SORT

Statistics and Operations Research Transactions

Volume
48

Number 2, July-December 2024

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 48, Number 2, July-December 2024

eISSN: 2013-8830

Articles

Information for authors

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) —formerly Qüestiió— is an international journal launched in 2003 and distributed in printed form as well as in digital form online. From 2024 it will be published in digital form only. It is published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society, the Catalan Statistical Society and the Departament de Recerca i Universitats, of the Generalitat de Catalunya. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Editor in Chief

David V. Conesa, *Universitat de València, Dept. d'Estadística i Investigació Operativa*

Executive Editors

Michela Cameletti, *Università degli Studi di Bergamo, Dipt. di Scienze Economiche*
Esteve Codina, *Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa*
María L. Durbán, *Universidad Carlos III de Madrid, Depto. de Estadística y Econometría*
Guadalupe Gómez, *Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa*
Montserrat Guillén, *Universitat de Barcelona, Dept. d'Econometria, Estadística i Economia Espanyola (Past Editor in Chief 2007-2014)*
Enric Ripoll*, Institut d'Estadística de Catalunya*

Production Editor

Michael Greenacre, *Universitat Pompeu Fabra, Dept. d'Economia i Empresa*

Layout manager

Mercè Aicart

Responsible for the Secretary of SORT

Elisabet Aznar, *Institut d'Estadística de Catalunya*

Editorial Advisory Committee

Carmen Armero	*Universitat de València, Dept. d'Estadística i Investigació Operativa*
Eduard Bonet	*ESADE-Universitat Ramon Llull, Dept. de Mètodes Quantitatius*
Carles M. Cuadras	*Universitat de Barcelona, Dept. d'Estadística (Past Editor in Chief 2003–2006)*
Pedro Delicado	*Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa*
Paul Eilers	*Erasmus University Medical Center*
Laureano F. Escudero	*Universidad Miguel Hernández, Centro de Investigación Operativa*
Elena Fernández	*Universidad de Cádiz, Depto. de Estadística e Investigación Operativa*
Ubaldo G. Palomares	*Universidad Simón Bolívar, Dpto. de Procesos y Sistemas*
Jaume García	*Universitat Pompeu Fabra, Dept. d'Economia i Empresa*
Montserrat Herrador	*Instituto Nacional de Estadística*
Maria Jolis	*Universitat Autònoma de Barcelona, Dept. de Matemàtiques*
Pierre Joly	*Conseil d'Analyse Economique*
Ludovic Lebart	*Centre Nationale de la Recherche Scientifique*
Richard Lockhart	*Simon Fraser University, Dept. of Statistics & Actuarial Science*
Glòria Mateu	*Universitat de Girona, Dept. d'Informàtica, Matemàtica Aplicada i Estadística*
Geert Molenberghs	*Leuven Biostatistics and Statistical Bioinformatics Centre*
Eulalia Nualart	*Universitat Pompeu Fabra, Dept. d'Economia i Empresa*
Josep M. Oller	*Universitat de Barcelona, Dept. d'Estadística*
Maribel Ortego	*Universitat Politècnica de Catalunya, Dept. d'Enginyeria Civil i Ambiental*
Javier Prieto	*Universidad Carlos III de Madrid, Dpto. de Estadística y Econometría*
Pere Puig	*Universitat Autònoma de Barcelona, Dept. de Matemàtiques (Past Editor in Chief 2015-2020)*
José María Sarabia	*Universidad de Cantabria, Dpto. de Economía*
Albert Satorra	*Universitat Pompeu Fabra, Dept. d'Economia i Empresa*
Albert Sorribas	*Universitat de Lleida, Dept. de Ciències Mèdiques Bàsiques*
Vladimir Zaiats	*Universitat Pompeu Fabra, Dept. d'Economia i Empresa*

Institut d'Estadística de Catalunya

The mission of the Statistical Institute of Catalonia (Idescat) is to provide high-quality and relevant statistical information, with professional independence, and to coordinate the Statistical System of Catalonia, with the aim of contributing to the decision making, research and improvement to public policies.

Management Committee

President
Xavier Cuadras Morató — *Director of the Statistical Institute of Catalonia*

Secretary
Cristina Rovira — *Deputy Director General of Production and Coordination*

Editor in Chief
David V. Conesa — *Universitat de València, Dept. d'Estadística i Investigació Operativa*

Representatives of the Statistical Institute of Catalonia
Cristina Rovira — *Deputy Director General of Production and Coordination*
Josep Maria Martínez — *Head of Department of Standards and Quality*
Josep Sort — *Deputy Director General of Information and Communication*
Elisabet Aznar — *Responsible for the Secretary of SORT*

Representative of the Universitat Politècnica de Catalunya
Guadalupe Gómez — *Department of Statistics and Operational Research*

Representative of the Universitat de Barcelona
Jordi Suriñach — *Department of Econometrics, Statistics and Spanish Economy*

Representative of the Universitat de Girona
Javier Palarea-Albaladejo — *Department of Informatics, Applied Mathematics and Statistics*

Representative of the Universitat Autònoma de Barcelona
Xavier Bardina — *Department of Mathematics*

Representative of the Universitat Pompeu Fabra
David Rossell — *Department of Economics and Business*

Representative of the Universitat de Lleida
Albert Sorribas — *Department of Basic Medical Sciences*

Representative of the Universitat Rovira i Virgili
Josep Domingo-Ferrer — *Department of Computer Engineering and Maths*

Representative of the Catalan Statistical Society
Núria Pérez — *Fight Against AIDS Foundation*

Secretary

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

© Institut d'Estadística de Catalunya
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
www.idescat.cat/sort/

eISSN: 2013-8830
SORT 48 (2) July-December (2024)

2024 CALIDAD REVISTAS CIENTÍFICAS ESPAÑOLAS
FECYT-073/2024
Fecha de certificación: 20 de mayo de 2011 (2ª convocatoria)
Válido hasta: 24 de julio de 2025

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística
Departament de Recerca i Universitats

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 48
Number 2
July-December 2024
eISSN: 2013-8830

Articles

SORT 48 (2) July-December 2024, 155-182 DOI: 10.57645/20.8080.02.17

Patient-reported outcomes and survival analysis of chronic obstructive pulmonary disease patients: a two-stage joint modelling approach

Cristina Galán-Arcicollar[1,2], Josu Najera-Zuloaga[2] and Dae-Jin Lee[1,3]

Abstract

Joint modelling has gained attention in longitudinal studies incorporating biomarkers and survival data. In the context of chronic diseases, patient evolution is often tracked through multiple assessments, with patient-reported outcomes playing a crucial role. The Beta-Binomial distribution is suggested as a suitable model for these longitudinal variables. However, its integration into joint modelling remains unexplored. This study introduces an estimation procedure for analyzing longitudinal patient-reported outcomes and survival data together. We compare different estimation approaches through simulation experiments, including the proposed model. Furthermore, the methodologies are applied to real data from a follow-up study on chronic obstructive pulmonary disease patients.

MSC: *62N02, 62N03.*

Keywords: *Joint modelling, Beta-Binomial regression, Patient-reported outcomes, Survival analysis.*

1. Introduction

In recent years, there has been an increased focus on placing patients at the centre of health care research and evaluating clinical care (Weldring and Smith, 2013). For instance, patient-reported outcomes (PROs) are helpful tools that provide information on the patient's health status considering their health, quality of life, or functional status associated with the health care or treatment they received (Weldring and Smith, 2013). PROs have gradually become a significant source of information as they evaluate a wide

[1] Applied Statistics Research Line, Basque Center for Applied Mathematics, Bizkaia, Spain.
[2] Department of Mathematics, University of the Basque Country UPV/EHU, Bizkaia, Spain.
[3] School of Science and Technology, IE University, Madrid, Spain.

Received: June 2023
Accepted: December 2023

range of outcomes, such as pain, fatigue or vitality. Its use is strongly recommended in combination with clinical indicators to provide a more comprehensive patient assessment, especially in chronic illnesses (Speight and Barendse, 2010). This information is characterized by coming directly from the patient, without interpretation of the patient's response by a clinician or anyone else (US Department of Health and Human Services, 2006). The PRO measurements are usually collected by supplying validated questionnaires to patients. Thus, PROs are often built as a sum of responses to several items of the questionnaire, so they can be considered as discrete and bounded random variables. It has been shown in the literature that due to subject-specific characteristics, PROs are usually overdispersed (i.e., the mean-variance relationship fails mainly due to presence of an unexpected source of variation). In this context, the Beta-Binomial distribution has been proposed in the literature as an adequate distribution to fit overdispersed discrete and bounded outcomes, particularly in PRO analysis (Arostegui, Núñez-Antón and Quintana, 2007).

Most clinical studies that consider health-related quality of life (HRQoL) involve the follow-up of patients where longitudinal data are collected to evaluate patient worsening or changes in the health-status over time. Typically, it is also considered survival or time-to-event analysis during follow-up studies, when there is an event of interest such as death or illness relapse/recovery. The primary purpose of this article is to propose a methodology that allows answering the question of whether an association exists between the survival data and the serial measurements of HRQoL. In this framework, several studies have analyzed the impact of HRQoL on mortality (or survival) under a cross-sectional approach, as in Domingo-Salvany et al. (2002), or simply by fitting a model for the vital status instead for the survival times, such as Esteban et al. (2022). Other works have focused on the association of HRQoL with clinical measurements such as the number of hospitalizations or body mass index (BMI) in order to assess the evolution over time without considering survival data, like in Esteban et al. (2020). However, it is well known that when interest relies on the two outcomes (i.e., the longitudinal evolution and time-to-event), separate analyses are not the best modelling options because they do not consider the dependencies between them.

The statistical literature uses the term "joint modelling" to refer to those methods that simultaneously analyze longitudinal measurements and time-to-event outcomes. Nowadays, joint models of longitudinal and survival data have received much attention in the literature, particularly in medical studies, where these data frequently arise together in practice (Wu et al., 2012). The classical approach to joint modelling consists of a full likelihood formulation that integrates the two fitted models for longitudinal and survival to jointly estimate the parameter set. Regardless joint model's popularity, a complete analysis that includes longitudinal discrete and bounded outcome has not been thoroughly studied. Moreover, when these studies are carried out, the nature of data is not usually considered, where due to computational complexities, linearity or some kind of data transformation is assumed to simplify the computations, for example in Ibrahim, Chu and Chen (2010); Wu et al. (2012); Li, Tosteson and Bakitas (2013). In addition,

two-stage joint models consisting of estimating the longitudinal submodel and then plug the shared information into the survival submodel has been proposed as a simple solution in Self and Pawitan (1992). Although they are mainly known for their biased results when compared to fully likelihood methods (Wu et al., 2012), they are a well-known and a flexible methodology in the joint modelling framework. Bayesian methods have also been proposed to extend joint models to generalized linear mixed models, although it is desirable to check if the final results are sensitive to the choices of prior distributions (Armero, 2021).

In this work, although we do not overcome the possible bias of the two-stage methodology, we argue and show that it is preferable to consider a proper distribution for the longitudinal outcome rather than assuming a linear mixed model (i.e., a Gaussian response) when performing a joint model with a direct focus on PROs. Particularly, we aim to assess the effect of the longitudinal PRO measurements and the time until patient's death occurs. However, for the case of PROs, the fact that the Beta-Binomial distribution does not belong to the exponential family of distributions makes its inclusion into the joint modelling framework not straightforward. Our proposal provides an easy way to account for the evolution of PRO questionnaires of patients and the estimation of the survival probabilities by means of a joint modelling which includes the Beta-Binomial distribution for the analysis of longitudinal discrete and bounded data with overdispersion. A two-stage approach is considered in this work, where the first step consists of fitting a longitudinal Beta-Binomial mixed-effects model that estimates the impact of observed covariates and the subjects' evolution through time by following the approach described in Najera-Zuloaga, Lee and Arostegui (2019). Then, in a second step, the estimated linear predictors are included into a Cox proportional hazards regression model. We developed an unified methodology that couple these two models and show through simulation studies the better performance among other alternatives. Finally, in the suplementary material, we provide the R code to implement our approach using the functions `BBmm` in the R package `PROreg` (Najera-Zuloaga, Lee and Arostegui, 2022) for the longitudinal modelling of the PROs and `coxph` function of the `survival` R package (Therneau and Grambsch, 2000).

This research was motivated by an analysis of the health-status of patients with chronic obstructive pulmonary disease (COPD). Researchers at the Respiratory Service at Galdakao Hospital in Biscay (northern Spain) designed the COPD study, which was a longitudinal clinical trial that recorded measurements of the health status and evolution of patients being treated for COPD, see Esteban et al. (2020) for further details. One of the main objectives of the study was to measure the relationship between HRQoL and mortality of COPD patients. The hypothesis to assess this relationship is based on the fact that COPD is not only an airway obstruction disease, it is a complex, heterogeneous and multisystem disease (Vanfleteren et al., 2016) whose overall impact on individuals is many-sided. Thus, its severity is not fully captured by clinical parameters, it often needs to be supplemented by other indicators from a patient's perspective, such as those associated with HRQoL. This work is intended to provide clinicians and researchers on PROs

the statistical tools for modelling the HRQoL evolution of patients and its survival probabilities in an unified framework and discuss on joint modelling approaches in this context.

The paper is organized as follows. In Section 2, we introduce the details of our COPD study data set. Section 3 is dedicated to present the modelling of HRQoL evolution over time. In Section 4, we present two popular methodologies that incorporate longitudinal measurements in survival analysis and we provide a new approach for PRO data. The COPD data analysis with the presented methodologies is supplied in Section 5, with clinical interpretation of relevant results. In Section 6, we perform a simulation study based on the COPD data that compares the detailed methodologies in Section 4, including our proposal. Finally, in Section 7 it is given some conclusions and discussion.

