}_k Cov(\mathbf{X}_0, \mathbf{X}_0), \tag{7}$$

(b) The correlation matrix of random variables $\mathbf{X}_n$ and $\mathbf{X}_{n-k}$, $k \in \{0,1,\ldots,n\}$, is given as

$$Corr(\mathbf{X}_k, \mathbf{X}_0) = \begin{bmatrix} \sqrt{\frac{Var(X_0)}{Var(X_k)}} & 0 \\ 0 & \sqrt{\frac{Var(Y_0)}{Var(Y_k)}} \end{bmatrix} \mathbf{A}_1\mathbf{A}_2\ldots\mathbf{A}_k Corr(\mathbf{X}_0, \mathbf{X}_0), \tag{8}$$

where $Var(X_i) = \frac{\mu_{z_i}}{1+\mu_{z_i}}$ and $Var(Y_i) = \frac{\nu_{z_i}}{1+\nu_{z_i}}$.

The proof is given in the Appendix.

Following theorem contains the results of the conditional expectations and variances.

Theorem 2 *a) The conditional expectation of the random variable $\mathbf{X}_{n+k}$ on $\mathbf{X}_n$ is given by*

$$E(\mathbf{X}_{n+k}|\mathbf{X}_n) = \mathbf{A}_{n+1}\mathbf{A}_{n+2}\ldots\mathbf{A}_{n+k}[\mathbf{X}_n - \boldsymbol{\mu}_n] + \boldsymbol{\mu}_{n+k},\ k \in \mathbb{N}_0, \tag{9}$$

b) The conditional variance of X_{n+k} on $\mathbf{X}_n$ is given by

$$\begin{aligned} Var(X_{n+k}|X_n,Y_n) = \Bigg\{ & \alpha_{z_{n+1}}(1+\alpha_{z_{n+1}}) \left(\prod_{s=2}^{k} \alpha_{z_{n+s}}^2 I\{k>1\} + I\{k=1\} \right) \\ & + \sum_{i=2}^{k-1} \left(\prod_{s=1}^{i-1} \alpha_{z_{n+s}} \right) \alpha_{z_{n+i}}(1+\alpha_{z_{n+i}}) \left(\prod_{s=i+1}^{k} \alpha_{z_{n+s}}^2 \right) I\{k>2\} \\ & + \left(\prod_{s=1}^{k} \alpha_{z_{n+s}} \right) \alpha_{z_{n+k}}(1+\alpha_{z_{n+k}}) I\{k>1\} \Bigg\} (X_n - \mu_{z_n}) \\ & + \mu_{z_{n+k}}(1+\mu_{z_{n+k}}) - \left(\prod_{s=1}^{k} \alpha_{z_{n+s}}^2 \right) \mu_{z_n}(1+\mu_{z_n}), \end{aligned}$$

and the conditional variance of Y_{n+k} on $\mathbf{X}_n$ is analogous.

c) The conditional probability mass function is given by

$$\begin{aligned} & P(X_n = x_n, Y_n = y_n | X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}, Z_n = z_n, Z_{n-1} = z_{n-1}) \\ & = P(X_n = x_n | X_{n-1} = x_{n-1}, Z_n = z_n, Z_{n-1} = z_{n-1}) \\ & \cdot P(Y_n = y_n | Y_{n-1} = y_{n-1}, Z_n = z_n, Z_{n-1} = z_{n-1}), \end{aligned}$$

where

$$\begin{aligned} & P(X_n = x_n | X_{n-1} = x_{n-1}, Z_n = z_n, Z_{n-1} = z_{n-1}) \\ & = \sum_{k=0}^{x_n} \binom{x_{n-1}+k-1}{x_{n-1}-1} \frac{\alpha_{z_{n-1}}^k}{(1+\alpha_{z_{n-1}})^{k+x_{n-1}}} \\ & \cdot \left[\left(1 - \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}} \right) \frac{\mu_{z_n}^{x_n-k}}{(1+\mu_{z_n})^{x_n-k+1}} + \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}} \cdot \frac{\alpha_{z_n}^{x_n-k}}{(1+\alpha_{z_n})^{x_n-k}} \right] I_{\{x_{n-1}\neq 0\}} \\ & + \left[\left(1 - \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}} \right) \frac{\mu_{z_n}^{x_n}}{(1+\mu_{z_n})^{x_n+1}} + \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}} \cdot \frac{\alpha_{z_n}^{x_n}}{(1+\alpha_{z_n})^{x_n}} \right] I_{\{x_{n-1}=0\}}, \end{aligned}$$

and the analogous formula holds for Y_n.

The proofs are given in the Appendix.

Remark 1 Regarding the correlation between $\{X_n(z_n)\}$ and $\{Y_n(z_n)\}$, the following can be said. Values of the processes $\{X_n(z_n)\}$ and $\{Y_n(z_n)\}$ are determined by the random process realization. Namely, certain parameter values of one component may only occur with the corresponding parameter values of another component. This explains the correlation between $\{X_n\}$ and $\{Y_n\}$, which cannot be calculated, since it is not a correlation in the standard sense and definition. However, as $\{z_n\}$ is determined by the clustering of the observed counting processes, it is actually this sequence, $\{z_n\}$, that contains in-

formation about this kind of correlation. Beside this, the standard covariance function $Cov(X_n, Y_m)$, for some $m,n \in \mathbb{N}$, can be different from zero, which is in fact used in the section about the Yule-Walker method of estimation of the unknown process parameters.

4 Parameter estimation

Let $X_1, X_2, \ldots, X_N$ and $Y_1, Y_2, \ldots, Y_N$ be samples from the RrNGINAR$_1(\mathscr{M}_1, \mathscr{A}_1, \mathscr{P}_1)$ and RrNGINAR$_1(\mathscr{M}_2, \mathscr{A}_2, \mathscr{P}_2)$ processes, respectively, where $\mathscr{M}_1 = \{\mu_1, \mu_2, \ldots, \mu_r\}$, $\mathscr{M}_2 = \{\nu_1, \nu_2, \ldots, \nu_r\}$, $\mathscr{A}_1 = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$, $\mathscr{A}_2 = \{\beta_1, \beta_2, \ldots, \beta_r\}$ and $\mathscr{P}_1 = \mathscr{P}_2 = \{1\}$ are the corresponding sets of unknown parameters. In the following subsections, two methods for parameter estimation are given: the Yule-Walker method and the conditional maximum likelihood method.

4.1 Yule-Walker estimation

The Yule-Walker parameter estimators are defined matching theoretical and empirical values of the correlation structure of the model. Recall that, usually the Yule-Walker estimation method (YW) assumes that the process is stationary. Since this assumption does not hold for the models with a random environment, because they have different states, it is necessary to define the Yule-Walker estimators using some parts of the sample, which can be considered stationary.

Let us define the set $I_k = \{i \in \{1,2,\ldots,N\} | z_i = z_{i+1} = k\}$ of indices i of the process elements $X_i(z_i)$ and $Y_i(z_i)$ corresponding to the state k, whose followers $X_{i+1}(z_{i+1})$ and $Y_{i+1}(z_{i+1})$ are also in the same state k and denote its number of elements by $n_k = |I_k|$.

$$\widehat{\mu}_k = \frac{1}{n_k}\sum_{i\in I_k} X_i(k), \quad \widehat{\gamma}_0^{X,k} = \frac{1}{n_k}\sum_{i\in I_k}(X_i(k)-\widehat{\mu}_k)^2,$$

$$\widehat{\nu}_k = \frac{1}{n_k}\sum_{i\in I_k} Y_i(k), \quad \widehat{\gamma}_0^{Y,k} = \frac{1}{n_k}\sum_{i\in I_k}(Y_i(k)-\widehat{\nu}_k)^2,$$

$$\widehat{\gamma}_1^{X,k} = \frac{1}{n_k}\sum_{i\in I_k}(X_{i+1}(k)-\widehat{\mu}_k)(X_i(k)-\widehat{\mu}_k),$$

$$\widehat{\gamma}_1^{Y,k} = \frac{1}{n_k}\sum_{i\in I_k}(Y_{i+1}(k)-\widehat{\nu}_k)(Y_i(k)-\widehat{\nu}_k),$$

$$\widehat{\gamma}_{10}^{X,Y,k} = \frac{1}{n_k}\sum_{i\in I_k}(X_{i+1}(k)-\widehat{\mu}_k)(Y_i(k)-\widehat{\nu}_k),$$

$$\widehat{\gamma}_{01}^{X,Y,k} = \frac{1}{n_k}\sum_{i\in I_k}(Y_{i+1}(k)-\widehat{\nu}_k)(X_i(k)-\widehat{\mu}_k),$$

$$\widehat{\gamma}_{00}^{X,Y,k} = \frac{1}{n_k}\sum_{i\in I_k}(X_i(k)-\widehat{\mu}_k)(Y_i(k)-\widehat{\nu}_k),$$

where

$$\begin{aligned}
&\gamma_1^{X,k} = Cov(X_n, X_{n+1}),\ \ \gamma_1^{Y,k} = Cov(Y_n, Y_{n+1}) \ \text{ if } \ z_n = z_{n+1} = k,\\
&\gamma_0^{X,k} = Var(X_n),\ \ \gamma_0^{Y,k} = Var(Y_n) \ \text{ if } \ z_n = k,\\
&\gamma_{00}^{X,Y,k} = Cov(X_n, Y_n) \ \text{ if } \ z_n = k,\\
&\gamma_{01}^{X,Y,k} = Cov(X_n, Y_{n+1}),\ \ \gamma_{10}^{X,Y,k} = Cov(X_{n+1}, Y_n) \ \text{ if } \ z_n = z_{n+1} = k.
\end{aligned}$$