2. Motivation study

Chronic obstructive pulmonary disease (COPD) is a major cause of chronic morbidity and mortality worldwide (Pauwels and Rabe, 2004). COPD is a respiratory system disease with irreversible damage which causes physiological discomfort and psychosocial impact on individuals, thus it is also associated with high level of disability. Although medical assessment of COPD mainly involves clinical biomarker measurements, the overall impact of COPD on individuals is multifaceted and it is not completely reflected by them. Thus, the assessment of the PROs and HRQoL is considered part of the standard care in the treatment of COPD (Wiklund, 2004). There is good evidence that COPD exacerbations can have a large and sustained impact on patients' symptoms and health status (Jones and Higenbottam, 2007). For this reason, tools such as PROs are needed to evaluate all different aspects of the disease, as they supplement clinical biomarkers by other kind of indicators from a patients perspective.

Our COPD study is an observational study that was designed at the Respiratory Unit at Galdakao Hospital in Biscay, Spain. A sample of 543 patients were consecutively included during the first year and in the second half of the study. The study is conducted in five years follow-up period for each patient with a maximum of four clinical examination and interviews per patient. Thus, the number of measurements per patient ranges between one and four. Figure 1 summarize this feature taking into account patients division according to the occurrence of event.

Notice that most patients had the maximum number of measurements but we also find that there are only one measurement recorded for some of them, being an unbalanced longitudinal data set. Moreover, patients' entry time, where baseline (first) measurement was recorded, did not take place at the same time due to consecutively entry of patients. In addition, measurement times were unequally space intervals because we found that the second measurement is one year apart from the fist one, as well as the third measurement from the second one, but the fourth measurement is three years apart from the third one.

The health-status in the COPD study was measured with both, generic and disease-specific questionnaires, named respectively Short Form-36 Health Survey (SF-36) and

St. George's Respiratory Questionnaire (SGRQ). Questionnaires often provide information about different health aspects, thus they are usually divided in dimensions according to the information referred. Particularly, SF-36 was constructed to represent eight health dimensions, which are the physical functioning (PF), role physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role emotional (RE), and mental health (MH). The first four dimensions are mainly physical, whereas the last four measure mental aspects of health. Each dimension scale score range from 0 to 100, where a higher score indicates a better health status. This standardized scoring system is detailed in Ware et al. (1993). On the other hand, SGRQ consist of three dimensions namely, symptoms (SYMP), impact (IMP) and activity (ACT) where higher scores refer to worse health status (Jones, Quirk and Baveystock, 1991). Each of the three dimensions of the questionnaire is separately standardized in the range 0 to 100.

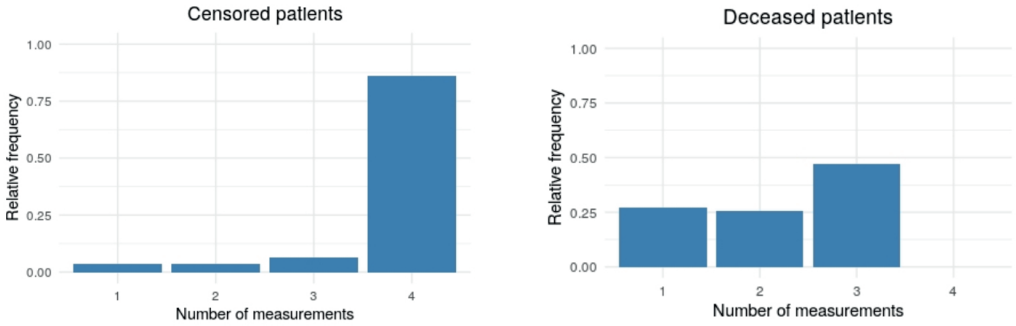

Figure 1. *Relative frequencies of patient's measurements. Left-side picture represents relative frequency of patients with no event while right-side represents relative frequency of patients with presence of the event during the study.*

In this work, we considered a re-scaled form of the patients' scores to interpret them in a binomial form, which makes its interpretation easier. The SF-36 re-scaling is motivated by the work of Arostegui, Núñez-Antón and Quintana (2013) and SGRQ is based on the idea that a 4-points change in the 0-100 scale is considered a clinically significant change (Jones, 2005). Based on this re-scaling, patients' average scores are shown in Table 1 according to the maximum score of each dimension.

In our COPD study, survival data was also collected and the time-to-event variable considered was the patient's date of death. Survival times are frequently influenced by right censoring in which the event time differs from the observed one. Particularly, in the COPD study administrative censoring was applied, which occurs when the study observational period ends without the presence of the event, i.e, when the patient has four measurements. Furthermore, we discovered censorship as a result of withdrawals, also known as loss to follow-up. During the study, 167 events were recorded, so 376 patients were censored. Because 324 patients completed the study's follow-up with no event, its censoring was administrative. Thus, 52 patients were censored as a result of the withdrawal process, corresponding approximately to a 10% of random censoring.

Table 1. *HRQoL scores of each test dimension are presented as mean (sd) for each year of the study. Number of patients at each year is also shown.*

		HRQoL mean measurements			
Test	**Dimension**	**Baseline**	**1 year**	**2 years**	**5 years**
	(max score)	n=543	n=484	n=425	n=320
SF-36	PF (20)	11.55 (4.88)	11.63 (4.99)	11.56 (4.94)	11.29 (4.99)
	RP (4)	2.62 (1.56)	2.44 (1.60)	2.51 (1.60)	2.22 (1.65)
	BP (9)	6.59 (2.58)	6.28 (2.71)	6.43 (2.66)	6.37 (2.57)
	GH (20)	8.81 (4.27)	8.54 (4.55)	8.32 (4.38)	8.22 (4.07)
	VT (20)	11.87 (4.99)	11.65 (4.80)	11.93 (4.68)	11.52 (4.78)
	SF (8)	6.53 (1.96)	6.39 (2.07)	6.57 (1.93)	6.23 (2.09)
	RE (3)	2.41 (1.08)	2.22 (1.18)	2.30 (1.14)	2.11 (1.23)
	MH (13)	9.47 (2.91)	9.42 (2.78)	9.44 (2.80)	9.25 (2.94)
SGRQ	SYMP (24)	10.63 (5.54)	10.11 (5.60)	10.35 (5.81)	10.52 (5.83)
	IMP (24)	7.52 (5.21)	7.11 (5.29)	7.06 (5.08)	7.12 (5.21)
	ACT (24)	11.63 (6.19)	10.95 (6.19)	11.21 (6.14)	11.34 (6.31)

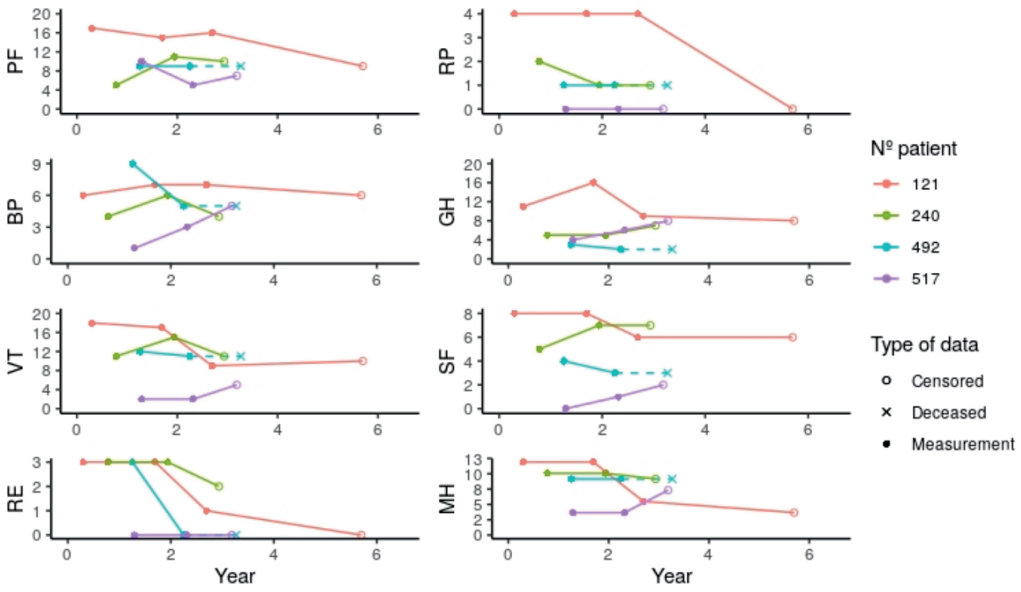

Figure 2. *SF-36 longitudinal measurements according to the survival data recorded of four COPD patients. The straight dashed line indicates the extrapolation of the patient's score from the last recorded measurement to the observed date of death.*

Figure 2 and Figure 3 illustrate the longitudinal measurements of the eight SF-36 dimensions and the three SGRQ dimensions respectively of four patients of the database.

According to the survival data recorded, we indicated the dates of death with a cross symbol and joined with the previous measurements with a dashed line because at patient's date of death no HRQoL data was recorded. This is an example of the different types of data recorded according to the two kinds of censorship and the event recorded. For instance, patients 240 and 517 only recorded 3 measurements and then no more information was collected, so they are censored due to loss of follow-up, while patient 121 completed the four measurements and is administratively censored.

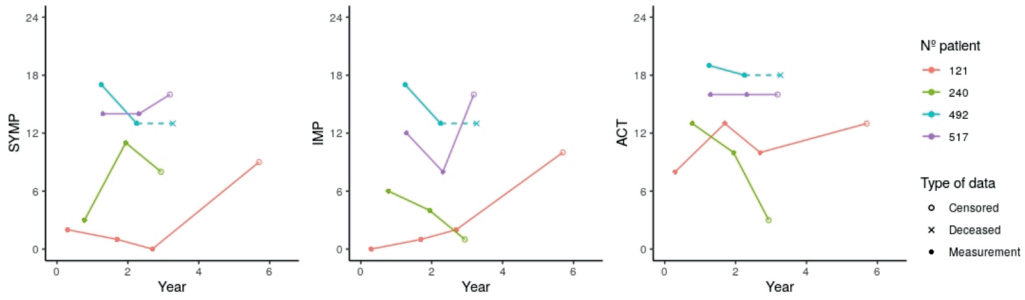

Figure 3. *SGRQ longitudinal measurements according to the survival data recorded of four COPD patients. We used the straight dashed line to joint the date of death with the recorded measurements.*

3. Modelling the evolution of patients' HRQoL

This article is focused on self-reported outcomes, which usually have discrete and bounded distributions. A particular case of self-reported outcomes in medicine are the patient-reported outcomes (PROs). It is known that this kind of measurements usually lead to floor or ceiling effect, i.e, typically accumulate values in one or both edges of the score scale due to subject-specific characteristics, leading to different shapes such as U, J or inverse J-shaped (Najera-Zuloaga, Lee and Arostegui, 2018). The usual exponential family distributions are not able to fit them properly and particularly, the normality assumption for the longitudinal variable in classic joint models will not be frequently satisfied for this kind longitudinal data. The Beta-Binomial distribution has been proposed in literature as the proper distribution to analyze discrete and bounded outcomes with overdispersion (Arostegui et al., 2007). Generally, the Beta-Binomial distribution is defined as a mixture of the Binomial and Beta distributions. It consists of a finite sum of Bernoulli variables whose probability parameter is random and follows a Beta distribution. The Beta-Binomial distribution preserves the characteristics of the Binomial distribution which suits the nature of discrete and bounded data, but it also displays the flexibility of the Beta distribution. We denote that variable Y follows a Beta-Binomial distribution as $Y \sim BB(m, p, \phi)$ with parameters m, p and ϕ, where parameter m makes reference to the maximum number of trials, p is the probability parameter and ϕ the correlation/dispersion parameter. The density function of the Beta-Binomial distribution does not belong to the exponential family distributions but it has a closed-form equation,

see Arostegui et al. (2007) for further details. The article Najera-Zuloaga et al. (2018) proposes a marginal regression model with the Beta-Binomial distribution for measuring the effect of explanatory variables on discrete and bounded response variables. This Beta-Binomial regression is performed by connecting the probability parameter to a vector of regression parameters by means of a logit link function model.

PROs are usually measured in a longitudinal framework in which individuals are followed up for a certain period. The extension of Beta-Binomial regression models to the longitudinal framework is performed in terms of mixed-effects that associate all measurements for the same subject. Beta-Binomial mixed-effect model (BBMM) includes random effects into the linear predictor to accommodate the dependency of repeated measurements. As in mixed model methodology, conditioned on the random effects, the repeated measurements are independent and drawn from a Beta-Binomial distribution. Let $y_i = (y_{i1}, ..., y_{in_i})$ be the repeated measurements for subject i, then $y_{ij}|u_i \sim BB(m_i, p_{ij}, \phi)$, $\forall j \in \{1, ..., n_i\}$, and u_i its corresponding vector of random effects $u_i \sim N(0, D)$ that describe the subject-specific characteristics. The probability parameter of the Beta-Binomial distribution is also linked with logit link function to the fixed and random parameters such as:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = x_{ij}\beta + z_{ij}u_i = \eta_{ij}, \quad i = 1, ..., n_i, \tag{1}$$

where x_{ij} and z_{ij} are the fixed and random covariates respectively for the jth measurement of subject i with β and u_i the corresponding effect parameters. See Najera-Zuloaga et al. (2019) for further details.

4. Relating longitudinal patient reported outcomes and survival analysis

During follow-up studies in clinical trials, it is of particular interest to collect several biomarker measurements as well as time-to-event outcomes, such as death, illness relapse, recovery or development of some disease (Papageorgiou et al., 2019). Thus, longitudinal data and survival data frequently appear together in practice, and they are often associated in some ways (Wu et al., 2012). The longitudinal measurement of biomarkers is useful for characterizing the occurrence of an outcome of interest because they can predict treatment outcomes or be related to the event process and prognosis (Arisido et al., 2019). It is then of particular interest to evaluate and investigate its relationship (Ibrahim et al., 2010). In this section, we present two well-known methodologies for estimating the relationship between these outcomes, as well as define our proposed approach.

4.1. A Time-Varying Cox-Model

The first proposed approach to analyze the relationship between longitudinal and survival outcomes consisted of extending the classical proportional hazards model (Ri-

zopoulos, 2012). The main objective was to allow the inclusion of time-dependent covariates into this known survival model. This methodology is usually referred as Time-Varying Cox-Model (TVCM). The risk function for patient i is defined quite similar to the original Cox model and it can be written as follow:

$$h_i(t|\mathcal{Y}_i(t), w_i) = h_0(t)R_i(t)\exp\{\gamma w_i + \alpha y_i(t)\}, \tag{2}$$

where $h_i(t)$ and $h_0(t)$ respectively denote the subject and baseline risk function at each time t, $\mathcal{Y}_i(t) = \{y_i(s) : s \leq t\}$ is the history of longitudinal covariate up to time t and w_i refers to the vector of baseline covariates for patient i. Parameters γ and α measure the impact of baseline and longitudinal covariates respectively into the risk function. Finally, it is included a risk indicator function $R_i(\cdot)$ with $R_i(t) = 1$ if subject i is at risk at time t, and $R_i(t) = 0$ otherwise.

TVCM is known for being a flexible semi-parametric methodology for fitting survival analysis because parameters estimation is based on proportional hazards. Thus, partial likelihood is used to perform parameters estimation, where baseline risk function is left unspecified. However, it is assumed in TVCM that time-dependent covariates are predictable processes, measured without error, and have their complete path fully specified (Rizopoulos, 2012). Notice that the whole longitudinal history is not available because information is recorded only at some measurement times. In order to overcome this issue, TVCM is based on the so-called 'last value carried forward' (LVCF) approach, where the marker values are considered constant between measurement points. In Figure 4, we considered the PF and MH dimensions of SF-36 COPD data shown in Figure 2 and we showed a graphic interpretation of the stepwise approach of the TVCM methodology. We can observe that the longitudinal trends are not considered in this method.

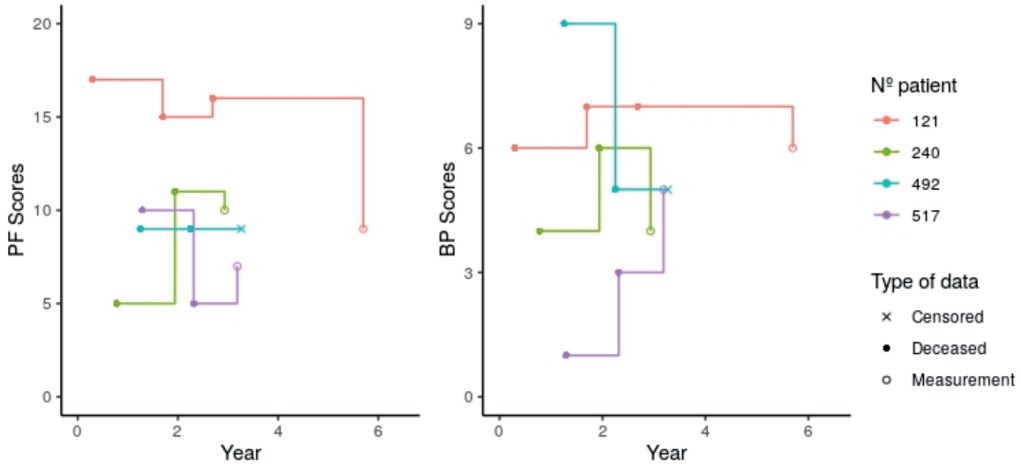

Figure 4. *Graphic interpretation of the last value carried forward used in Time-Varying Cox Model.*

Moreover, it is important to distinguish between the different kinds of time-varying covariates because it is known that TVCM is not appropriate when time-dependent covariates are of endogenous nature (Rizopoulos, 2012). Endogenous covariates, such as biomarkers, typically require the survival of the subject for their existence and thus, they satisfy $S_i(t|\mathcal{Y}_i(t)) = Pr(T_i^* > t|\mathcal{Y}_i(t)) = 1$, where $S_i(\cdot)$ and T_i^* make reference to the survival function and the random time-to-event variable for subject i respectively. Then, for instance, if the date of death for the subject is s, the covariate will not be recorded at each time $t \geq s$. This fact of endogenous covariates is not taken into account in TVCM. In addition, despite TVCM flexibility, several studies have performed a variety of simulated scenarios to show that it can lead to really biased results due to all assumptions mentioned above, see for instance Arisido et al. (2019).

The implementation of this methodology in R software makes used of displayed data sets using the counting process format $(start, stop, status)$, holding the information on the longitudinal variable for a specific time interval. The start and stop times denote the limits of the time intervals during which measurements were recorded and the status indicate the ocurrence of the event at the end of the interval. The `tmerge` function aids in the creation of such data sets and the model is fitted in the `coxph` function, both functions from the `survival` package (Therneau and Grambsch, 2000).

4.2. Shared parameter joint modelling

In order to overcome TVCM limitations, an alternative modelling framework has been introduced in the literature, known as the joint modelling for longitudinal and time-to-event data (Faucett and Thomas, 1996). The motivating idea behind these joint models is to couple the survival model with a suitable model for the repeated endogenous measurements. The longitudinal model is usually fitted by using Linear Mixed Model (LMM) to describe covariates impact together with time evolution, and survival model is usually performed in terms of the Cox model, leading to a definition of risk function as follows:

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t)\exp\{\gamma w_i + \alpha m_i(t)\}. \tag{3}$$

It reminds quite similar to the subject's risk function defined in TVCM shown in Equation (2), but the values included in the survival model are not the observed ones $y_i(t)$, instead it is incorporated the term $m_i(t)$ that denotes the true and unobserved value. This true value is computed according to the longitudinal model assumed, thus:

$$m_i(t) = \mathbb{E}(y_i(t)|u_i) = x_i(t)\beta + z_i(t)u_i, \tag{4}$$

where $x_i(t)$ are the fixed parameters with β the corresponding fixed effect and $z_i(t)$ the random parameters with u_i the corresponding vector of random effects having a multivariate normal distribution with mean zero and covariance matrix D, i.e, $u_i \sim N(0, D)$. One of the most popular estimation method that has been proposed for the described joint model consisted of computing a full maximum likelihood of the observed outcomes. Both models, longitudinal and survival, are set and then, they are linked using a shared

latent structure, due to this feature joint models belong to the class of shared-parameter models. The key assumption of this methodology is full conditional independence, i.e, it is assumed that conditional to random effects, time-to-event and longitudinal outcomes are independent, as well as the different measurements for the same individual (Rizopoulos, 2012). Considering the classical survival analysis notation, let T_i be the recorded time-to-event and δ_i the event indicator, that takes value 1 if the observed time T_i equals to the event time T_i^* and 0 otherwise, the assumptions can be written as follow:

$$f_\theta(T_i, \delta_i, y_i|u_i) = f_\theta(T_i, \delta_i|u_i) f_\theta(y_i|u_i),$$

$$f_\theta(y_i|u_i) = \prod_{j=1}^{n_i} f_\theta(y_i(t_{ij})|u_i),$$

where θ denotes the set of all, survival and longitudinal parameters, i.e., $\theta = (\theta_t, \theta_y, \theta_u)$ and y_i denotes the vector for the n_i measurements for subject i taken at time points $\{t_{ij} : j = 1, ..., n_i\}$. Under these assumptions the log-likelihood contribution for the ith subject can be formulated in this way:

$$\begin{aligned}\log f_\theta(T_i, \delta_i, y_i) &= \log \int f_\theta(T_i, \delta_i, y_i|u_i) f_{\theta_u}(u_i) du_i \\ &= \log \int f_{\theta_t,\beta}(T_i, \delta_i|u_i) \big[\prod_j f_{\theta_y}(y_i(t_{ij})|u_i)\big] f_{\theta_u}(u_i) du_i.\end{aligned}$$

As it was mentioned, time-to-event outcome is usually fitted by Cox model and the density is included in the log-likelihood characterized by the hazard and survival functions:

$$\begin{aligned}f_{\theta_t}(T_i, \delta_i|u_i) &= h_i(T_i|\mathcal{M}_i(T_i))^{\delta_i} S_i(T_i|\mathcal{M}_i(T_i)) \\ &= \big[h_0(T_i)\exp\{\gamma w_i + \alpha m_i(T_i)\}\big]^{\delta_i} \times \exp\left(-\int_0^{T_i} h_0(s)\exp\{\gamma w_i + \alpha m_i(s)\} ds\right).\end{aligned}$$

Notice that for joint modelling, it is necessary to set a baseline hazard function. Weibull or exponential functions are frequently used as baseline hazard in survival analysis, but also it can be modeled flexibly using piecewise-constant or B-splines (Rosenberg, 1995) methods. Furthermore, it is necessary to set the longitudinal model to compute $m_i(\cdot)$ according to Equation (4) and also for the inclusion of the density function into the full likelihood, which is the normal density function. The literature also considered the use of generalized linear mixed models (GLMM) for longitudinal data into the joint modelling framework. This extension is straightforward (Rizopoulos, 2012) by choosing the density function as a member of the exponential family. The difficulty of this extension of joint models to incorporate GLMM is that computation becomes more demanding because of the nonlinearity of the longitudinal models (Wu et al., 2012). Moreover, this extension is not implemented in the available R software packages. Due to that fact, it is common to assume linearity measurements, sometimes ignoring the nature of data, being Joint Modelling with longitudinal normality assumption one of the

widely used methodology in the literature. Moreover, it is worth mentioning that the inclusion of several longitudinal responses into the model also leads to computational complexities, being this another limitation.

The implementation of this methodology can be easily performed by using `jointModel` function from `JM` R-package (Rizopoulos, 2010), it is needed first to fit separately the linear mixed-effects model with `lme` from `nlme` package (Pinheiro et al., 2013) and Cox model with `coxph` from `survival` package (Therneau and Grambsch, 2000). In this article, we denote this methodology as JM.

4.3. A two-stage approach for discrete bounded outcomes

In this subsection, we present our proposal approach where we incorporate the Beta-Binomial distribution into the joint modelling framework by performing a two-stage methodology. Our main goal is to emphasize the need of an adequate model that assesses the relationship between longitudinal and a time-to-event outcomes which provides an easy interpretation for PROs. With that aim we propose a joint model methodology that includes Beta-Binomial distribution for a suitable fit of this kind of longitudinal data, which relies on a better estimation of the association parameter between the longitudinal and the time-to-event outcome.

The first step consists of fitting a longitudinal Beta-Binomial mixed-effects model to evaluate the impact of some observed covariates and also its evolution over time including subject-specific effects that account for non observable characteristics that are different for each individual. Then, in a second step, the estimated linear predictor is computed according to the Beta-Binomial regression model and it is included in a Cox proportional hazards survival model as observed covariates.

For a sample of N individuals, let $y_i = (y_{i1}, ...y_{in_i})$ be the n_i measurements for subject i such that $y_{ij}|u_i \sim BB(m, p_{ij}, \phi)$, taken at measurement times $t_i = (t_{i1}, ...t_{in_i})\ \forall i = 1, ..., N$. Then, following BBMM approach the probability parameter and the linear predictor are linked by means of *logit* function as indicated in Equation (1), where p_{ij} denotes the probability parameter for patient i at time t_{ij}, i.e., $p_{ij} = p_i(t_{ij})$. Fixed covariates x_{ij}, can be time-dependent or baseline covariates and random effects $u_i \sim \mathcal{N}(0, D(\sigma))$ are assumed to follow a multivariate Normal with zero mean and variance-covariance nonsingular matrix D, which depends on a vector of variance parameters σ.

In this first step where BBMM is fitted, parameters β, u_i, σ and ϕ are estimated. The analysis of these parameters provides an assessment of the impact of fixed effects, heterogeneity among subjects and overdispersion of the longitudinal responses. Next, for the second step, the parameter estimations are used to compute the fitted values for the longitudinal outcome at each time t following the Beta-Binomial methodology. Let p_{it} denote the probability parameter of subject i at each time t and $\hat{\eta}_{it}$ its corresponding estimated linear predictor, then:

$$\hat{y}_i(t) = m\hat{p}_{it} = m \cdot \text{logit}^{-1}(\hat{\eta}_{it}) = \frac{m}{1 + \exp(-\hat{\eta}_{it})}. \tag{5}$$

Once the fitted values are computed, in the second step, they are inserted into the classic Cox model as if they were observed values. Thus, we proceed by fitting the Cox model as usual survival analysis defining the risk function as:

$$\lambda_i(t) = \lambda_0(t)\exp(\gamma w_i + \alpha \hat{y}_i(t)). \tag{6}$$

In this study, we focused on the main idea of estimating joint models using two-stage approaches, which avoid integral computation difficulties and also allow a greater flexibility when setting the longitudinal model. Additionally, this methodology avoids the transformation of the HRQoL questionnaire scores by means of square root, logarithmic or Box-Cox. These data transformation of the questionnaries' scores is usually performed to set linearity assumption, although this approach makes interpreting coefficients more difficult. The inclusion of BBMM allows the analysis of the association parameter with hazard ratio interpretation in terms of 1-point scoring, which is quite intuitive and natural instead of incorporating the standardized data or data transformations. Moreover, the performance of a BBMM longitudinal regression provides an easy way to interpret covariates' influence for 1-unit increment in terms of the odds-ratio, akin to hazard ratio analysis.