These estimators are all strongly consistent, which can be shown by a similar proof as in Nastić et al. (2016). From the covariance properties, analysed in the previous section, it follows that

$$\begin{aligned}
Cov(X_{n+1}(z_{n+1}), Y_n(z_n)) &= \alpha_{z_{n+1}} Cov(X_n(z_n), Y_n(z_n)),\\
Cov(X_{n+1}(z_{n+1}), X_n(z_n)) &= \alpha_{z_{n+1}} Cov(X_n(z_n), X_n(z_n)),
\end{aligned}$$

so we can write

$$\alpha_{z_{n+1}} = \frac{1}{2}\left(\frac{Cov(X_{n+1}(z_{n+1}), Y_n(z_n))}{Cov(X_n(z_n), Y_n(z_n))} + \frac{Cov(X_{n+1}(z_{n+1}), X_n(z_n))}{Cov(X_n(z_n), X_n(z_n))}\right).$$

Let us now consider $X_n(z_n)$, such that $n \in I_k$. Then,

$$\alpha_k = \frac{1}{2}\left(\frac{Cov(X_{n+1}(k), Y_n(k))}{Cov(X_n(k), Y_n(k))} + \frac{Cov(X_{n+1}(k), X_n(k))}{Cov(X_n(k), X_n(k))}\right).$$

Therefore, we can estimate α_k in the following way

$$\widehat{\alpha}_k = \frac{1}{2}\left(\frac{\widehat{\gamma}_{10}^{X,Y,k}}{\widehat{\gamma}_{00}^{X,Y,k}} + \frac{\widehat{\gamma}_1^{X,k}}{\widehat{\gamma}_0^{X,k}}\right).$$

Similarly, we get

$$\widehat{\beta}_k = \frac{1}{2}\left(\frac{\widehat{\gamma}_{01}^{X,Y,k}}{\widehat{\gamma}_{00}^{X,Y,k}} + \frac{\widehat{\gamma}_1^{Y,k}}{\widehat{\gamma}_0^{Y,k}}\right).$$

From the consistency of the modified sample covariances follows the consistency of $\widehat{\alpha}_k$ and $\widehat{\beta}_k$.

4.2 Conditional maximum likelihood estimation

We also consider likelihood-based estimation method (CML), where we conduct the maximization of the log-likelihood function for the given sample $\{(X_1(z_1),Y_1(z_1)),\ldots,(X_N(z_N),Y_N(z_N))\}$. The function that needs to be maximized is of the form

$$\log L = \sum_{i=2}^{N} \log P((X_i,Y_i)=(x_i,y_i)|(X_{i-1},Y_{i-1})=(x_{i-1},y_{i-1})).$$

The conditional probability mass function is given by Theorem 2, where values X_0 and Y_0 are treated as known. The maximization procedure is conducted numerically using the `optim` function in the programming language `R`.

5 Analysis of prediction errors

In this section, we give the equations for the analysis of one-step-ahead prediction errors. Since model's prediction is conducted with two processes, survival and innovation, we analyse the prediction errors of these two processes separately. Since these two processes are unobservable, we will discuss their prediction errors in terms of conditional expectations. Namely, knowing the realization of the processes $\{X_n\}$ and $\{Y_n\}$ at the moment n, we calculate the conditional expectations of survival and innovation processes for that moment. This approach was discussed in detail in Freeland and McCabe (2004) for the univariate case and in Popović, Nastić and Ristić (2018) for the bivariate case. Here we use the similar methodology as in Popović et al. (2018). Notice that the survival and the innovation processes are mutually independent for known realization of the process $\{Z_n\}$.

Knowing all states up to moment n, we want to determine $P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m|X_n = x_n, Y_n = y_n, Z_n = z_n, X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}, Z_{n-1} = z_{n-1})$ and $P(\varepsilon_n(z_n,z_{n-1}) = x_n - m|X_n = x_n, Y_n = y_n, Z_n = z_n, X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}, Z_{n-1} = z_{n-1})$, and similarly for $\beta_{z_n} * Y_{n-1}(z_{n-1})$ and $\eta_n(z_n,z_{n-1})$. As stated above, we consider a model which is based on a realization of the process $\{Z_n\}$ i.e. $\{z_n\}$. Thus,

$$\begin{aligned} &P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m|X_n = x_n, Y_n = y_n, Z_n = z_n, X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}, Z_{n-1} = z_{n-1}) \\ &= \frac{P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m, X_n = x_n|Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})}{P(X_n = x_n|Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})} = f(m), \quad (10) \end{aligned}$$

where we have in mind that the survival component of the process X_n is independent of Y_n for known X_n, X_{n-1}, Z_n and Z_{n-1}. Function $f(m)$ is introduced here for practical reasons. The denominator is given in Theorem 2. Further, we calculate the nominator having in mind the definition of the process X_n, i.e. $X_n = \alpha_{z_n} * X_{n-1}(z_{n-1}) + \varepsilon_n(z_n,z_{n-1})$. Thus, for known X_n and X_{n-1}, the probability $P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m, X_n = x_n)$ is the

same as $P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m, \varepsilon_n(z_{n-1}, z_n) = x_n - m)$. According to the definition of the process, X_{n-1} is independent from ε_n (statement (C2)), so the nominator of the above equation is obtained as

$$\begin{aligned} &P(\alpha_{z_n} * X_{n-1}(z_{n-1}) = m, \varepsilon_n(z_{n-1}, z_n) = x_n - m | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1}) \\ &= P(NB(x_{n-1}, \alpha_{z_n}) = m) \cdot P(\varepsilon_n(z_{n-1}, z_n) = x_n - m). \end{aligned}$$

$NB(x_{n-1}, \alpha_{z_n})$ stands for a random variable with a negative binomial distribution with stated parameters. The probability of the random variable $\varepsilon_n(z_{n-1}, z_n)$ is given by equation (5) and it is equal to

$$P(\varepsilon_n(z_{n-1}, z_n) = x_n - m) = \left(1 - \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}}\right) \frac{\mu_{z_n}^{x_n - m}}{(1+\mu_{z_n})^{x_n - m + 1}} + \frac{\alpha_{z_n}\mu_{z_{n-1}}}{\mu_{z_n} - \alpha_{z_n}} \frac{\alpha_{z_n}^{x_n - m}}{(1+\alpha_{z_n})^{x_n - m + 1}}.$$

Further, the conditional distribution of the innovation process can be obtained following computations similar to those presented above for equation (10). Thus we have

$$\begin{aligned} &P(\varepsilon_n(z_{n-1}, z_n) = m | X_n = x_n, Y_n = y_n, Z_n = z_n, X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}, Z_{n-1} = z_{n-1}) \\ &= f(x_n - m). \end{aligned} \tag{11}$$

By using equations (10) and (11), we can derive the conditional expectations for the survival and innovation components, respectively. With $\mathscr{F}_n$, we denote the σ-algebra generated with (X_n, Y_n, Z_n), $(X_{n-1}, Y_{n-1}, Z_{n-1})$, $\ldots$, (X_0, Y_0, Z_0). Then, we have that

$$\begin{aligned} E(\alpha_{z_n} * X_{n-1}(z_{n-1}) | \mathscr{F}_n) = &\frac{x_{n-1}\alpha_{z_{n-1}}}{P(X_n = x_n | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})} \\ &\cdot P(X_n = x_n - 1 | Z_n = z_n, X_{n-1} = x_{n-1} + 1, Z_{n-1} = z_{n-1}), \end{aligned} \tag{12}$$

and

$$\begin{aligned} E(\varepsilon_n(z_{n-1}, z_n) | \mathscr{F}_n) = &\frac{1}{P(X_n = x_n | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})} \\ &\cdot \left[x_n P(X_n = x_n | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1}) \right. \\ &\left. - x_{n-1}\alpha_{z_{n-1}} P(X_n = x_n - 1 | Z_n = z_n, X_{n-1} = x_{n-1} + 1, Z_{n-1} = z_{n-1})\right]. \end{aligned} \tag{13}$$

The detailed derivations of equations (12) and (13) are given in Appendix. The analogue equations stand for $E(\beta_{z_n} * Y_{n-1}(z_{n-1}) | \mathscr{F}_n)$ and $E(\eta_n(z_n, z_{n-1}) | \mathscr{F}_n)$.

According to equation (9), the one-step-ahead prediction error at moment n, denoted as r_n, is equal to

$$\begin{aligned} r_n &= x_n(z_n) - (\alpha_{z_n} x_{n-1}(z_{n-1}) + \mu_{z_n} - \alpha_{z_n}\mu_{z_{n-1}}) \\ &= E(x_n(z_n)|\mathscr{F}_n) - \alpha_{z_n} x_{n-1}(z_{n-1}) - \mu_{z_n} + \alpha_{z_n}\mu_{z_{n-1}} \\ &= E(\alpha_{z_n} * x_{n-1}(z_{n-1}) + \varepsilon_{z_n}(z_{n-1}, z_n)|\mathscr{F}_n) - \alpha_{z_n} x_{n-1}(z_{n-1}) - \mu_{z_n} + \alpha_{z_n}\mu_{z_{n-1}} \\ &= E(\alpha_{z_n} * x_{n-1}(z_{n-1})|\mathscr{F}_n) - \alpha_{z_n} x_{n-1}(z_{n-1}) + E(\varepsilon_{z_n}(z_{n-1}, z_n)|\mathscr{F}_n) - \mu_{z_n} + \alpha_{z_n}\mu_{z_{n-1}}. \end{aligned}$$

We can conclude that the prediction error can be decomposed into two components. The first one is the prediction error of the survival process $r_{sur} = E(\alpha_{z_n} * X_{n-1}(z_{n-1})|\mathscr{F}_n) - \alpha_{z_n} X_{n-1}(z_{n-1})$ and the second one is the prediction error of the innovation process $r_{inn} = E(\varepsilon_{z_n}(z_{n-1}, z_n)|\mathscr{F}_n) - \mu_{z_n} + \alpha_{z_n}\mu_{z_{n-1}}$.