Our proposal can be easily implemented with the available R software. The first step is performed by using `BBmm` function from `PROreg` package (Najera-Zuloaga et al., 2022) and second step is fitted with `coxph` function from `survival` package (Therneau and Grambsch, 2000). See the R code provided in the supplementary material.

5. Application to COPD study

In the COPD study, one of the main objectives was to describe HRQoL and its evolution. This objective was studied in Esteban et al. (2020), where they analyzed the impact of sociodemographic variables and clinical indicators on the HRQoL, although only the SF-36 questionnaire was considered. Another key objective was to collect survival data based on time-to-event and the corresponding event indicators to assess HRQoL's relation to patient's risk of death. This objective was considered for SGRQ scores in Esteban et al. (2022). However, in this work they did not perform a survival analysis, because a logistic regression was fitted for vital status and only one-year periods were considered without including the time-to-event nature of the study. The original study aimed to investigate both, how HRQoL evolved and its relationship with the survival data collected. Nevertheless, these two outcomes have not been thoroughly studied jointly, where the inclusion of BBMM into the joint modelling framework may improve the results of the association parameter and provide an easier clinical interpretation. It is our goal to provide a complete analysis of both outcomes of interest for a more comprehensive evaluation. Apart from a joint analysis of both outcomes being preferable to a separate one, we also want to emphasize the importance of PRO data being adequately fitted.

The first step of our proposal can deal with one of the main objectives of the COPD study, which consists of measuring the evolution of patients' HRQoL scores over time

to determine trends. We performed a Beta-Binomial mixed-effect model by considering time as observable covariate to evaluate population and subject-specific evolution by including random intercept and slope as follow:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = (\beta_0 + u_{i0}) + (\beta_1 + u_{i1})\, t_{ij} = \eta_{ij}, \tag{7}$$

such that the vector of random effects satisfies $u_i = (u_{i0}, u_{i1}) \sim \mathcal{N}(0, D(\sigma))\ \forall i = 1, ..., 543$, with D a diagonal matrix with entries $\sigma = (\sigma_{u_0}, \sigma_{u_1})$. The components of vector u_i represent respectively baseline and subject-specific effects in the evolution of patient i. Results are shown in Table 2 for the fixed slope and the standard deviation of random effects. It should be mentioned that the algorithm failed to estimate the variance of the random slopes, σ_{u_1}, for the dimensions BP, SF, MH and IMP. In fact, for the aforementioned dimensions, due to the limited number of measurements per subject, the model is not able to capture differences in the longitudinal trends of individuals. Therefore, we adjusted the model by removing the random slopes effects from the linear predictor and evaluated it again in these dimensions.

Table 2. *Univariate BBMM results. Fixed slope parameter estimates are shown, each with standard error, p-value and odds-ratio (OR) according to one-unit increment in time covariate. Standard deviations of random effects (intercept and slope) are also included with the corresponding standard errors. The symbol * represents tendency to zero deviance.*

		Fixed coefficients			Random sd	
		β_1 (se)	p-value	OR	σ_{u_0} (se)	σ_{u_1} (se)
SF36	PF	-0.06 (0.01)	<0.001	0.94	1.03 (0.03)	0.12 (0.01)
	RP	-0.13 (0.03)	<0.001	0.88	1.67 (0.06)	0.31 (0.02)
	BP	-0.03 (0.02)	0.1347	0.97	1.16 (0.04)	*
	GH	-0.04 (0.01)	<0.001	0.96	0.80 (0.03)	0.06 (0.00)
	VT	-0.04 (0.01)	<0.001	0.96	0.92 (0.03)	0.10 (0.01)
	SF	-0.07 (0.02)	<0.001	0.93	1.53 (0.06)	*
	RE	-0.10 (0.04)	0.007	0.90	2.19 (0.10)	0.53 (0.03)
	MH	-0.03 (0.01)	0.004	0.97	0.99 (0.03)	*
SGRQ	SYMP	0.01 (0.01)	0.499	1.01	0.76 (0.03)	0.10 (0.01)
	IMP	0.00 (0.01)	0.670	1.00	1.06 (0.03)	*
	ACT	0.03 (0.01)	0.001	1.03	1.18 (0.04)	0.13 (0.01)

The results of both questionnaires, generic (SF-36) and disease-specific (SGRQ), show that the health status of the patients is worsening over time. For instance, one year of evolution in the RP dimension is associated with an odds-ratio of 0.88 (i.e. $\exp(-0.13) \approx 0.88$). As a result, for each year of evolution, the patient is approximately 12% less likely to score one more point in the RP dimension, which means that RP patients' scores will decrease resulting in poorer health status.

Table 3. *Univariate Cox model results for hazard ratio including 95% confidence intervals. JM included standardized 0-100 data divided in order to compare results with TSBB and TVCM included binomial form data. Significant results according to p-value are in bold.*

SF-36 scores association with patients' mortality						
	TSBB		TVCM		JM	
	HR	95% CI	HR	95% CI	HR	95% CI
PF (20)	**0.91**	**0.88, 0.94**	**0.88**	**0.84, 0.90**	**0.87**	**0.83, 0.90**
RP (4)	0.96	0.85, 1.09	**0.82**	**0.74, 0.90**	**0.74**	**0.63, 0.88**
BP (9)	0.97	0.88, 1.07	**0.90**	**0.85, 0.96**	0.91	0.81, 1.02
GH (20)	0.98	0.93, 1.02	**0.94**	**0.90, 0.98**	**0.94**	**0.90, 0.99**
VT (20)	**0.94**	**0.90, 0.98**	**0.92**	**0.89, 0.95**	**0.91**	**0.87, 0.95**
SF (8)	0.94	0.84, 1.03	**0.86**	**0.80, 0.93**	**0.82**	**0.72, 0.93**
RE (3)	0.93	0.80, 1.08	**0.79**	**0.70, 0.90**	**0.69**	**0.54, 0.87**
MH (13)	0.96	0.90, 1.03	**0.89**	**0.85, 0.94**	**0.92**	**0.85, 0.99**

Once longitudinal parameters are estimated, fitted values are computed and included in a classic Cox model to evaluate the relationship between HRQoL scores evolution and patient's risk of death. Therefore, the second step of our proposal deals with the second objective of the COPD study, which consisted of evaluating the association between HRQoL results and patient's risk of death. We considered both questionnaires in the survival analysis. First, for SGRQ, we considered a multivariate approach that incorporates the questionnaire's three dimensions in the survival model. However, due to the existing correlation between the dimensions, parameter interpretation can be misleading as it is shown in Table 4 and will be explained in the next paragraphs. It is worth mentioning that the Cox model can easily handle multiple longitudinal covariates, but `JM` package can only incorporate one longitudinal covariate per model and, thus, only univariate models can be performed. Second, in order to avoid misleading interpretation due to covariate correlation and as an illustrative way to show the diversity in the covariate inclusion that the proposed model offers, for generic SF-36 questionnaire we considered an univariate survival approach and thus, we incorporated each dimension separately into the survival models. Furthermore, we included the transformed binomial form data in TVCM for its easy interpretation, but we considered the standardized 0–100 data in the JM methodology because it is more intuitive to assume a normality-based longitudinal model. However, to facilitate a comparison of the estimations of the different methods, we divided the standardized 0–100 scores such that each dimension's maximum score coincides with that in its binomial form.

Results are displayed in Table 3 for the SF-36 survey and Table 4 for the SGRQ. The results are shown according to the association coefficient α of Equation (6), significant results are in bold and they are graphically shown in a forest plot in Figure 5 and 6 for SF-36 and SGRQ respectively. The association coefficients are interpreted in terms of Hazard ratio (HR$=\exp(\alpha)$). Thus, at any particular time point t, HR denotes the

relative increase in the risk at time t that results from one unit increase in the longitudinal variable at the same time point. This interpretation is akin to the odds-ratio interpretation of coefficients in Beta-Binomial regression.

Table 4. *Multivariate Cox model results for hazard ratio including 95% confidence intervals. JM included standardized 0–100 data divided in order to compare results with TSBB and TVCM that included binomial form data. Significant results according to p-value are in bold. (*) Univariate models were considered in JM.*

SGRQ scores association with patients' mortality						
	TSBB		TVCM		JM (*)	
	HR	95% CI	HR	95% CI	HR	95% CI
SYMP (24)	**0.94**	**0.89, 0.99**	0.98	0.94, 1.01	1.03	0.99, 1.07
IMP (24)	0.99	0.93, 1.06	0.98	0.93, 1.04	**1.07**	**1.03, 1.10**
ACT (24)	**1.11**	**1.06, 1.17**	**1.11**	**1.06, 1.17**	**1.09**	**1.06, 1.13**

The results show that the PF (measures mobility disability) and VT (measures energy and fatigue) dimensions were statistically significant in predicting patients' risk of death according to our TSBB methodology. Because PF and VT are two dimensions of the generic SF-36 test, higher scores are associated with lower risk of death, with PF having the greatest impact. In particular, our TSBB proposal relates one more point in the PF dimension with a 9% lower risk. These two significant dimensions of the SF-36 deal with daily patients' activity and how do they feel about it, so the patients' perception about their physical activity is related with a significant impact on mortality.

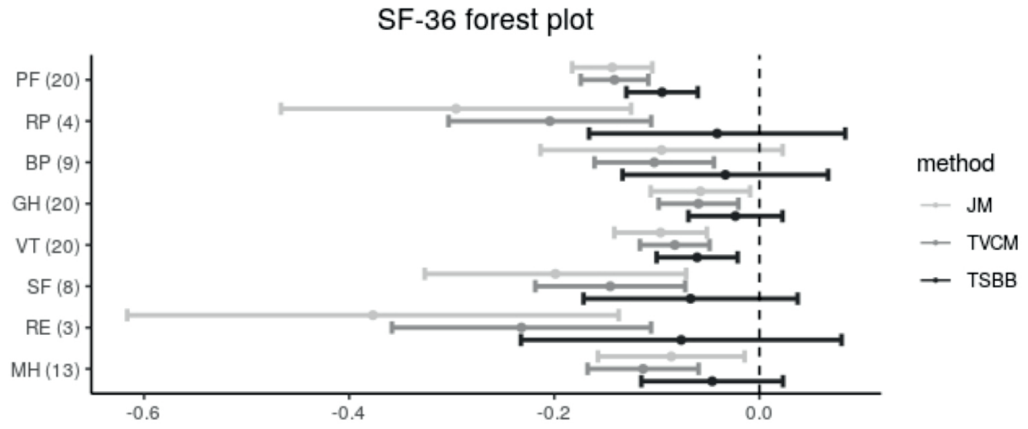

Figure 5. *Forest plot of association coefficient for each SF-36 dimension according to each methodology.*

TVCM results are quite different because all dimensions had signficative effect on patients' risk. It is important to remark that, as it was mentioned in Section 4.1, this methodology extends the previous longitudinal outcome until the event occurrence and

therefore less variability is assumed. This leads into lower standard deviation and smaller confidence intervals, which can be the source because of all the results remain significant.

JM methodology is dealing with standardized data that we transformed in order to compare the one-point change in the dimensions' score and thus, results are mainly different in all dimensions that those obtained from TVCM and TSBB. The results of JM for PF and VT dimensions showed high decrease in patient's risk for one more point scored in these dimensions compared to TSBB and TVCM. This fact will be due to overestimation of this methodology shown in the simulation study in Section 6.

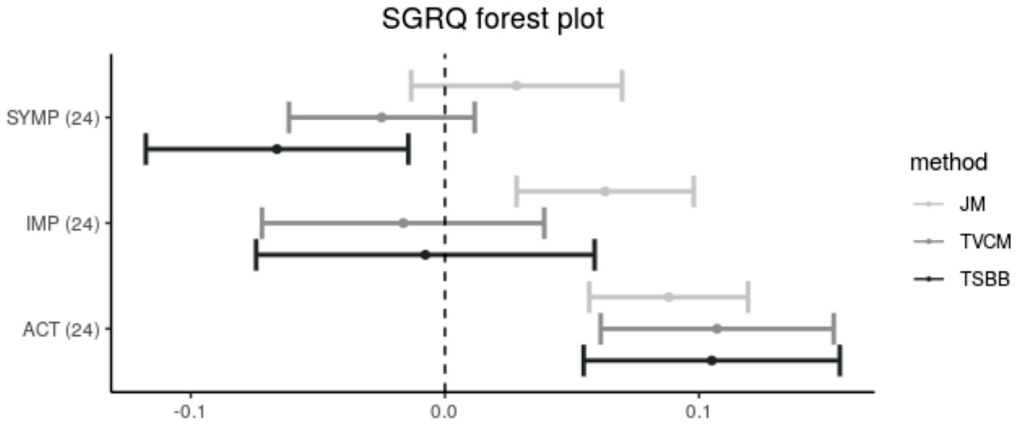

Figure 6. *Forest plot of association coefficient for each SGRQ dimension according to each methodology.*

The SGRQ scores are related with lower health status of the subjects. Particularly, the ACT dimension (the effect of disturbances to mobility and physical activity) was significant in the three methodologies having the greatest impact on patients mortality. Our TSBB methodology and TVCM coincides in associating one more point in the score of this dimension with 11% higher risk, being a 9% risk increment in JM method. We found that the TSBB multivariate model associates an additional point in the SYMP dimension (quantifying distress due to respiratory symptoms) with 6% lower risk. This misleading effect occurs because of covariate correlation. In fact, both approaches that incorporate the three dimensions in the same model, TSBB and TVCM, attribute a decreasing risk effect to the increment of SYMP dimension, although in TCVM the effect is not statistically significant. On the other hand, when JM univariate analysis was performed, IMP dimension showed significant effect on risk where one added point is associated with 7% higher risk of death. Finally, we can conclude that in both questionnaires, the patients' perception of their physical activity and how they feel about it could be a great indicator to take into account in COPD patients' mortality.

6. Simulation Study

In this section, we present a simulation study to assess the performance of the proposed method and compare it with two methodologies widely used in the literature to deal with the analysis of the parameter association between the longitudinal and time-to-event outcomes: TVCM and JM with longitudinal normality assumption. Generally, in literature, TVCM is performed when the main interest is to assess the relationship between the two outcomes because of its popularity and flexibility, leaving the longitudinal model unspecified. The performance of the joint model based in normality assumption is the most commonly used in literature even for discrete and bounded longitudinal data, where standardization or data transformation is performed, avoiding data nature. The simulations have two main aims: (a) validate our proposal approach under several parameter conditions and (b) compare the performance of these three approaches in controlled scenarios offering a variability of situations according to longitudinal shape and the relationship among outcomes.

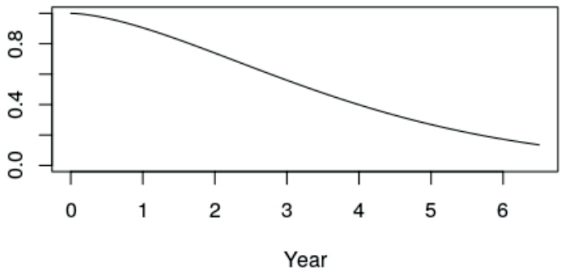

Figure 7. *Baseline survival function for the Weibull baseline hazard with scale and shape parameters* $w_1 = 0.1$, $w_2 = 1.6$

The overall scenario settings are mainly based on COPD study, which was detailed in Section 4. We are considering the same maximum number of measurements per patient, which is four. The longitudinal outcome was computed at the entry and measurement times, dispersed as in the COPD study. Thus, the measurements times are generated as follow: The first measurement corresponds to the subject's entry time which is performed according to a uniform distribution in the interval $(0, 1.5)$. The second measurement is one year apart from the first, as well as the third measurement from the second. However, the fourth measurement is three years apart from the third measurement. The described times are not equally spaced, as in the COPD study, where the data are collected at irregularly spaced times. The overall follow-up period is of five years since the simulated subject's entry.

Time-to-event or survival times, denoted as T_i^*, are generated by evaluating the inverse of a cumulative hazard from Equation (3), see Crowther and Lambert (2013). To that aim, we assumed the Cox proportional hazard model with Weibull baseline risk function $h_0(t) = w_1 w_2 t^{w_2 - 1}$, where w_1 and w_2 denote the scale and shape parameter re-

spectively. See Figure 7 for the parameters $w_1 = 0.1$ and $w_2 = 1.6$. By choosing this baseline hazard function we aim to simulate a baseline survival function that don't reach 0 at the end of the follow-up period, as not all patients die at the end of the COPD study.

We also considered an administrative censoring time A_i according to the COPD study, such that if they reach the fourth measurement, the five-year follow-up period ends without event observation, thus it satisfies $A_i = t_{i4}$. Besides, in order to perform possible dropouts, a loss of follow-up censoring time C_i is performed with a uniform distribution between patient time entry and patient last measurement time with around 10% of censored individuals, which is the censoring rate we found in the COPD study. Finally, the observed time for each subject is computed as $T_i = \min\{T_i^*, A_i, C_i\}$. Furthermore, an event indicator δ_i is recorded, i.e, if $T_i = T_i^*$, then $\delta_i = 1$ and 0 otherwise. Generated marker values $y_i(t_{ij})$ with $t_{ij} > T_i$ were disregarded.

The experiment consisted of 200 random simulations with 250 subjects. We considered a maximum of four measurements per subject, resulting in 1000 observations per simulation of a longitudinal variable distributed as a Beta-Binomial with a fixed maximum score m, a probability parameter p, and a dispersion parameter ϕ. For the sake of clarity, probability parameter p is computed according to the model assumed in Equation (7), considering subjects' overall and specific evolution over time.

Fitted values in Equation (3) of the survival model are computed according to the Beta-Binomial distribution model following Equation (5), which using linear predictor from Equation (7) leads to:

$$m_i(t) = \frac{m}{1 + \exp\{-((\hat{\beta}_0 + \hat{u}_{i0}) + (\hat{\beta}_1 + \hat{u}_{i1})t)\}}. \tag{8}$$

In order to achieve both objectives of the simulation study, all longitudinal parameters were set except for the dispersion parameter ϕ and the association parameter α of the survival model. By giving different values to the dispersion parameter ϕ, longitudinal data can adopt a wide variety of shapes and varying parameter α allows us to evaluate a small, moderate and strong association between the two outcomes of interest.

Two main scenarios were considered in order to simulate longitudinal responses based on two dimensions of the HRQoL questionnaires considered in the COPD study, one dimension of the SF-36 and another one of the SGRQ. This scenario division will allow us to evaluate positive and negative association among outcomes, like it happens in the SF-36 and the SGRQ respectively. In one scenario we considered 24 as response maximum score and positive association parameter while in the other 8 was set up as maximum score and negative association parameter. See Table 5 for a summary of both main scenarios.

Table 5. *Setting parameters to perform two main simulation scenarios.*

	β_0	β_1	σ_{u_0}	σ_{u_1}	m	α
Scenario 1	-0.19	0.03	1.2	0.05	24	> 0
Scenario 2	0.40	-0.15	1.5	0.3	8	< 0

Once the two main scenarios were set, we provided scenario division by varying dispersion parameter using $\phi \in \{0.05, 0.5, 1\}$, so that, longitudinal distribution takes different shapes. See Figure 8 and 9 for a graphic image of the different generated shapes for Scenario 1 and Scenario 2 respectively. Notice that we have generated bell, flat, U, J and inverse J shapes.

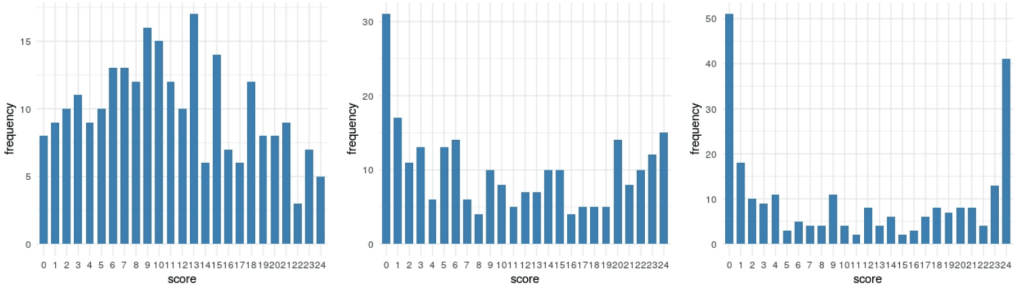

Figure 8. *Frequency of the simulated longitudinal scores based on Scenario 1 with* $\phi = 0.05, 0.50, 1$ *from left to right.*

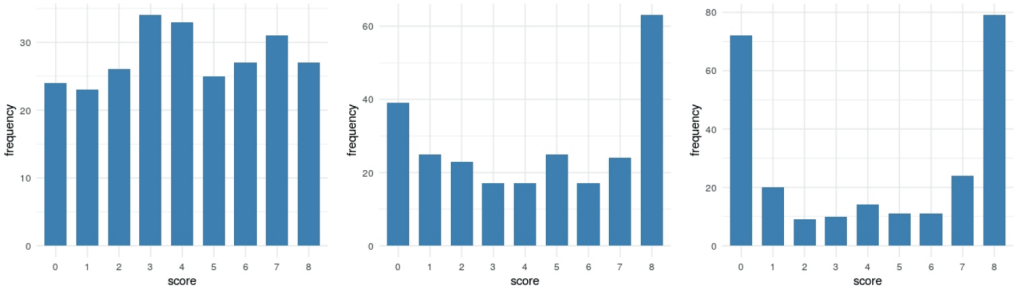

Figure 9. *Frequency of the simulated longitudinal scores based on Scenario 2 with* $\phi = 0.05, 0.50, 1$ *from left to right.*

Furthermore, we want to compare the performance of the methods by varying the relationship between the two outcomes of interest and thus the association parameter was set in order to perform scenarios corresponding to small, moderate and strong association between $m_i(t)$ and $h_i(t)$. As association parameter interpretation is done considering the hazard ratio in terms of scoring one additional point, it is important to take into account the maximum score number, because increasing an additional point does not have the same effect when the maximum score is 8 or 24. Thus, according to maximum number of trials, we set the association parameter as $\alpha \in (0.01, 0.05, 0.10)$ for Scenario 1 and $\alpha \in (-0.05, -0.10, -0.15)$ for Scenario 2. These modifications of ϕ and α lead to 9 sub-scenarios for each of the two main set above. Methods are represented as TSBB (Two-Stage Beta-Binomial) for our proposed methodology, TVCM for Time-Varying Cox Model and JM for Joint Model normality-based.

Although the estimates of all parameters (longitudinal and survival) have been obtained by both joint model methodologies, only results for α will be shown in detail.

First, because the association parameter between outcomes is the one estimated by the three methodologies, as TVCM does not provide a longitudinal analysis. Besides, longitudinal coefficients cannot be entirely compared for JM and TSBB because the longitudinal model differs. Lastly, because the simulation objective is to compare the methods' performance when estimating the relationship between the two outcomes of interest. Moreover, related to baseline hazard function, as TSBB and TVCM do not need to specify it, we set piecewise method in JM baseline hazard selection in order to allow a flexible semi-parametric methodology.

The results of the simulation study are summarised in Table 6 for Scenario 1 and Table 7 for Scenario 2. Results in both tables are based on the following statistics: the *%Bias* ($bias/\alpha$), the empirical standard deviation (ESD), the average standard deviation (ASD) and the 95% coverage probability (CP). It is worth mentioning that convergence problems related to the normality assumption in the longitudinal model were found when applying the JM methodology. Mainly it occurred in Scenario 1 for $\alpha = 0.1$, where there is a higher risk of death and, therefore, fewer longitudinal data. The convergence percentage was only 34% for $\phi = 0.5$ sub-scenario, which makes this methodology hardly applicable in case studies. Only in Scenario 2 for $\alpha = -0.05$ and $\phi = 0.05$ this methodology converged 100% times. To compare the three different approaches, we contemplated those 200 realizations where all methods converged.

Table 6. *Results of the association parameter α obtained from the proposed method (TSBB), the TVCM and joint model (JM) fitted to data considering Scenario 1 in Table 5 with $\alpha \in (0.01, 0.05, 0.10)$ and $\phi \in (0.05, 0.5, 1)$. Percentage bias (%Bias), empirical standard deviation (ESD), average standard deviation (ASD) and 95% coverage probability (CP) are shown.*

		Scenario 1											
		$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.10$			
ϕ		*%Bias*	ESD	ASD	CP	*%Bias*	ESD	ASD	CP	*%Bias*	ESD	ASD	CP
0.05	TSBB	0.11	0.012	0.012	98.0	0.37	0.014	0.012	72.0	0.36	0.018	0.014	33.0
	TVCM	0.18	0.010	0.011	96.0	0.12	0.012	0.011	89.0	0.05	0.011	0.012	94.0
	JM	0.43	0.012	0.013	95.5	0.48	0.016	0.014	63.5	0.45	0.018	0.018	33.5
0.5	TSBB	-0.01	0.013	0.013	96.0	0.23	0.018	0.013	76.0	0.09	0.025	0.014	73.5
	TVCM	-0.37	0.008	0.008	92.5	-0.36	0.009	0.008	38.5	-0.47	0.008	0.008	0.0
	JM	0.46	0.015	0.016	95.0	0.51	0.022	0.019	75.0	0.24	0.034	0.024	84.0
1	TSBB	-0.16	0.015	0.012	86.5	0.04	0.016	0.012	85.0	-0.10	0.025	0.012	60.0
	TVCM	-0.47	0.008	0.007	88.5	-0.50	0.008	0.007	7.5	-0.59	0.007	0.007	0.0
	JM	0.50	0.018	0.017	92.5	0.48	0.024	0.021	87.5	0.21	0.046	0.029	77.5

First, both tables show that TVCM has the lowest ESD in all scenarios due to the LVCF approach that assumes longitudinal outcomes constant among measurement times. According to Table 6, we can see that this methodology produces highly biased

results, specially when the dispersion and association parameters are increased, as evidenced by the poor CP results, which reached zero in some sub-scenarios. Table 7 demonstrates that TVCM produced less biased results when there is small association between $m_i(t)$ and $h_i(t)$. Otherwise, its results are generally not the least biased in the moderate and strong association scenarios, which, in addition to its low ESD, is reflected in its lower CP when compared to other methods.