6 Model simulations

In this section, we test two methods for estimating the parameters of the BRrNGINAR(1) model on simulated data sets. The first method is the conditional maximum likelihood method where the conditional likelihood function can be obtained from Theorem 2, statement c). The second one is the Yule-Walker method presented in Section 4.1.

We simulate 100 samples of lengths 100, 500, 1000 and 5000. Using the Monte Carlo method, we generate a time series that evolves according to equations (3) and (4). The values for ε_n and η_n are picked randomly from the distribution determined by equations (5) and (6), respectively. Further, the values for components $\alpha_{z_n} * X_{n-1}(z_{n-1})$ and $\beta_{z_n} * Y_{n-1}(z_{n-1})$ are random numbers generated from the appropriate negative binomial distribution (where we take $X_0 = \varepsilon_0$ and $Y_0 = \eta_0$ as initial values).

The following parameters were used for the simulation procedure: a) $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $\beta_1 = 0.15$, $\beta_2 = 0.25$, $\mu_1 = 1$, $\mu_2 = 2$, $\nu_1 = 1$, $\nu_2 = 3$; b) $\alpha_1 = 0.45$, $\alpha_2 = 0.5$, $\beta_1 = 0.55$, $\beta_2 = 0.65$, $\mu_1 = 2$, $\mu_2 = 3$, $\nu_1 = 4$, $\nu_2 = 5$; c) $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $\alpha_3 = 0.25$, $\beta_1 = 0.15$, $\beta_2 = 0.25$, $\beta_3 = 0.25$, $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\nu_1 = 1$, $\nu_2 = 2$, $\nu_3 = 3$; d) $\alpha_1 = 0.35$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$, $\beta_1 = 0.4$, $\beta_2 = 0.25$, $\beta_3 = 0.35$, $\mu_1 = 2$, $\mu_2 = 3$, $\mu_3 = 4$, $\nu_1 = 3$, $\nu_2 = 4$ $\nu_2 = 5$. These values were chosen according to our experience in testing BRrNGINAR(1) as well as other bivariate models. We tried to determine the sets of parameters that are most likely to be found with real data sets. In all cases, we take into account the appropriate boundaries for the thinning parameters. The random environment processes with 2 and 3 random states are considered. For the cases a) and b), the probability vector of states is $(0.5, 0.5)$, while this vector has values $(0.3, 0.4, 0.3)$ for cases c) and d), so all the states are nearly equally probable. We set the transition probability matrix from state i to state j as $\begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$ for cases a) and b), and $\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$ for cases c) and d). They are chosen in such way that diagonal elements are the biggest in the matrices, so that the corresponding processes stay in the same state long enough. The estimated values obtained with the YW method are presented in Table 2 and Table 4, and with the

CML method in Table 1 and Table 3. Besides the estimated values of the parameters, there are also standard deviations of the estimates.

Table 1: *Estimated values of unknown parameters for the BRrNGINAR(1) model with two states using the conditional maximum likelihood method. The standard errors of estimates are given in the brackets.*

a)	$\alpha_1=0.1$	$\alpha_2=0.2$	$\beta_1=0.15$	$\beta_2=0.25$	$\mu_1=1$	$\mu_2=2$	$\nu_1=1$	$\nu_2=3$
100	0.1149	0.1879	0.1438	0.2118	0.9979	1.9679	0.9589	2.4212
	(0.095)	(0.0959)	(0.078)	(0.080)	(0.145)	(0.260)	(0.129)	(0.524)
500	0.1082	0.1982	0.1439	0.2345	1.0044	2.022	0.9773	2.715
	(0.045)	(0.058)	(0.034)	(0.046)	(0.083)	(0.162)	(0.095)	(0.367)
1000	0.099	0.202	0.1459	0.2423	1.0067	2.0106	0.9905	2.7554
	(0.0337)	(0.0350	(0.024)	(0.036)	(0.067)	(0.123)	(0.071)	(0.327)
5000	0.0988	0.1909	0.1527	0.2441	1.0022	1.9972	0.9982	2.9955
	(0.0138)	(0.014)	(0.012)	(0.015)	(0.022)	(0.042)	(0.028)	(0.051)
b)	$\alpha_1=0.45$	$\alpha_2=0.5$	$\beta_1=0.55$	$\beta_2=0.65$	$\mu_1=2$	$\mu_2=3$	$\nu_1=4$	$\nu_2=5$
100	0.4316	0.4454	0.5283	0.5894	1.989	2.9757	3.9493	4.9333
	(0.057)	(0.064)	(0.064)	(0.064)	(0.159)	(0.212)	(0.248)	(0.444)
500	0.4489	0.4745	0.5416	0.6257	2.0079	2.992	4.0002	4.9824
	(0.041)	(0.030)	(0.040)	(0.032)	(0.103)	(0.127)	(0.133)	(0.141)
1000	0.4516	0.4803	0.5444	0.6344	2.007	2.9871	4.0101	4.9915
	(0.029)	(0.025)	(0.027)	(0.027)	(0.096)	(0.109)	(0.130)	(0.155)
5000	0.4425	0.4877	0.5431	0.6401	1.985	2.988	3.9874	4.9854
	(0.021)	(0.012)	(0.020)	(0.014)	(0.064)	(0.065)	(0.074)	(0.092)

Table 2: *Estimated values of unknown parameters for BRrNGINAR(1) model with two states using Yule-Walker method. The standard errors of estimates are given in the brackets.*

a)	$\alpha_1=0.1$	$\alpha_2=0.2$	$\beta_1=0.15$	$\beta_2=0.25$	$\mu_1=1$	$\mu_2=2$	$\nu_1=1$	$\nu_2=3$
100	0.1681	0.2175	0.1671	0.2017	0.9671	1.9477	1.0168	2.8605
	(0.123)	(0.127)	(0.127)	(0.123)	(0.247)	(0.359)	(0.214)	(0.560)
500	0.1163	0.2004	0.1459	0.2382	1.0035	1.9938	1.0155	2.9531
	(0.075)	(0.079)	(0.073)	(0.074)	(0.097)	(0.186)	(0.104)	(0.262)
1000	0.1061	0.195	0.1516	0.2402	1.0007	1.9983	1.016	3.0034
	(0.057)	(0.058)	(0.048)	(0.055)	(0.065)	(0.119)	(0.077)	(0.221)
5000	0.1009	0.2029	0.1484	0.2529	0.994	1.9956	1.0003	3.0076
	(0.025)	(0.024)	(0.025)	(0.024)	(0.027)	(0.057)	(0.031)	(0.091)
b)	$\alpha_1=0.45$	$\alpha_2=0.5$	$\beta_1=0.55$	$\beta_2=0.65$	$\mu_1=2$	$\mu_2=3$	$\nu_1=4$	$\nu_2=5$
100	0.4149	0.449	0.5316	0.5641	1.9574	2.972	4.0829	5.1293
	(0.169)	(0.169)	(0.178)	(0.159)	(0.513)	(0.739)	(0.930)	(1.279)
500	0.4409	0.4803	0.5359	0.6358	1.9907	2.9747	3.9991	5.0536
	(0.085)	(0.077)	(0.076)	(0.075)	(0.235)	(0.358)	(0.447)	(0.588)
1000	0.4457	0.4877	0.5396	0.6361	1.9744	2.9839	4.0285	5.0532
	(0.064)	(0.057)	(0.054)	(0.053)	(0.163)	(0.245)	(0.288)	(0.398)
5000	0.4472	0.4988	0.5477	0.6486	1.9963	2.9922	3.9944	5.0088
	(0.027)	(0.029)	(0.025)	(0.024)	(0.064)	(0.100)	(0.130)	(0.197)