For the TSBB approach, the ESD remains quite similar in all the scenarios of Table 6, being slightly larger when dispersion or association parameters increase, which also happens in Table 7. This conclusion results quite logical as increasing variability in the outcomes also increases the uncertainty of the estimations. Its ESD results are higher than the ones in TVCM in all sub-scenarios because fitting longitudinal outcomes when the event takes place includes more variability rather than considering the last value recorded constant. Concerning bias, TSBB presents the lowest *%Bias* in Table 6 for most of the scenarios, and it does so in Table 7 except for the small-association case, where only around 33 results were statistically significant. However, in Table 6, we find low CP compared with other methodologies, mainly in sub-scenario $\alpha = 0.1$ that might be due to high patient risk, that few longitudinal data is taken into account and coefficient estimation is poorer. Then, there is greater variability in estimates of the longitudinal sub-model. However, in the second step we include the fitted value in the Cox model without incorporating the estimated variability of the longitudinal part. Then the difference between the ESD and the ASD produces the low CP values in those sub-scenarios.

Table 7. *Results of the association parameter α obtained from the proposed method (TSBB), the TVCM and joint model (JM) fitted to data considering Scenario 2 in Table 5 with $\alpha \in (-0.05, -0.10, -0.15)$ and $\phi \in (0.05, 0.5, 1)$. Percentage bias (%Bias), empirical standard deviation (ESD), average standard deviation (ASD) and 95% coverage probability (CP) are shown.*

		Scenario 2											
		$\alpha = -0.05$				$\alpha = -0.10$				$\alpha = -0.15$			
ϕ		*%Bias*	ESD	ASD	CP	*%Bias*	ESD	ASD	CP	*%Bias*	ESD	ASD	CP
0.05	TSBB	-0.32	0.033	0.031	90.5	0.09	0.032	0.033	95.0	0.23	0.037	0.038	89.0
	TVCM	0.17	0.031	0.028	90.0	0.26	0.032	0.031	87.5	0.30	0.035	0.034	75.0
	JM	0.74	0.037	0.035	81.5	0.78	0.039	0.039	51.5	0.83	0.049	0.046	23.0
0.5	TSBB	-0.41	0.027	0.030	91.5	-0.02	0.031	0.033	97.5	0.07	0.035	0.037	96.0
	TVCM	-0.12	0.023	0.024	95.0	-0.09	0.026	0.026	95.0	-0.08	0.029	0.029	92.5
	JM	0.79	0.038	0.040	85.0	0.80	0.045	0.045	59.5	0.85	0.056	0.055	32.0
1	TSBB	-0.43	0.029	0.029	88.0	-0.12	0.032	0.032	94.0	0.00	0.035	0.035	95.0
	TVCM	-0.23	0.022	0.022	91.0	-0.23	0.023	0.024	86.0	-0.19	0.026	0.027	81.0
	JM	0.87	0.044	0.043	84.5	0.83	0.053	0.050	62.5	0.92	0.069	0.061	33.5

Lastly, the JM approach ends with the highest *%Bias* in almost all the scenarios, specially in Table 7 where the maximum score number of the longitudinal outcome was set equal to 8. Normality assumption tends to fail as the maximum score number is lower in binomial data. Therefore, the simulation study shows the importance of considering an adequate distribution for the longitudinal outcome. Moreover, we can see that JM results are overestimated in all sub-scenarios which, in fact, the same effect can be observed in Section 5 where real data estimations with JM methodology were more extreme compared to those obtained with TSBB or TVCM methodologies.

Based on the simulation study results, we can conclude that although our proposal is based on a two-stage methodology, which is known as a biased approach, it performed better estimations in most scenarios. The first reason is that TVCM results are more biased in most cases and tend to underestimate the association parameter. Lastly, a Joint Model with a normality-based assumption for longitudinal data leads to really skewed results and high ESD. Moreover, it has significant convergence problems due to the normality assumption of the discrete and bounded longitudinal data. Hence, we emphasize the importance of considering a complete analysis of both outcomes and remark on the consequences of considering an inaccurate distribution of longitudinal data.

7. Conclusion and further research

In biological or clinical trials, which usually involve the follow-up of subjects, longitudinal and time-to-event variables are often recorded. Traditionally, although interest relied primarily on clinical indicators, PRO measurements have gained relevance in recent years and are now highly recommended in patients' assessments (Deshpande et al., 2011). This tool allows researchers to collect patient-perceived information about several topics, usually regarding health-related quality of life (HRQoL). Several studies are arising to examine HRQoL as a surrogate indicator of prognosis concerning mortality, showing the usefulness of HRQoL in the prediction of prognosis (Esteban et al., 2022). However, we found that some analysis were performed under a cross-sectional framework that did not take into account the repeated examinations and others that are only focused on measure patient's HRQoL evolution without considering survival data.

In the literature, we realized that there are two common methodologies for analyzing the relationship between longitudinal and time-to-event outcomes: Time-Varying Cox Model (TVCM) and the classic Joint Model based on shared-random effects (JM). TVCM incorporates the repeated measurements as covariates considering its value constant between the measurement time intervals, whereas JM jointly fits both outcomes where it is considered a longitudinal normality-based model. However, these two popular approaches are not able to incorporate PRO characteristics. Firstly because TVCM is not considered adequate for endogenous measurements that require the survivality of the patient to be recorded. Lastly, because the classic Joint Model leads to computational complexities if normal distribution is not assumed for the longitudinal data. To fill this gap in the state of the art, the main objective of this work is to perform a complete anal-

ysis of longitudinal HRQoL measurements and survival data that incorporates the PRO discrete, bounded and overdispersed essence.

In this study, we proposed a joint model based on two-stage methodology that allows the inclusion of the Beta-Binomial distribution for longitudinal data and also avoids computational complexities. We focused on the idea of a suitable fitting for repeated measurements that enhance the estimation of the relationship between longitudinal and survival data. Moreover, we emphasized how critical it is to select a wrong longitudinal model when considering a joint model performance. Our proposal includes a Beta-Binomial mixed-effect model to accommodate the correlation structure of longitudinal overdispersed discrete and bounded outcomes. In addition, the semi-parametric Cox model is set for survival data where longitudinal fitted values are included as covariates. Although the two stage approach is not a novel methodology for the joint modelling framework, the estimation procedure that we propose is the first one in the literature that entirely incorporates the nature of the PRO data.

This approach also provides a longitudinal fit of the PRO data that TVCM does not consider and also set an adequate model for repeated PRO measurements that causes computational complexities in the JM methodology. Moreover, several longitudinal PRO measurements can be easily incorporated in the survival model which is not possible in the JM approach. To compare the performance of our proposed methodology with those that are widely used in the literature for assessing the relationship between longitudinal PRO measurements and survival data, we carried out a simulation study. The simulation study showed that in most of the scenarios, the proposed method obtained better performance with a smaller bias and a greater coverage probability of the parameter than the other methods. Besides, results showed that JM presented significant convergence problems due to normality assumption, which makes it hardly applicable for real data. This method also lead to really biased estimations. The TVCM was mainly affected by PRO overdispersion because when variability is increased in the longitudinal outcomes its bias worsens due to the last value carried forward assumption.

We also applied the methodologies to the detailed COPD study to analyze real data and show that our methodology presents an easy interpretation of the results in terms of odds-ratio and hazard ratio. Relevant conclusions were exposed for both generic and disease-specific questionnaires. It was shown that the patient's perception of its activity levels and its feelings about it has a significant impact on the patient's risk of death.

Our proposed methodology has potential limitations. The two-stage methodology is mainly known for being a biased approach and it is shown in the simulation study results seemed to be biased. Moreover, a few number of repeated measurements per subject can lead into a to poor estimates of the longitudinal model that will be reflected in the estimation of the association parameter. Furthermore, the second step of the process does not account for the variability of longitudinal estimations, which can underestimate the variance of the estimations as it was shown in simulation study by ESD and ASD differences. Consequently, it sometimes results in low CP results despite of having a low bias. For further work, we aim to investigate a Bayesian approach that allow us to incorpo-

rate Beta-Binomial distribution for repeated measurements and jointly fits longitudinal and survival models. We also aim to include the longitudinal estimation variability in order to avoid low CP results and correct the potential bias that standard two-stage methods present. In the literature, several proposals exist for mechanisms to reduce the bias inherent in the two-stage joint modeling methodology while preserving its main advantages: the ability to address complex structures flexibly and the reduced computational demand. See, for instance, the works of Leiva-Yamaguchi and Alvares (2021) and Alvares and Leiva-Yamaguchi (2023).

As concluding remarks, we recommend the used of our proposed two-stage methodology that incorporates a Beta-Binomial mixed-effects model into the joint modelling framework. This methodology provides a complete analysis of longitudinal PRO and survival data with an adequate fit over popular methodologies.

Acknowledgements

This research is supported by AEI and EJ-GV under Grants PID2020-115882RB-I00 / AEI / 10.13039/5011 00011033 with acronym "S3M1P4R", Departamento de Educacion, Política Lingüística y Cultura del Gobierno Vasco [IT1456-22] and also by the Basque Government through the BERC 2022-2025 program and by the Ministry of Science and Innovation: BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033. We also gratefully acknowledge Dr. Cristobal Esteban for providing the COPD data and Inmaculada Arostegui for the scientific discussion.

References

Alvares, D. and Leiva-Yamaguchi, V. (2023). A two-stage approach for Bayesian joint models: reducing complexity while maintaining accuracy. *Statistics and Computing*, 33(5):1–11. ISSN: 15731375. DOI: 10.1007/s11222-023-10281-9. URL: https://doi.org/10.1007/s11222-023-10281-9.

Arisido, M. W., Antolini, L., Bernasconi, D. P., Valsecchi, M. G. and Rebora, P. (2019). Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint. *BMC Medical Research Methodology*, 19(1):1–13. ISSN: 14712288. DOI: 10.1186/s12874-019-0873-y.

Armero, C. (2021). Bayesian Joint Models for Longitudinal and Survival Data. *Wiley StatsRef: Statistics Reference Online*, pages 1–6. DOI: https://doi.org/10.1002/9781118445112.stat08129.

Arostegui, I., Núñez-Antón, V. and Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine*, 26(June):1318–1342. DOI: 10.1002/sim.

Arostegui, I., Núñez-Antón, V. and Quintana, J. M. (2013). On the recoding of continuous and bounded indexes to a binomial form: An application to quality-of-

life scores. *Journal of Applied Statistics*, 40(3):563–582. ISSN: 02664763. DOI: 10.1080/02664763.2012.749845.

Crowther, M. J. and Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23):4118–4134. ISSN: 0277-6715. DOI: 10.1002/sim.0000.

Deshpande, P. R., Sudeepthi, B. L., Rajan, S. and Nazir, C. P. A. (2011). Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Research*, 2(4):137. ISSN: 2229-3485. DOI: 10.4103/2229-3485.86879.

Domingo-Salvany, A., Lamarca, R., Ferrer, M., Garcia-Aymerich, J., Alonso, J., Félez, M., Khalaf, A., Marrades, R. M., Monsó, E., Serra-Batlles, J. and Antó, J. M. (2002). Health-related quality of life and mortality in male patients with chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 166(5):680–685. ISSN: 1073449X. DOI: 10.1164/rccm.2112043.

Esteban, C., Arostegui, I., Aramburu, A., Moraza, J., Aburto, M., Aizpiri, S., Chasco, L. and Quintana, J. M. (2022). Changes in health-related quality of life as a marker in the prognosis in COPD patients. *ERJ Open Research*, 8(1):00181–2021. DOI: 10.1183/23120541.00181-2021. URL: http://dx.doi.org/10.1183/23120541.00181-2021.

Esteban, C., Arostegui, I., Aramburu, A., Moraza, J., Najera-Zuloaga, J., Aburto, M., Aizpiri, S., Chasco, L. and Quintana, J. M. (2020). Predictive factors over time of health-related quality of life in COPD patients. *Respiratory Research*, 21(1):1–11. ISSN: 1465993X. DOI: 10.1186/s12931-020-01395-z.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*, 15(15):1663–1685. ISSN: 02776715. DOI: 10.1002 / (SICI) 1097-0258(19960815)15:15¡1663::AID-SIM294¿3.0.CO;2-1.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796–2801. ISSN: 0732183X. DOI: 10.1200/JCO.2009.25.0654.

Jones, P. and Higenbottam, T. (2007). Quantifying of severity of exacerbations in chronic obstructive pulmonary disease adaptations to the definition to allow quantification. *Proceedings of the American Thoracic Society*, 4(8):597–601. ISSN: 15463222. DOI: 10.1513/pats.200707-115TH.

Jones, P. W. (2005). St. George's respiratory questionnaire: MCID. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 2(1):75–79. ISSN: 15412555. DOI: 10.1081/COPD-200050513.

Jones, P. W., Quirk, F. H. and Baveystock, C. M. (1991). The St George's Respiratory Questionnaire. *Respiratory Medicine*, 85:25–31. ISSN: 09546111. DOI: 10.1016/S0954-6111(06)80166-6.

Leiva-Yamaguchi, V. and Alvares, D. (2021). A two-stage approach for bayesian joint models of longitudinal and survival data: Correcting bias with informative prior. *Entropy*, 23(1):1–10. ISSN: 10994300. DOI: 10.3390/e23010050.

Li, Z., Tosteson, T. D. and Bakitas, M. A. (2013). Joint modeling quality of life and survival using a terminal decline model in palliative care studies. *Statistics in Medicine*, 32(8):1394–1406. ISSN: 02776715. DOI: 10.1002/sim.5635.

Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2018). Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. *Statistical Methods in Medical Research*, 27(10):2989–3009. ISSN: 14770334. DOI: 10.1177/0962280217690413.

Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2019). A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes. *Biometrical Journal*, 61(3):600–615. ISSN: 15214036. DOI: 10. 1002 / bimj . 201700251.

Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2022). *PROreg: Patient Reported Out- comes Regression Analysis*. R package version 1.2. URL: https://CRAN.R-project.org/package=PROreg.

Papageorgiou, G., Mauff, K., Tomer, A. and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application*, 6(August 2018):223–240. ISSN: 2326831X. DOI: 10.1146/annurev-statistics-030718-105048.

Pauwels, R. A. and Rabe, K. F. (2004). Burden and clinical features of chronic obstructive pulmonary disease (COPD). *Lancet*, 364(9434):613–620. ISSN: 01406736. DOI: 10.1016/S0140-6736(04)16855-4.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-108.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33. ISSN: 15487660. DOI: 10.18637/jss.v035.i09.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. CRC Press, p. 274. ISBN: 1439872864.

Rosenberg, P. S. (1995). Hazard Function Estimation Using B-Splines. *Biometrics*, 51(3):874. ISSN: 0006341X. DOI: 10.2307/2532989.

Self, S. and Pawitan, Y. (1992). Modeling a Marker of Disease Progression and Onset of Disease. *AIDS Epidemiology*, (1991):231–255. DOI: 10.1007/978-1-4757-1229-2_11.

Speight, J. and Barendse, S. M. (2010). FDA guidance on patient reported outcomes. *BMJ (Online)*, 341(7772):518. ISSN: 17561833. DOI: 10.1136/bmj.c2921.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN: 0-387-98784-3.

US Department of Health and Human Services (2006). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health and Quality of Life Outcomes*, 4:1–20. ISSN: 14777525. DOI: 10.1186/1477-7525-4-79.

Vanfleteren, L. E. G. W., Spruit, M. A., Wouters, E. F. M. and Franssen, F. M. E. (2016). Management of chronic obstructive pulmonary disease beyond the lungs.

The Lancet Respiratory Medicine, 4(11):911–924. ISSN: 22132619. DOI: 10.1016/ S2213-2600(16)00097-7. URL: http://dx.doi.org/10.1016/ S2213-2600(16)00097-7.

Ware, J. E., Snow, K. K., Kosinski, M. and Gandek, B. (1993). SF36 Health Survey, Manual and Interpretation Guides.

Weldring, T. and Smith, S. M. S. (2013). Article Commentary: Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health Services Insights*, 6:61–68. ISSN: 11786329. DOI: 10.4137/HSI.S11093.

Wiklund, I. (2004). Assessment of patient-reported outcomes in clinical trials: The example of health-related quality of life. *Fundamental and Clinical Pharmacology*, 18(3):351–363. ISSN: 07673981. DOI: 10.1111/j.1472-8206.2004.00234.x.

Wu, L., Liu, W., Yi, G. Y. and Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012. ISSN: 1687952X. DOI: 10.1155/2012/640153.

SORT 48 (2) July-December 2024, 183-208 DOI: 10.57645/20.8080.02.18

Non-parametric estimation of the covariate-dependent bivariate distribution for censored gap times

Ewa Strzalkowska-Kominiak[1], Elisa M. Molanes-López[2] and Emilio Letón[3]

Abstract

In many biomedical studies, recurrent or consecutive events may occur during the follow-up of the individuals. This situation can be found, for example, in transplant studies, where there are two consecutive events which give rise to two times of interest subject to a common random right-censoring time, the first one being the elapsed time from acceptance into the transplantation program to transplant, and the second one the time from transplant to death. In this work, we incorporate the information of a continuous covariate into the bivariate distribution of the two gap times of interest and propose a non-parametric method to cope with it. We prove the asymptotic properties of the proposed method and carry out a simulation study to see the performance of this approach. Additionally, we illustrate its use with Stanford heart transplant data and colon cancer data.

MSC: *62N01, 62N02, 62P10, 62G05.*

Keywords: *Bivariate distribution, Copula function, Covariate, Serial dependence, Random censoring, Kernel estimation.*

1. Introduction

In many biomedical studies, recurrent or consecutive events may occur during the follow-up study of the individuals. In this setting, it is of interest to study the time between consecutive events subject to a common censoring variable. In the literature, these consecutive times are known as gap times. Examples of this situation can be found in the recurrence of breast cancer, bleeding episodes in patients with liver cirrhosis, AIDS studies or transplantation in heart studies. Several authors have already dealt with estimating

[1] Department of Statistics, Universidad Carlos III de Madrid, Spain, email: estrzalk@est-econ.uc3m.es (corresponding author).

[2] Department of Statistics and Operations Research, Facultad de Medicina, Universidad Complutense de Madrid (UCM), Madrid, Spain.

[3] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Spain.

Received: July 2023

Accepted: December 2023

the distribution function of such gap times (see, e.g., Lin, Sun and Ying (1999), Meira-Machado and Roca-Pardiñas (2011), Serrat and Gómez (2007) or Huang and Louis (1998), among others). Bivariate consecutive data under interval sampling have been considered in Zhu and Wang (2012), where the authors indicated the importance of applying these models to cancer study. Moreover, a bivariate estimation under censoring and with semi-competing risk data, has been studied in Fine, Jiang and Chappell (2001) and Wang (2003). Nevertheless, none of these authors consider the influence of the covariates. In the case of complete data, the estimation of the conditional bivariate distribution has been studied in terms of conditional copula function by Gijbels, Veraverbeke and Omelka (2011). In our paper, we introduce a continuous covariate into the censored gap times setup and propose a new method to estimate the conditional bivariate distribution function. This new methodology is an adaptation and a mixture of the methods proposed by Beran (1981) and van Keilegom (2004), which are further used to estimate as well the bivariate conditional density and the marginal distributions. It provides also a strong basis to study the conditional copulas and measures of conditional dependence.

The paper is organized as follows. In Section 2, we introduce the model. In Section 3, we propose our estimator of the conditional joint distribution function of two gap times. In Section 4, we derive kernel type estimators for the conditional distribution function of the two marginal times. In Section 5, we propose a likelihood based bandwidth selector. We check the behaviour of the proposed methods through a simulation study in Section 6. Finally, we illustrate their use with two real data examples in Section 7.

2. Model description

Let T_1 and T_2 be two consecutive times subject to a common random right-censoring variable, C. Denote by $\tilde{T}_1$ and $\tilde{T}_2$ the observed times, that is, $\tilde{T}_1 = \min(T_1, C)$, $\tilde{T}_2 = \min(T_2, C_2)$, where $C_2 = (C - T_1)1_{\{T_1 \leq C\}}$. Moreover, set $\delta_1 = 1_{\{T_1 \leq C\}}$ and $\delta_2 = 1_{\{T_2 \leq C_2\}}$ as the observed censoring indicators. The following three situations may occur:

a) $T_1 + T_2 \leq C \Rightarrow T_1$ and T_2 are observed, that is, $\tilde{T}_1 = T_1$, $\tilde{T}_2 = T_2$, $\delta_1 = 1$ and $\delta_2 = 1$.

b) $T_1 \leq C < T_1 + T_2 \Rightarrow \tilde{T}_1 = T_1, \delta_1 = 1$, $\tilde{T}_2 = C - T_1$ and $\delta_2 = 0$.

c) $C < T_1 \Rightarrow \tilde{T}_1 = C, \delta_1 = 0$, $\tilde{T}_2 = 0$ and $\delta_2 = 0$.

See Figure 1, where $T = T_1 + T_2$, to get a visual insight of these three possible situations.

The joint cumulative distribution function (cdf) of the pair of consecutive times (T_1, T_2) has been previously studied by several authors: Wang and Wells (1998) under the assumption of independence between the censoring time and the gap times; Visser (1996) considered that censoring may depend upon previous gap times, although requiring discrete censoring time and gap times; de Uña-Álvarez and Meira-Machado (2008), Strzalkowska-Kominiak and Stute (2010) and de Uña-Álvarez and Amorim (2011) studied consecutive survival times with a common censoring variable. In this work, we

consider this last setting with the novelty of including extra information given by a one-dimensional continuous covariate, let say X, to gain in estimation efficiency and study its influence on survival times.

To illustrate our setting, we will consider the following examples. The first one is based on the Stanford heart transplant and a colon cancer data set, described in Meira-Machado and Roca-Pardiñas (2011) and Kalbfleisch and Prentice (1980), among others. In this dataset, there are two times of interest, $T_1 =$ time from acceptance into the transplantation program to transplant (in months), and $T_2 =$ time from transplant to death (in months), and some covariates available such as age, year of acceptance and surgery (prior bypass surgery). Additionally, there are other two variables ($delta$ and $status$) that specify if the individual has received a transplant (or not), $delta = 1$ ($delta = 0$), and if he/she has died (or not), $status = 1$ ($status = 0$). Those individuals with both times observed correspond with those that have $delta = 1$ and $status = 1$. In the case that the first time is observed but not the second time, the individuals have $delta = 1$ and $status = 0$. For the rest, with $delta = 0$, none of the two times are observed. The second example is based on a colon cancer dataset, previously studied in Lawless and Yilmaz (2011). For this dataset, the two times of interest are $T_1 =$ time from study registration to recurrence (in years), and $T_2 =$ time from recurrence to death (in years), and the covariates available are treatment and age. Analogously to the first example, there are censoring indicators that state if none, one or both times are censored. In both examples, we aim to assess the effect of age on the vector (T_1, T_2). In the second example, we are also interested in studying the treatment effect. In both cases, the censoring variable C is the time until end of the study which, for a given age, can be assumed independent of the lifetimes of interest.

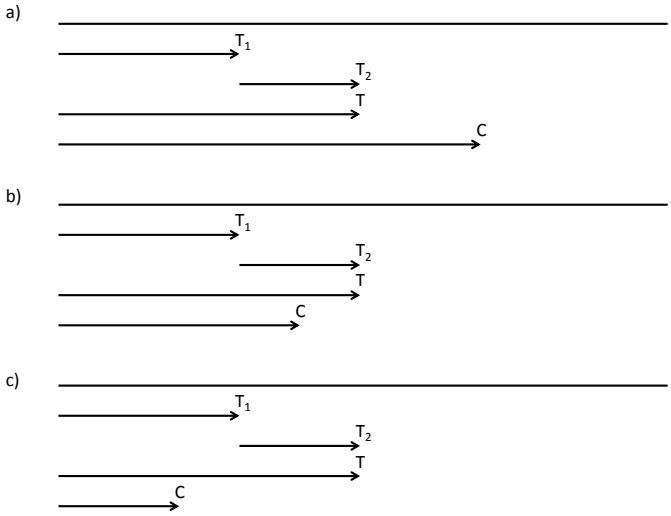

Figure 1. *The three possible situations for T_1, T_2, T and C.*

Based on n i.i.d. replicates, $\{(\tilde{T}_{1i},\tilde{T}_{2i},\delta_{1i},\delta_{2i}),i=1,...,n\}$, of the random vector $(\tilde{T}_1,\tilde{T}_2,\delta_1,\delta_2)$, our goal in this paper is to estimate the bivariate cdf of (T_1,T_2) given that $X=x$, that is,

$$\mathbf{F}(y_1,y_2|x)=\mathbb{P}(T_1\leq y_1,T_2\leq y_2|X=x), \tag{1}$$

under the following assumption:

A1: (T_1,T_2) is independent of C given X.

Note that under assumption A1, it follows the following condition:

A2: T_2 is independent of $C_2=(C-T_1)1_{\{T_1\leq C\}}$ given T_1 and X.

From here on, we denote by $F_T(t)$ the cdf of T, $G(t)$ the cdf of C, and by $G(t|x)$ the conditional cdf of C given X, that is, $G(t|x)=\mathbb{P}(C\leq t|X=x)$. We assume throughout the paper that the densities $\mathbf{f}(y_1,y_2|x)$, $f_X(x)$ and $g(t|x)$, related to $\mathbf{F}(y_1,y_2|x)$, $F_X(x)=\mathbb{P}(X\leq x)$ and $G(t|x)$, exist and are continuous, and that $F_1(y_1|x)=\mathbf{F}(y_1,\infty|x)$ and $F_X(x)$ are differentiable up to order two. Note that $\mathbf{F}$ is identifiable only in the region $\{(t_1,t_2):t_1+t_2\leq\tau_c(x)\}$, where $\tau_c(x)$ is the right hand side of the support of C given $X=x$, that is, $\tau_c(x)=\inf\{t:G(t|x)=1\}$.

3. Conditional joint estimators

The aim of this section is to propose a nonparametric estimator of (1).

Let's $f_{21}(y_2|t_1,x)$ and $f_1(t_1|x)$ denote the densities related to the distributions $F_{21}(y_2|t_1,x)$ and $F_1(t_1|x)$, respectively, where $F_{21}(y_2|t_1,x)=\mathbb{P}(T_2\leq y_2|T_1=t_1,X=x)$. Hence

$$\begin{aligned}\mathbf{F}(y_1,y_2|x)&=\int_0^{y_1}\int_0^{y_2}f(t_1,t_2|x)dt_1dt_2=\int_0^{y_1}\int_0^{y_2}f_{21}(t_2|t_1,x)f_1(t_1|x)dt_2dt_1\\&=\int_0^{y_1}F_{21}(y_2|t_1,x)f_1(t_1|x)dt_1.\end{aligned}$$

Hence, our estimator is based on the following, more general, identity

$$\mathbf{F}(y_1,y_2|x)=\int_0^{y_1}F_{21}(y_2|t_1,x)F_1(dt_1|x),$$

where

$$F_1(y_1|x)=\mathbb{P}(T_1\leq y_1|X=x). \tag{2}$$

Moreover, under assumption A1, we have that

$$\begin{aligned}\mathbb{P}(T_2\leq y_2|T_1=t_1,\delta_1=1,X=x)&=\mathbb{P}(T_2\leq y_2|T_1=t_1,C\geq t_1,X=x)\\&=\mathbb{P}(T_2\leq y_2|T_1=t_1,X=x)\end{aligned}$$

and therefore

$$F_{21}(y_2|t_1,x)=\mathbb{P}(T_2\leq y_2|T_1=t_1,\delta_1=1,X=x). \tag{3}$$

In the following subsections, we propose nonparametric estimators of (2), (3) and (1), respectively.