Table 3: *Estimated values of unknown parameters for the BRrNGINAR(1) model with three states using the conditional maximum likelihood method. The standard errors of estimates are given in the brackets.*

c)	$\alpha_1=0.1$	$\alpha_2=0.2$	$\alpha_3=0.25$	$\beta_1=0.15$	$\beta_2=0.25$	$\beta_3=0.25$	$\mu_1=1$	$\mu_2=2$	$\mu_3=3$	$\nu_1=1$	$\nu_2=2$	$\nu_3=3$
100	0.1424	0.1901	0.2078	0.1648	0.2044	0.2018	1.0127	1.985	2.9899	1.0063	1.9992	3.0107
	(0.067)	(0.052)	(0.052)	(0.060)	(0.049)	(0.048)	(0.097)	(0.129)	(0.081)	(0.057)	(0.046)	(0.115)
500	0.1162	0.1965	0.2231	0.1526	0.2255	0.2179	1.0156	1.9939	2.9887	1.0197	1.9997	2.9795
	(0.039)	(0.033)	(0.031)	(0.036)	(0.030)	(0.032)	(0.060)	(0.068)	(0.101)	(0.060)	(0.070)	(0.097)
1000	0.1048	0.2003	0.2272	0.1441	0.2337	0.2288	1.018	1.9874	2.9805	1.0161	2.0023	2.9915
	(0.031)	(0.030)	(0.024)	(0.035)	(0.022)	(0.023)	(0.040)	(0.059)	(0.083)	(0.042)	(0.050)	(0.075)
5000	0.099	0.1989	0.2403	0.1534	0.2372	0.2365	1.0029	2.0027	2.9992	1.0039	1.9941	3.0019
	(0.012)	(0.018)	(0.015)	(0.013)	(0.011)	(0.014)	(0.017)	(0.029)	(0.033)	(0.021)	(0.031)	(0.037)
d)	$\alpha_1=0.35$	$\alpha_2=0.4$	$\alpha_3=0.4$	$\beta_1=0.4$	$\beta_2=0.25$	$\beta_3=0.35$	$\mu_1=2$	$\mu_2=3$	$\mu_3=4$	$\nu_1=3$	$\nu_2=4$	$\nu_2=5$
100	0.3211	0.343	0.3357	0.3571	0.2844	0.3417	2.0001	3.0044	3.9411	3.0038	4.0015	4.98
	(0.070)	(0.059)	(0.071)	(0.078)	(0.072)	(0.072)	(0.098)	(0.237)	(0.264)	(0.135)	(0.192)	(0.191)
500	0.3422	0.3659	0.3666	0.3827	0.2644	0.3446	2.009	3.0076	3.9788	2.9994	3.9926	4.9844
	(0.034)	(0.036)	(0.036)	(0.043)	(0.039)	(0.043)	(0.100)	(0.100)	(0.174)	(0.139)	(0.221)	(0.227)
1000	0.3471	0.3797	0.3746	0.3965	0.2586	0.3475	1.9973	3.004	3.9739	3.0007	3.9998	4.9766
	(0.029)	(0.025)	(0.028)	(0.037)	(0.033)	(0.035)	(0.051)	(0.097)	(0.131)	(0.080)	(0.09)	(0.123)
5000	0.3489	0.3926	0.3835	0.3961	0.2518	0.3418	2.0004	2.9981	3.9959	3.0011	4.0003	5.0014
	(0.011)	(0.012)	(0.015)	(0.014)	(0.016)	(0.016)	(0.035)	(0.042)	(0.042)	(0.040)	(0.050)	(0.052)

Table 4: *Estimated values of unknown parameters for the BRrNGINAR(1) model with three states using the Yule-Walker method. The standard errors of estimates are given in the brackets.*

c)	$\alpha_1=0.1$	$\alpha_2=0.2$	$\alpha_3=0.25$	$\beta_1=0.15$	$\beta_2=0.25$	$\beta_3=0.25$	$\mu_1=1$	$\mu_2=2$	$\mu_3=3$	$\nu_1=1$	$\nu_2=2$	$\nu_3=3$
100	0.2347	0.3121	0.3808	0.2396	0.3024	0.2697	1.0124	1.973	2.9647	1.0669	1.9626	2.9263
	(0.241)	(0.203)	(0.249)	(0.224)	(0.259)	(0.164)	(0.250)	(0.445)	(0.637)	(0.283)	(0.429)	(0.639)
500	0.1532	0.1906	0.2612	0.1478	0.2529	0.2612	0.9999	1.9868	2.9664	0.998	2.0213	2.9839
	(0.107)	(0.102)	(0.122)	(0.097)	(0.147)	(0.117)	(0.111)	(0.205)	(0.311)	(0.117)	(0.203)	(0.276)
1000	0.1221	0.1816	0.2544	0.1384	0.2597	0.2377	1.0099	1.9862	2.9945	0.9963	2.0264	2.987
	(0.073)	(0.090)	(0.086)	(0.083)	(0.102)	(0.096)	(0.085)	(0.144)	(0.232)	(0.090)	(0.151)	(0.200)
5000	0.0978	0.1949	0.2461	0.1492	0.2534	0.2492	0.9983	1.9849	2.9967	1.0083	2.0047	3.0044
	(0.045)	(0.040)	(0.039)	(0.041)	(0.044)	(0.041)	(0.036)	(0.068)	(0.090)	(0.043)	(0.073)	(0.090)
d)	$\alpha_1=0.35$	$\alpha_2=0.4$	$\alpha_3=0.4$	$\beta_1=0.4$	$\beta_2=0.25$	$\beta_3=0.35$	$\mu_1=2$	$\mu_2=3$	$\mu_3=4$	$\nu_1=3$	$\nu_2=4$	$\nu_2=5$
100	0.35	0.4054	0.3785	0.4183	0.3304	0.3102	1.9593	2.9636	3.8851	3.0181	3.9338	5.0608
	(0.172)	(0.203)	(0.236)	(0.235)	(0.207)	(0.217)	(0.501)	(0.708)	(0.895)	(0.719)	(0.809)	(1.313)
500	0.3356	0.4027	0.3935	0.3819	0.2583	0.3533	2.0347	3.0377	4.0598	2.9842	3.9981	5.0265
	(0.134)	(0.108)	(0.120)	(0.118)	(0.117)	(0.121)	(0.230)	(0.335)	(0.407)	(0.322)	(0.447)	(0.525)
1000	0.3311	0.4026	0.3979	0.386	0.2496	0.355	2.0031	3.0253	4.0108	3.0082	4.0012	5.033
	(0.094)	(0.094)	(0.081)	(0.089)	(0.081)	(0.082)	(0.162)	(0.231)	(0.292)	(0.216)	(0.284)	(0.351)
5000	0.3458	0.4028	0.4004	0.3987	0.2452	0.3486	1.9961	2.9871	4.0118	3.0009	3.9956	4.9814
	(0.043)	(0.041)	(0.039)	(0.040)	(0.038)	(0.037)	(0.068)	(0.110)	(0.135)	(0.102)	(0.126)	(0.154)

From the presented results, we can conclude that the estimates converge to the true values with the growth of the sample length, which is followed by the decrease of the standard deviation of the estimates. We can notice that both methods perform better when the true values of the parameters are larger (this will be important when we discuss the results from the application section). A probable reason for that is that when the parameters take small values, the generated series have a lot of zeros. So, these methods need bigger samples to estimate parameters of such "flattened" series.

The CML method provides good results even for samples of length 100. Also, there is no influence on the estimates with respect to the number of random states. On the other hand, the number of random states has a large impact on YW estimates when the sample length is 100. When there are only two states, YW performs similarly as the CML method. With three states, YW provides quite unprecise estimates for samples of length 100. The reason for that lies in the fact that the correlation functions are calculated on small sub-samples, thus their values are not very reliable. So, we can notice large deviations from the true values in the test c). The estimates are much better when the length of the sample is 500 or larger. The estimates of the parameters μ_i and ν_i, $i \in \{1,2,3\}$ converge very quickly with both methods, regardless of the number of states.

The probability vector of states and the transition probability matrix are estimated regardless of YW and CML methods. The probability vector is estimated by dividing the number of occurrences of a state by the length of the sample, while for the transition probability matrix, the number of transitions from state i to state j is divided by the total number of occurrences of states i. This way, we obtain very precise results for all studied samples, thus we omit a detailed discussion here.

We can conclude that CML is much more reliable for small samples (when the length of series is 100). On the other hand, a disadvantage of the CML method is that CML estimates are obtained numerically, thus the CML method is much more time consuming. The YM method provides estimates quite close to the real values when the sample size is 500 or greater and, since it has the analytical solution, it proves to be a better choice than CML for large samples.

The estimation procedure was conducted by using the Monte Carlo simulation. Thus, for each of 100 sample paths we estimate the model parameters. So, for each of these parameters, we get series of 100 values. The mean values and the sample standard deviations of these series are presented respectively as the estimations and their standard errors in Tables 1-4.

7 Application

This section is devoted to the practical aspect of the model. We test the model on a real data set and compare the results to some other known bivariate models. The comparison is based on the ability of the model to predict a value one step ahead for the observed

time series. The goodness of fit is measured in terms of the root mean square error (RMS). Also, we provide values for the Akaike information criterion (AIC), but since we are focused on the forecasting ability of the model, the main attention is paid to the values of RMS.

Parameters of the model are estimated using the conditional maximum likelihood method. As we have concluded in the previous section, the YW method is not very reliable for samples of length 100. Since the series that we deal with in this section have between 105 and 144 observations, we will only use the CML method for parameter estimation.

We compare the BRrNGINAR(1) model with three other bivariate models. Two of these models are with constant coefficients and dependent innovation processes, where one evolves under the Poisson bivariate distribution (BVPOIBINAR(1) model) and the other evolves under the negative binomial distribution (BVNBIBINAR(1) model). Both models were presented in Pedeli and Karlis (2011). The third model that we use for comparison was presented in Popović et al. (2016), it has random coefficients and independent innovation processes (BVGGINAR(1) model).

We test our model on three data sets. First, we consider the data set that contains two series of different events observed in the same region. Then, we focus on bivariate time series composed of data of the same event, observed in different regions. The third test considers two series of data of the same type of event that evolve in the same environment. In all three cases, we assume that the same factors influence both observed series.

7.1 Different events observed in the same region

First, we will test our model on the same data series as in Popović et al. (2016). These series are monthly counts of robberies (ROBB) and aggravated assaults (AGGASS) from January 1990 to December 2001 (for more details about these time series, see Popović et al. (2016)). The observed series together with their ACF and PACF are given in Figure 1. The bar plots in Figure 1 imply a higher level of activities in the first half compared to the end of the series. The series fluctuate around different means during two periods, so the BRrNGINAR(1) model might be appropriate since it has the ability to capture these changes of the frequency.

The results can be found in Table 5. It can be noticed that RMS for both series is the lowest for the BRrNGINAR(1) model. For the observed series, we have detected two states. According to this conclusion, we define the BRrNGINAR(1) model. For the BRrNGINAR(1) model, the main drawback is the number of parameters, but as we can see the model produces the lowest prediction errors, especially for the AGGASS series, and the lowest AIC value.

Table 5: *Parameter estimates of INAR models, RMS and AIC for ROBB and AGGASS data series.*

Model	CML estimates	RMS ROBB	RMS AGGASS	AIC
BRrNGINAR(1)	$\hat{\alpha}_1 = 0.515(0.008)$, $\hat{\alpha}_2 = 0.568(0.021)$ $\hat{\beta}_1 = 0.259(0.105)$, $\hat{\beta}_2 = 0.37(0.059)$ $\hat{\mu}_1 = 2.388(0.001)$, $\hat{\mu}_2 = 3.205(0.001)$ $\hat{\nu}_1 = 1.117(0.001)$, $\hat{\nu}_2 = 2.018(0.001)$	2.376	1.648	1044.45
BVGGINAR(1)	$\hat{\alpha} = 0.499(0.052)$, $\hat{p} = 0.887(0.12)$, $\hat{a} = 2.877(0.328)$ $\hat{\beta} = 0.281(0.058)$, $\hat{q} = 0.805(0.192)$, $\hat{b} = 1.765(0.187)$	2.496	1.827	1065.83
BVPOIBINAR(1)	$\hat{\alpha}_1 = 0.413(0.042)$, $\hat{\lambda}_1 = 1.664(0.148)$ $\hat{\alpha}_2 = 0.21(0.053)$, $\hat{\lambda}_2 = 1.389(0.128)$, $\hat{\phi} = 0.443(0.107)$	2.541	1.857	1183.76
BVNBIBINAR(1)	$\hat{\alpha}_1 = 0.413(0.046)$, $\hat{\lambda}_1 = 1.665(0.205)$ $\hat{\alpha}_2 = 0.169(0.061)$, $\hat{\lambda}_2 = 1.461(0.182)$, $\hat{\beta} = 0.883(0.176)$	2.541	1.88	1077.39

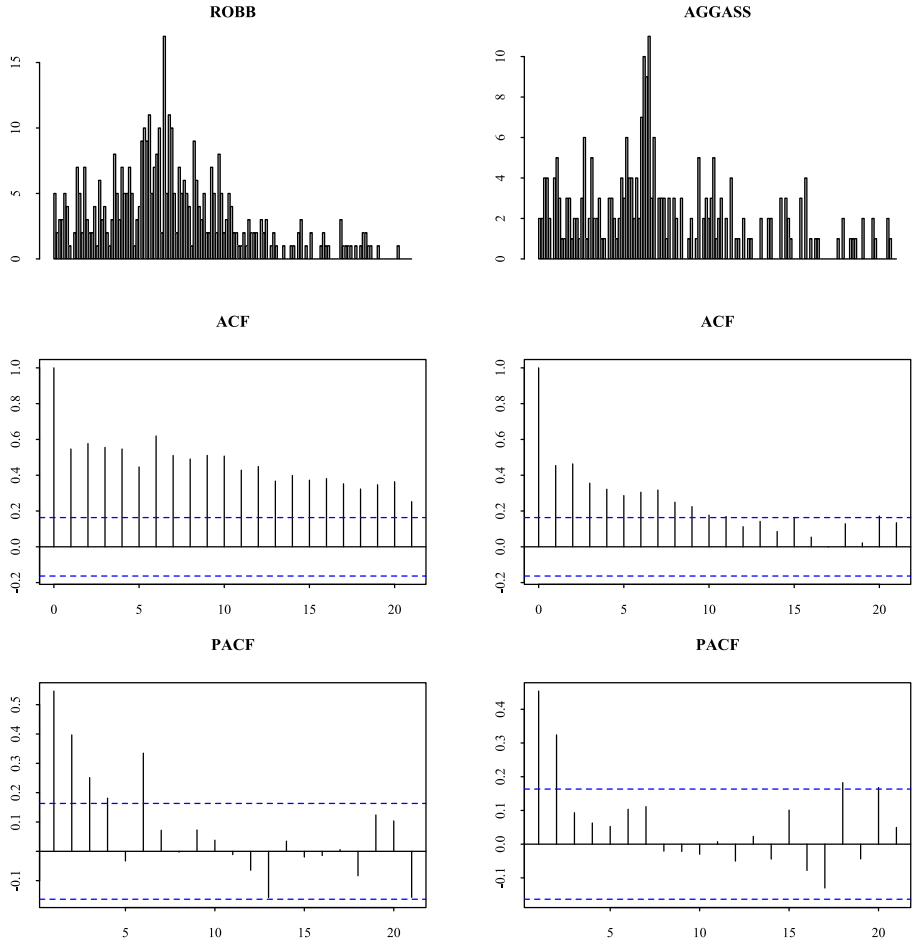


Figure 1: *Bar plots, autocorrelation and partial autocorrelation functions of robberies and aggravated assaults recorded in one police station.*

Table 5 contains the estimated values of model parameters as well as the standard errors of these estimates. Since the estimates are obtained with the CML method, these standard errors are computed as the square root of the diagonal elements of the inverse of the Hessian (the `optim` function from the programming language `R` can return the Hessian). The same holds for Table 6 and Table 7 that are going to be discussed in the next two subsections.

Table 6: *Parameter estimates of INAR models, RMS and AIC for SIMPASS-A and SIMPASS-B data series.*

Model	CML estimates	RMS SIMPASS-A	RMS SIMPASS-B	AIC
BRrNGINAR(1)	$\hat{\alpha}_1 = 0.502(0.131)$, $\hat{\alpha}_2 = 0.507(0.001)$ $\hat{\beta}_1 = 0.52(0.088)$, $\hat{\beta}_2 = 0.63(0.191)$ $\hat{\mu}_1 = 1.768(0.001)$, $\hat{\mu}_2 = 2.485(0.001)$ $\hat{\nu}_1 = 3.877(0.552)$, $\hat{\nu}_2 = 4.52(0.454)$	1.448	2.164	1066.63
BVGGINAR(1)	$\hat{\alpha} = 0.492(0.001)$, $\hat{p} = 0.558(0.075)$, $\hat{a} = 2.076(0.001)$ $\hat{\beta} = 0.65(0.041)$, $\hat{q} = 0.291(0.053)$, $\hat{b} = 2.075(0.001)$	1.612	2.48	1118.61
BVPOIBINAR(1)	$\hat{\alpha}_1 = 0.315(0.065)$, $\hat{\lambda}_1 = 1.544(0.171)$ $\hat{\alpha}_2 = 0.294(0.064)$, $\hat{\lambda}_2 = 2.96(0.303)$, $\hat{\phi} = 0.42(0.201)$	1.588	2.243	1053.74
BVNBIBINAR(1)	$\hat{\alpha}_1 = 0.33(0.067)$, $\hat{\lambda}_1 = 1.512(0.183)$ $\hat{\alpha}_2 = 0.345(0.066)$, $\hat{\lambda}_2 = 2.744(0.319)$, $\hat{\beta} = 0.168(0.065)$	1.584	2.238	1043.91

Table 7: *Parameter estimates of INAR models, RMS and AIC for Bitfinex and Kraken data series.*

Model	CML estimates	RMS Bitfinex	RMS Kraken	AIC
BRrNGINAR(1)	$\hat{\alpha}_1 = 0.819(0.001)$, $\hat{\alpha}_2 = 0.767(0.084)$ $\hat{\beta}_1 = 0.83(0.001)$, $\hat{\beta}_2 = 0.829(0.011)$ $\hat{\mu}_1 = 20.782(1.647)$, $\hat{\mu}_2 = 24.362(1.619)$ $\hat{\nu}_1 = 10.056(1.252)$, $\hat{\nu}_2 = 11.117(0.985)$	9.611	4.142	1333.53
BVGGINAR(1)	$\hat{\alpha} = 0.497(0.012)$, $\hat{p} = 0.783(0.069)$, $\hat{a} = 23.361(0.001)$ $\hat{\beta} = 0.433(0.001)$, $\hat{q} = 0.233(0.054)$, $\hat{b} = 10.138(0.001)$	11.415	4.727	1522.22
BVPOIBINAR(1)	$\hat{\alpha}_1 = 0.515(0.022)$, $\hat{\lambda}_1 = 11.409(0.583)$ $\hat{\alpha}_2 = 0.558(0.037)$, $\hat{\lambda}_2 = 4.529(0.405)$, $\hat{\phi} = 4.465(0.416)$	10.802	4.287	1603.54
BVNBIBINAR(1)	$\hat{\alpha}_1 = 0.611(0.023)$, $\hat{\lambda}_1 = 9.142(1.141)$ $\hat{\alpha}_2 = 0.666(0.026)$, $\hat{\lambda}_2 = 3.432(0.466)$, $\hat{\beta} = 1.198(0.252)$	10.509	4.263	1273.17

7.2 The same event observed in the different regions

The BRrNGINAR(1) model evolves under hidden time series that represents certain states of the observed series. Thus, the observed series are affected by some common factor. To find the most realistic scenario, we will focus on the time series of the same event that took place in different regions. From the database that can be found on website `http://www.forecastingprinciples.com`, we examine the number of

simple assaults recorded in two police stations located in Rochester. These data were recorded from January 1990 to December 2000 in police stations number 36055009401 and 36055009602, so we denote the data series SIMPASS-A and SIMPASS-B, respectively. The mean values of these series are 2.24 and 4.23, while the variances are 3.11 and 5.79, respectively. The correlation coefficient between the two series is 0.29. The autocorrelation coefficients at lag one are 0.45 and 0.36 for SIMPASS-A and SIMPASS-B, respectively.

The bar plots of the observed series are given in Figure 2. We can notice some similar patterns in the evolution of these two series. The bar plots in Figure 2 imply a higher rate of activities at the beginning compared to the end of the observed data. This suggests the existence of two or more random states for the BRrNGINAR(1) model. The BRrNGINAR(1) with two random states shows better performance than the model with three random states in terms of RMS. Models with more than three random states are not adequate for these series, since the observed data set is not long enough to properly estimate all parameters of such models.

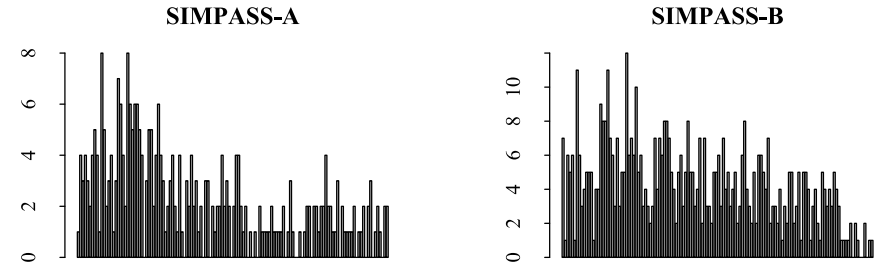


Figure 2: *Bar plots for simple assaults recorded in the two police stations.*

Since the random states have to be the same for both series, they are defined in the following way. The step one is to determine states for each series separately. This procedure is performed by using the quantiles of the observed series. Since we have only two states, we use the median as the boundary for determining states. Then, the states for the BRrNGINAR(1) model are determined as rounded average values of the states from step one for each observed moment. The states are given in Figure 3. In some cases, two observed values of one series have different states although they are equal. This is the consequence of determining random states for both series. But, in spite of this, it can be noticed that observed values are grouped into clusters.

Once again, we will compare the BRrNGINAR(1) model to three bivariate models mentioned above. The results are summarized in Table 6. We can notice that the BRrNGINAR(1) model achieved a much lower RMS for both observed series. Since we examine two time series of the same criminal activity, we can expect that the same factors affect the generation of these series. For example, unemployment or lack of police officers will encourage someone to commit a criminal act such as a simple as-

sault. Our model is based on this assumption, and as such it provides the best results from the forecasting point of view. We can notice that the AIC value is quite close to the BVPOIBINAR(1) and BVNBIBINAR(1) models. Since the BRrNGINAR(1) model depends on a larger number of parameters, it was expected to have a little bit larger value of the AIC.

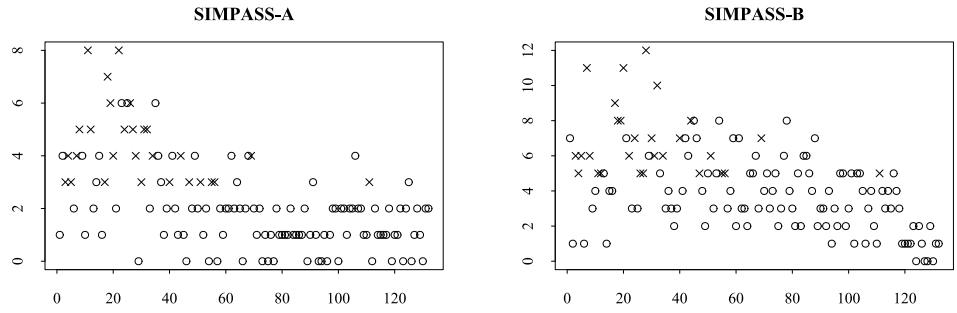


Figure 3: *States for SIMPASS-A and SIMPASS-B data series. The state one is denoted with* $\times$ *and the state two with* $\circ$*.*

For the analysis of the prediction error made by the BRrNGINAR(1) model, we use the approach discussed in Section 5. On the data sets SIMPASS-A and SIMPASS-B, the model makes the root mean square errors of 1.448 and 2.164, respectively. It can be said that these errors are produced by two sources, the prediction of the survival process and the prediction of the innovation process. We measure the prediction error of the survival component as the difference between the value calculated from equation (12) and the first addend of equation (14) when $k = 1$. Similarly, the prediction error of the innovation component is the difference between the value calculated from equation (13) and the second addend of equation (14) when $k = 1$. The residuals are presented in Figure 4.

The black line shows the series of the prediction errors created by the survival component, while the gray one represents the error of the innovation component. The dots are the actual prediction errors that we get when we apply the BRrNGINAR(1) model to the two observed series. It can be noticed that the two components produce errors with the opposite sign. Actually, the correlation coefficient between the two components for SIMPASS-A series is -0.55, and for SIMPASS-B it is -0.75. As a result of these negatively correlated errors, the actual prediction error is reduced. The most obvious consequence of this kind of behaviour can be noticed on the tenth observed value of the SIMPASS-A data set and on the seventh observed values of the SIMPASS-B data set.

It cannot be said that one or the other component produces larger errors. The behaviours of the survival and the innovation processes are quite similar. One of the most probable reasons for this is that they are both driven by the same hidden process which determines the states of the observed series.

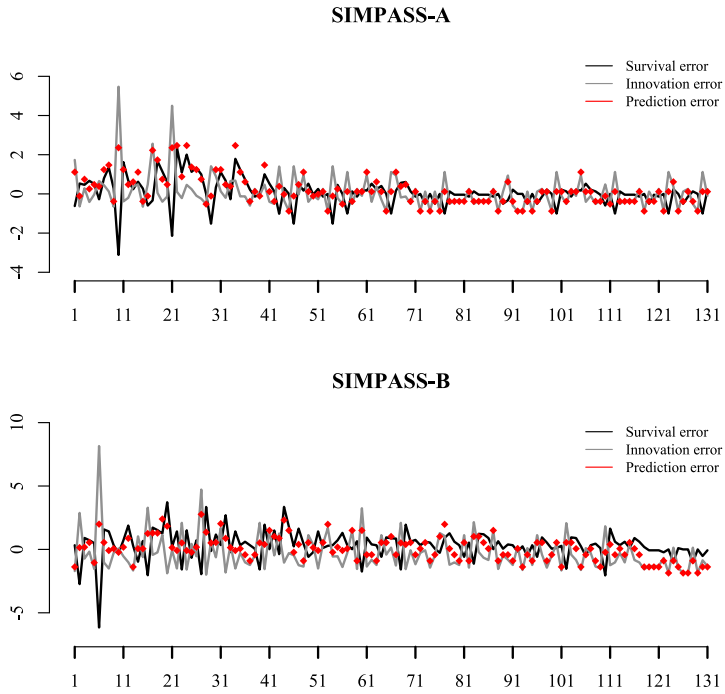


Figure 4: *Prediction errors produced by the survival and the innovation processes.*

7.3 The same type of event in the same environment

In order to more clearly motivate the introduction of the BRrNGINAR(1) model, we give another example where we test the model on two time series that nowadays spark a lot of interest among many people. Namely, we observe the volumes of two cryptocurrencies traded on weekly bases. The data set contains the traded volumes from the beginning of April 2017, until the end of March 2019, for cryptocurrencies Bitfinex and Kraken. We denote the smallest fraction of a coin that can be traded as a unit. Since these cryptocurrencies are traded in vary small fractions, the data that we present here are in 10^{12} units. So, the average values of these series are 23.36 and 10.14, respectively (which are actually 23.36×10^{12} and 10.14×10^{12} units). The standard deviations for the two series are 14.93 and 5.64, respectively, while the correlation between the series is 0.53. The autocorrelation coefficients on lag one are 0.71 and 0.64 for Bitfinex and Kraken, respectively.

Both series are presented in Figure 5. From the bar plots, we can conclude that similar factors influence weekly volumes for these two cryptocurrencies. We can clearly distinguish periods of high and low trading intensity. Thus, a stationary model for these two series would not be the best choice. Also, for the second series we can notice that the three periods of high volumes are followed by low market activities, which is the

usual trading behavior. These three peeks occurred on the last week of May 2017, the second week of February 2018 and the last week of November 2018. Thus, we cannot conclude that high trading volumes are connected to a specific time of year, nor that they occur after a certain period.

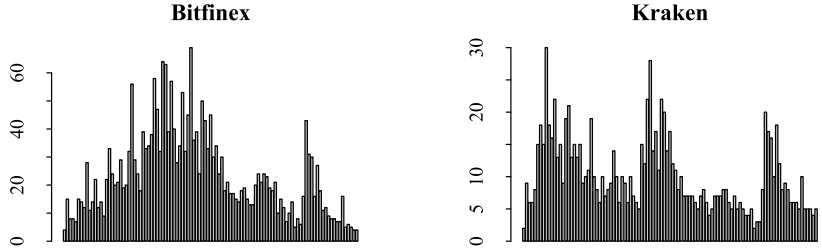


Figure 5: *Bar plots for weekly traded volumes of Bitfinex and Kraken cryptocurrencies.*

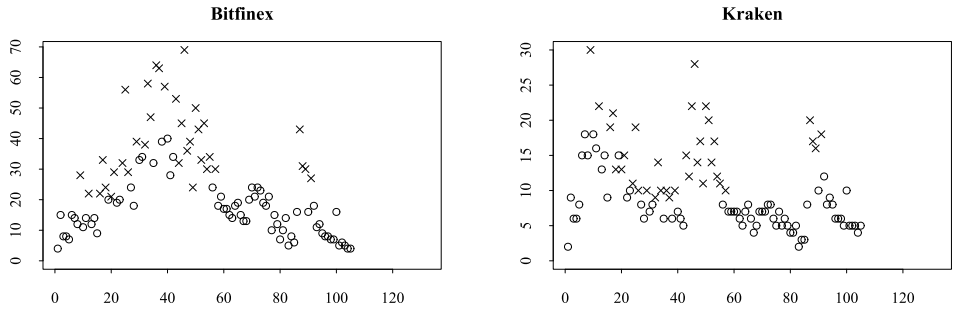


Figure 6: *States for weekly traded volumes of Bitfinex and Kraken cryptocurrencies. The state one is denoted with* $\times$ *and the state two with* $\circ$.

We notice two states of trading intensities for our model, which implies that we define the BRrNGINAR(1) with two random states, i.e. $r = 2$. These states are presented in Figure 6. Similarly as in two previous examples, following these states, we estimate the coefficient with the CML method and compare the results with the other three mentioned models. The obtained results are given in Table 7.

The values in Table 7 suggest that the BRrNGINAR(1) model has the smallest RMS. Thus, from the forecasting perspective, this model shows the best results. The advantage of the BRrNGINAR(1) model can be noticed especially with Bitfinex series, and some improvements are present with Kraken series as well. The reason for that probably lies in the fact that the observed series is non-stationary. As we can see with Bitfinex series, the BRrNGINAR(1) model estimated the mean value as $\hat{\mu}_1 = 20.782$ and $\hat{\mu}_2 = 24.362$, depending on the state. Other tested models have only one parameter for modelling the mean value. Even with the Kraken series where the difference between parameters $\hat{\nu}_1$ and $\hat{\nu}_2$ is not that big, we can see the improvements with RMS. The BRrNGINAR(1)

model achieves the second best value of AIC, which is a consequence of a larger number of parameters. Beside the fact that the number of parameters increases the AIC value, the estimated values of all these parameters have some deviation from the real values when the series are of length 105, as they are is this case. This fact also increases AIC value to a certain extent, having in mind the definition of AIC.

The purpose of all this testing is not to point out one model as the best model, but to demonstrate the type of series for which the BRrNGINAR(1) model is an adequate one. All observed series have in common that they fluctuate around different means during their evolution, which is expected to see when observing non-stationary series. This kind of behaviour looks like the series have trend, but not trend that can be easily captured with some linear or quadratic function, for example. These series take values from different intervals in different time frames which can be captured (in some degree) with the presented model.

8 Conclusion

The paper discusses a bivariate integer-valued autoregressive model of order one. The model is composed of two univariate models driven by the same hidden process. This hidden process is determined by the states that are assigned to the observed data. So, the hidden process allows the model to adjust itself to environment changes. As such, the model is non-stationary. Besides the main properties of the model, the focus is placed on its forecasting ability. Through tests on real data sets, it was shown that the model produces the smallest one-step-ahead prediction errors in terms of the root mean square error. Also, prediction errors are analysed in more detail by investigating prediction errors of each model component, the survival and the innovation component. These two components produce negatively correlated one-step-ahead prediction errors. This fact contributes to the reduction of the prediction errors which the model makes. The model contains a large number of parameters, so it requires a little bit larger data set for parameter estimation.

9 Appendix

Theorem 1

Proof.

a) Using the properties of the negative binomial thinning operator we have

$$Cov(X_k, Y_l) = \alpha_{z_k} Cov(X_{k-1}, Y_l), \quad Cov(X_k, X_l) = \alpha_{z_k} Cov(X_{k-1}, X_l),$$

$$Cov(X_k,Y_l)=\beta_{z_l}Cov(X_k,Y_{l-1}),\quad Cov(X_k,X_l)=\beta_{z_l}Cov(X_k,X_{l-1}).$$

Now, from these equalities, we simply get what is required.

b) This is obvious, based on the results given in a) and the fact that correlation is defined using the covariance. ∎

Theorem 2

Proof. It holds that

$$E(X_{n+k}|X_n)=\left(\prod_{s=1}^{k}\alpha_{z_{n+s}}\right)X_n+\sum_{l=1}^{k-1}\left(\prod_{s=1}^{k}\alpha_{z_{n+s}}\right)E(\varepsilon_{n+l})+E(\varepsilon_{n+k}),\qquad(14)$$

from the properties of the negative binomial thinning operator. If we take into account the distribution of the residuals, we get

$$E(X_{n+k}|X_n)=\left(\prod_{s=1}^{k}\alpha_{z_{n+s}}\right)(X_n-\mu_{z_n})+\mu_{z_{n+k}}.$$

The analogous relation holds for the Y component, so the required relation in a) is valid.

For the proof of b), the recurrent relation

$$\begin{aligned}Var(X_{n+k}|X_n,Y_n) &= \alpha^2_{z_{n+k}}Var(X_{n+k-1}|X_n,Y_n)\\ &\quad+\alpha_{z_{n+k}}(1+\alpha_{z_{n+k}})E(X_{n+k-1}|X_n)+Var(\varepsilon_{n+k})\end{aligned}$$

is used.

The statement c) follows from the fact that X_n and Y_n are independent for known Z_n, X_{n-1}, Y_{n-1} and Z_{n-1}. Also, Z_n is independent from X_{n-1} and Y_{n-1} for known Z_{n-1}. From the definition of the process $\{(X_n,Y_n)\}$, we have that

$$P(X_n=x_n|X_{n-1}=x_{n-1},Z_n=z_n,Z_{n-1}=z_{n-1})=P\left(\sum_{i=1}^{x_{n-1}}U_i^{z_{n-1}}+\varepsilon_n(z_{n-1},z_n)=x_n\right),$$

$$P(Y_n=y_n|Y_{n-1}=y_{n-1},Z_n=z_n,Z_{n-1}=z_{n-1})=P\left(\sum_{i=1}^{y_{n-1}}V_i^{z_{n-1}}+\eta_n(z_{n-1},z_n)=y_n\right).$$

Therefore, the statement c) is obtained using the above equations and properties of the residuals. ∎

Equation (12)

Proof. For simplicity, we will denote the probability mass function in the denominator as $P(X_n=x_n|Z_n=z_n,X_{n-1}=x_{n-1},Z_{n-1}=z_{n-1})=P(X_n=x_n|z_n,x_{n-1},z_{n-1})$. Now we

have

$$
\begin{aligned}
&E(\alpha_{z_n} * X_{n-1}(z_{n-1})|\mathscr{F}_n) = \sum_{j=0}^{x_n} j \cdot f(j) \\
&= \sum_{j=0}^{x_n} j \cdot \frac{P(NB(x_{n-1},\alpha_{z_n}) = j) \cdot P(\varepsilon_n(z_{n-1},z_n) = x_n - j)}{P(X_n = x_n|z_n,x_{n-1},z_{n-1})} \\
&= \frac{\sum_{j=0}^{x_n} j \binom{x_{n-1}+j-1}{j} \frac{\alpha_{z_{n-1}}^j}{(1+\alpha_{z_{n-1}})^{x_{n-1}+j}} P(\varepsilon_n(z_{n-1},z_n) = x_n - j)}{P(X_n = x_n|z_n,x_{n-1},z_{n-1})} \\
&= \frac{x_{n-1}\frac{\alpha_{z_{n-1}}}{1+\alpha_{z_{n-1}}} \sum_{j=1}^{x_n} \binom{x_{n-1}+j-1}{j-1} \frac{\alpha_{z_{n-1}}^{j-1}}{(1+\alpha_{z_{n-1}})^{x_{n-1}+j-1}} P(\varepsilon_n(z_{n-1},z_n) = x_n - j)}{P(X_n = x_n|z_n,x_{n-1},z_{n-1})} \\
&= \frac{x_{n-1}\frac{\alpha_{z_{n-1}}}{1+\alpha_{z_{n-1}}} \sum_{j=0}^{x_n-1} \binom{x_{n-1}+1+j-1}{j} \frac{\alpha_{z_{n-1}}^j}{(1+\alpha_{z_{n-1}})^{x_{n-1}+j}} P(\varepsilon_n(z_{n-1},z_n) = x_n - 1 - j)}{P(X_n = x_n|z_n,x_{n-1},z_{n-1})} \\
&= x_{n-1}\alpha_{z_{n-1}} \sum_{j=0}^{x_n-1} \frac{\binom{x_{n-1}+1+j-1}{j} \frac{\alpha_{z_{n-1}}^j}{(1+\alpha_{z_{n-1}})^{x_{n-1}+1+j}} P(\varepsilon_n(z_{n-1},z_n) = x_n - 1 - j)}{P(X_n = x_n|z_n,x_{n-1},z_{n-1})} \\
&= \frac{x_{n-1}\alpha_{z_{n-1}}}{P(X_n = x_n|Z_n = z_n,X_{n-1} = x_{n-1},Z_{n-1} = z_{n-1})} \\
&\cdot P(X_n = x_n - 1|Z_n = z_n,X_{n-1} = x_{n-1} + 1,Z_{n-1} = z_{n-1}),
\end{aligned}
$$

■

Equation (13)

Proof. For simplicity, we will introduce the following notation $P(A = a|Z_n = z_n,X_{n-1} = x_{n-1},Z_{n-1} = z_{n-1}) = P(A = a|z_n,x_{n-1},z_{n-1})$. Now we have

$$
\begin{aligned}
&E(\varepsilon_n(z_{n-1},z_n)|\mathscr{F}_n) = \sum_{i=0}^{x_n} i \cdot f(x_n - i) = \sum_{i=0}^{x_n} (x_n - i) \cdot f(i) \\
&= \frac{1}{P(X_n = x_n|z_n,x_{n-z},z_{n-1})} \\
&\cdot \sum_{i=0}^{x_n} (x_n - i) P(\varepsilon_n(z_{n-1},z_n) = x_n - i|z_n,x_{n-1},z_{n-1}) \cdot P(\alpha * X_{n-1}(Z_{n-1}) = i|z_n,x_{n-1},z_{n-1}) \\
&= \frac{1}{P(X_n = x_n|z_n,x_{n-z},z_{n-1})}
\end{aligned}
$$

$$\cdot \left[x_n \sum_{i=0}^{x_n} P(\varepsilon_n(z_{n-1}, z_n) = x_n - i | z_n, x_{n-1}, z_{n-1}) \cdot P(\alpha * X_{n-1}(Z_{n-1}) = i | z_n, x_{n-1}, z_{n-1}) \right.$$
$$\left. - \sum_{i=0}^{x_n} i \cdot P(\varepsilon_n(z_{n-1}, z_n) = x_n - i | z_n, x_{n-1}, z_{n-1}) \cdot P(\alpha * X_{n-1}(Z_{n-1}) = i | z_n, x_{n-1}, z_{n-1}) \right]$$
$$= \frac{1}{P(X_n = x_n | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})}$$
$$\cdot [x_n P(X_n = x_n | Z_n = z_n, X_{n-1} = x_{n-1}, Z_{n-1} = z_{n-1})$$
$$- x_{n-1} \alpha_{z_{n-1}} P(X_n = x_n - 1 | Z_n = z_n, X_{n-1} = x_{n-1} + 1, Z_{n-1} = z_{n-1})],$$

where the second term inside the brackets is derived in the same way as equation (12). ■

Theorem 1 from Laketa et al. (2018)

Let $\{X_n(z_n)\}$ be the RrNGINARmax$(\mathscr{M}, \mathscr{A}, \mathscr{P})$ or the RrNGINAR$_1(\mathscr{M}, \mathscr{A}, \mathscr{P})$ process. Let us suppose that $z_n = j$ and $z_{n-1} = i$ for some i and $j \in E_r$. If $0 \leq \alpha_j \leq \frac{\mu_j}{1+\max_{k \in E_r} \mu_k}$, then the distribution of the random variable $\varepsilon_n(i, j)$ can be written as a mixture of two geometric distributed random variables with means μ_j and α_j as follows

$$\varepsilon_n(i, j) \stackrel{d}{=} \begin{cases} Geom\left(\frac{\mu_j}{1+\mu_j}\right), & w.p.\ 1 - \frac{\alpha_j \mu_i}{\mu_j - \alpha_j}, \\ Geom\left(\frac{\alpha_j}{1+\alpha_j}\right), & w.p.\ \frac{\alpha_j \mu_i}{\mu_j - \alpha_j}. \end{cases} \tag{15}$$

Acknowledgements

The authors are grateful to Professor Miroslav Ristić for valuable suggestions. Also, the authors highly appreciate all the comments and instructions from the reviewers of the SORT journal, which significantly improved the quality of the manuscript. The first author acknowledges the grant of MNTR 174026 and the second and the third author acknowledge the grant of MNTR 174013 for carrying out this research.

References

Al-Osh, M.A. and Aly, E.E.A.A. (1992). First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics - Theory and Methods*, 21, 2483–2492.

Al-Osh, M.A. and Alzaid, A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8, 261–275.

Aly, E.E.A.A. and Bouzar, N. (1994). On some integer-valued autoregressive moving average models. *Journal of Multivariate Analysis*, 50, 132–151.

Alzaid, A.A. and Al-Osh, M.A. (1993). Some autoregressive moving average processes with generalized Poisson marginal distributions. *Annals of the Institute of Statistical Mathematics*, 45, 223–232.

Chakraborty, S. and Bhati, D. (2016). Transmuted geometric distribution with applications in modeling and regression analysis of count data. *SORT*, 40, 153–176.

Chakraborty, S. and Bhati, D. (2017). Corrigendum to "Transmuted geometric distribution with applications in modeling and regression analysis of count data". *SORT*, 41, 117-118.

Fernández-Fontelo, A., Cabana, A., Puig, P. and Morina, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine*, 35, 4875–4890.

Fernández-Fontelo, A., Fontdecaba, S. and Puig, P. (2017). Integer-valued AR processes with Hermite innovations and time-varying parameters: An application to bovine fallen stock surveillance at a local scale. *Statistical Modelling*, 17, 1–24.

Fokianos, K. (2011). Some recent progress in count time series. *Statistics*, 45, 49–58.

Freeland, R.K. and McCabe B.P.M. (2004). Analysis of low count time series data by poisson autoregression. *Journal of Time Series Analysis*, 25, 701–722.

Jazi, M.A., Jones, G. and Lai, C.D. (2012a). First-order integer-valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis*, 33, 954–963.

Jazi, M.A., Jones, G. and Lai, C.D. (2012b). Integer-valued AR with geometric innovations. *Journal of Iranian Statistical Society*, 2, 173–190.

Latour, A. (1998). Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis*, 19, 439–455.

Laketa, P.N., Nastić, A.S. and Ristić, M.M. (2018). Generalized random environment INAR models of higher order. *Mediterranean Journal of Mathematics*, 15, 1–9.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin*, 21, 645–650.

Moriña, D., Puig, P., Ríos, J., Vilella, A. and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, 30, 3125–3136.

Nastić, A.S., Laketa, P.N. and Ristić, M.M. (2016). Random environment integer-valued autoregressive process. *Journal of Time Series Analysis*, 37, 267–287.

Nastić, A.S., Laketa, P.N. and Ristić, M.M. (2019). Random environment INAR models of higher order. *REVSTAT-Statistical Journal*, 17, 35–65.

Pedeli, X. and Karlis, D. (2011). A bivariate INAR(1) process with application. *Statistical Modelling*, 11, 325–349.

Popović, P.M., Ristić, M.M. and Nastić, A.S. (2016). A geometric bivariate time series with different marginal parameters. *Statistical Papers*, 57, 731–753.

Popović, P.M., Nastić, A.S. and Ristić, M.M. (2018). Residual analysis with bivariate INAR(1) models. *Revstat*, 16, 349–363.

Qi, X., Li, Q. and Zhu, F. (2019). Modeling time series of count with excess zeros and ones based on INAR(1) model with zero-and-one inflated Poisson innovations. *Journal of Computational and Applied Mathematics*, 346, 572–590.

Ristić, M.M., Bakouch, H.S. and Nastić, A.S. (2009). A new geometric first-order integer-valued autoregressive (NGINAR(1)) process. *Journal of Statistical Planning and Inference*, 139, 2218–2226.

Ristić, M.M., Nastić, A.S., Jayakumar, K. and Bakouch, H.S. (2012). A bivariate INAR(1) time series model with geometric marginals. *Applied Mathematics Letters*, 25, 481–485.

Ristić, M.M., Nastić, S.A. and Miletić Ilić, V.A. (2013). A geometric time series model with dependent Bernoulli counting series. *Journal of Time Series Analysis*, 34, 466–476.

Rydén, J. (2017). Statistical modeling of warm-spell duration series using hurdle models. *SORT*, 41, 177–188.

Tang, M. and Wang, Y. (2014). Asymptotic behavior of random coefficient INAR model under random environment defined by difference equation. *Advances in Difference Equations*, 2014, 1–9.

Weiß, C.H. (2015). A Poisson INAR(1) model with serially dependent innovations. *Metrika*, 78, 829–851.

Weiß, C.H., Homburg, A. and Puig, P. (2019). Testing for zero inflation and overdispersion in INAR(1) models. *Statistical Papers*, 60, 823–848.

Zheng, H., Basawa, I.V. and Datta, S. (2006). Inference for *p*th-order random coefficient integer-valued autoregressive processes. *J. Time Ser. Anal.*, 27, 411–440.

Zheng, H., Basawa, I.V. and Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference*, 137, 212–229.

Zhu, R. and Joe, H. (2010). Negative binomial time series model based on expectation thinning operators. *Journal of Statistical Planning and Inference*, 140, 1974–1888.

Zhu, R. and Joe, H. (2006). Modelling count data time series with Markov processes based on binomial thinning. *Journal of Time Series Analysis*, 27, 725–738.