3.1. Estimation of $F_1(t_1|x)$

We propose to estimate $F_1(t_1|x)$ by the Beran (1981) estimator of the conditional cdf of T_1 given $X = x$, denoted by $F_{1n}(t_1|x)$. More precisely, F_{1n} is a standard version of the Beran estimator, defined as

$$F_{1n}(y|x) = 1 - \prod_{i=1}^{n} \left[1 - \frac{B_{in}(x) 1_{\{\tilde{T}_{1i} \leq y\}} \delta_{1i}}{\sum_{j=1}^{n} 1_{\{\tilde{T}_{1j} \geq \tilde{T}_{1i}\}} B_{jn}(x)} \right], \tag{4}$$

where

$$B_{in}(x) = \frac{K(\frac{x - X_i}{h_1})}{\sum_{j=1}^{n} K(\frac{x - X_j}{h_1})},$$

and h_1 is a bandwidth parameter. As usual, K is a bounded kernel function with the following properties:

$$\int uK(u)du = 0, \text{ and } \int u^2 K(u)du < \infty.$$

3.2. Estimation of $F_{21}(t_2|t_1,x)$

We propose to estimate $F_{21}(t_2|t_1,x)$ by the Beran estimator of the conditional cdf of T_2 given $(T_1, \delta_1, X) = (t_1, 1, x)$, denoted by $F_{21n}(t_2|t_1,x)$.

Let define

$$H^*(t|t_1,x) = \mathbb{P}(\tilde{T}_2 \leq t | T_1 = t_1, \delta_1 = 1, X = x).$$

and

$$\tilde{H}^*(t|t_1,x) = \mathbb{P}(\tilde{T}_2 \leq t, \delta_2 = 1 | T_1 = t_1, \delta_1 = 1, X = x).$$

Given that the assumption A1 is valid, the condition A2 holds and therefore, we obtain that

$$1 - H^*(t^-|t_1,x) = (1 - F_{21}(t^-|t_1,x))\mathbb{P}(C_2 \geq t | T_1 = t_1, \delta_1 = 1, X = x)$$

and

$$\tilde{H}^*(t|t_1,x) = \int_0^t \mathbb{P}(C_2 \geq s | T_1 = t_1, \delta_1 = 1, X = x) F_{21}(ds|t_1,x).$$

Hence, the hazard rate of F_{21} equals

$$d\Lambda_{21}(t|t_1,x) = \frac{F_{21}(dt|t_1,x)}{1 - F_{21}(t^-|t_1,x)} = \frac{\tilde{H}^*(dt|t_1,x)}{1 - H^*(t^-|t_1,x)}.$$

Finally, we obtain

$$F_{21n}(t_2|t_1,x) = 1 - \prod_{i=1}^{n} \left[1 - \frac{\tilde{B}_{in}(x,t_1) 1_{\{\tilde{T}_{2i} \leq t_2\}} \delta_{2i}}{\sum_{j=1}^{n} 1_{\{\tilde{T}_{2j} \geq \tilde{T}_{2i}\}} \tilde{B}_{jn}(x,t_1)} \right],$$

the Beran estimator based on the restricted sample (that is, only the data with $\delta_1 = 1$), where

$$\tilde{B}_{in}(x,t_1) = \frac{\delta_{1i} K(\frac{x-X_i}{h_1}) K(\frac{t_1-\tilde{T}_{1i}}{h_2})}{\sum_{j=1}^{n} \delta_{1j} K(\frac{x-X_j}{h_1}) K(\frac{t_1-\tilde{T}_{1j}}{h_2})}$$

and h_1 and h_2 are bandwidth parameters.

3.3. Estimation of $\mathbf{F}(y_1, y_2|x)$

Following the same ideas as in van Keilegom (2004), we propose to estimate $\mathbf{F}$ by plugging in appropriate estimators of F_1 and F_{21}, that is,

$$\mathbf{F}_n(y_1,y_2|x) = \int_0^{y_1} F_{21n}(y_2|t_1,x) F_{1n}(dt_1|x), \tag{5}$$

where $F_{1n}(t_1|x)$ and $F_{21n}(t_2|t_1,x)$ have been previously introduced in Subsections 3.1 and 3.2.

In order to estimate $\mathbf{f}(y_1,y_2|x)$, the density function associated to the bivariate cdf $\mathbf{F}(y_1,y_2|x)$, we consider an alternative way to write $\mathbf{F}_n(y_1,y_2|x)$ as follows

$$\mathbf{F}_n(y_1,y_2|x) = \sum_{i=1}^{n} W_{1i}(x) 1_{\{\tilde{T}_{1i} \le y_1\}} F_{21n}(y_2|\tilde{T}_{1i},x), \tag{6}$$

where

$$W_{1i}(x) = F_{1n}(\tilde{T}_{1i}|x) - F_{1n}(\tilde{T}_{1i}^-|x)$$

and we set the following weights

$$W_{ik}^B(x) = \mathbf{F}_n(\tilde{T}_{1i},\tilde{T}_{2k}|x) - \mathbf{F}_n(\tilde{T}_{1i}^-,\tilde{T}_{2k}|x) - \mathbf{F}_n(\tilde{T}_{1i},\tilde{T}_{2k}^-|x) + \mathbf{F}_n(\tilde{T}_{1i}^-,\tilde{T}_{2k}^-|x).$$

Taking into account that

$$W_{ik}^B(x) = W_{1i}(x) W_{2ki}(x),$$

where

$$W_{2ki}(x) = F_{2n}(\tilde{T}_{2k}|\tilde{T}_{1i},x) - F_{2n}(\tilde{T}_{2k}^-|\tilde{T}_{1i},x),$$

we propose to estimate $\mathbf{f}(y_1,y_2|x)$, the density function associated to the bivariate cdf $\mathbf{F}(y_1,y_2|x)$, by

$$\mathbf{f}_n(y_1,y_2|x) = \frac{1}{h_2 h_3} \sum_{i=1}^{n} \sum_{k=1}^{n} W_{ik}^B(x) K\left(\frac{y_1-\tilde{T}_{1i}}{h_2}\right) K\left(\frac{y_2-\tilde{T}_{2k}}{h_3}\right), \tag{7}$$

where h_3 is an additional bandwidth parameter.

Before we state our first theorem, we set

$$H(t|x) = \mathbb{P}(\tilde{T}_1 \le t | X = x),$$

and define $\tau_1(x)$ and $\tau_2(t_1,x)$ as follows

$$0 \leq \tau_1(x) < \inf\{t : H(t|x) = 1\} < \infty, \quad \tau_2(t_1,x) < \inf\{t : H^*(t|t_1,x) = 1\}.$$

Set

$$A(x) = \{(t_1,t_2) : t_1 \leq \tau_1(x), t_2 \leq \inf_{t \leq t_1} \tau_2(t,x)\}.$$

Theorem 3.1. *Let $(y_1,y_2) \in A(x)$ and $x \in \{u : f_X(u) > 0\}$, where f_X denotes the density of X. Under assumption A1, if $nh_1^5 \to 0$, $nh_1^3 \to \infty$ and $nh_2^5 \to c > 0$, we have*

$$\sqrt{nh_1}(\mathbf{F}_n(y_1,y_2|x) - \mathbf{F}(y_1,y_2|x)) \to \mathcal{N}(0, \sigma_1^2(y_1,y_2|x)),$$

where

$$\sigma_1^2(y_1,y_2|x) = \frac{\rho_1^2(y_1,y_2,x) \int K^2(t)dt}{f_X(x)}$$

and $\rho_1^2(y_1,y_2,x)$ is given in (A.12).

Proof. See Appendix A.

Remark 3.2. *The variance $\sigma_1^2(y_1,y_2|x)$ has a complicated structure since it reflects variability of three leading terms in the expansion of the process from Theorem 3.1 and hence in practical applications it needs to be approximated using bootstrap or jackknife. However, when no censoring is present, it reduces to the well known expression with $\rho_1^2(y_1,y_2,x) = \mathbf{F}(y_1,y_2|x)(1-\mathbf{F}(y_1,y_2|x))$.*

Remark 3.3. *The estimator $\mathbf{F}_n(y_1,y_2|x)$ given in (6) can be extended to the case where T_1 is, additionally to being censored from the right, left truncated by a random variable Z, independent of T_1 and T_2. In such a case, we only need to replace F_{1n} by the estimator introduced by Iglesias-Pérez and González-Manteiga (1999).*

4. Conditional marginal estimators

Our goal in this section is to estimate $F_1(t|x)$ and $F_2(t|x)$, where F_i denotes the conditional distribution function of T_i given $X = x$, for $i = 1,2$. Remark that $F_1(t|x)$ can be consistently estimated by a standard conditional Beran estimator as given in (4). See Beran (1981) or González-Manteiga and Cadarso-Suárez (1994), for details. Regarding $F_2(t|x)$, due to induced dependent censoring, we cannot use the standard Beran estimator. So we propose a new estimator derived from the proposed bivariate estimator $\mathbf{F}_n(t_1,t_2|x)$. We should recall, however, that we have the asymptotic properties of this estimator only for $(t_1,t_2) \in A(x)$. Hence, similarly to van Keilegom (2004), we define

$$F_2^{\tau_1}(t|x) = \mathbf{F}(\tau_1(x),t|x)$$

that, under the assumption $\inf\{t : F_1(t|x) = 1\} \leq \inf\{t : G(t|x) = 1\}$, can be made arbitrarily close to $F_2(t|x)$. Consequently, we define the following consistent estimator

$$F_{2n}^{\tau_1}(t|x) = \mathbf{F}_n(\tau_1(x), t|x).$$

Since, as mentioned before, the estimation of $F_1(t|x)$ has been widely studied in the literature, we focus in our simulation study on the new estimator that we have proposed for $F_2(t|x)$. In order to measure the performance of this conditional estimator, we consider the Kolmogorov-Smirnov (KS) distance, that is,

$$KS^B(x) = \sup_t |F_{2n}^{\tau_1}(t|x) - F_2^{\tau_1}(t|x)|. \tag{8}$$

5. Likelihood based bandwidth selection

In this section, we derive the likelihood function for two consecutive censored times, T_1 and T_2. To give some insights, we will consider first the case of two independent censored times C_1 and C_2, so that T_1 is censored by C_1 and T_2 is censored by C_2. For this case, the likelihood function is given by

$$\begin{aligned}\ell(\gamma) = \sum_{i=1}^{n} \Big[& \delta_{1i}\delta_{2i} \log(\mathbf{f}(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) + \delta_{1i}(1-\delta_{2i}) \log(g_1(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) \\ & + \delta_{2i}(1-\delta_{1i}) \log(g_2(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) + (1-\delta_{1i})(1-\delta_{2i}) \log(\bar{\mathbf{F}}(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) \Big],\end{aligned}$$

where the density and cdf of (T_1, T_2) given X are assumed to be known up to some parameter γ, and

$$\begin{aligned} g_1(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i) &= \int_{\tilde{T}_{2i}}^{\infty} \mathbf{f}(\tilde{T}_{1i}, t|X_i)dt, \\ g_2(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i) &= \int_{\tilde{T}_{1i}}^{\infty} \mathbf{f}(t, \tilde{T}_{2i}|X_i)dt, \\ \bar{\mathbf{F}}(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i) &= \mathbb{P}(T_1 > \tilde{T}_{1i}, T_2 > \tilde{T}_{2i}|X = X_i). \end{aligned}$$

Taking into account that in our setup there is one common censoring time C so that $(1-\delta_{1i})\delta_{2i} = 0$ and if $1-\delta_{1i} = 1$, then $1-\delta_{2i} = 1$ and $\tilde{T}_{2i} = 0$, we have that $\bar{\mathbf{F}}(\tilde{T}_{1i}, 0|X_i) = 1 - F_1(\tilde{T}_{1i}|X_i)$ and the likelihood function is reduced to

$$\begin{aligned}\ell(\gamma) = \sum_{i=1}^{n} \Big[& \delta_{1i}\delta_{2i} \log(\mathbf{f}(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) + \delta_{1i}(1-\delta_{2i}) \log(g_1(\tilde{T}_{1i}, \tilde{T}_{2i}|X_i)) \\ & + (1-\delta_{1i}) \log(1 - F_1(\tilde{T}_{1i}|X_i)) \Big],\end{aligned} \tag{9}$$

which is a generalization of the discrete case covered in Visser (1996).

Denoting $\mathbb{K}(y) = \int_y^\infty K(z)dz$ and replacing the unknown quantities in (9) by estimators involving three bandwidth parameters, h_1, h_2 and h_3, we derive an estimated log-likelihood function in terms of $\gamma = (h_1, h_2, h_3)$ to work with. Specifically, we denote by $\ell_n(h_1, h_2, h_3)$ the estimated log-likelihood function where $\mathbf{f}$ is estimated by $\mathbf{f}_n$ given in (7) and g_1 is estimated by

$$g_{1n}(y_1, y_2|x) = \frac{1}{h_2} \sum_{i=1}^{n} \sum_{k=1}^{n} W_{ik}^B(x) K\left(\frac{y_1 - \tilde{T}_{1i}}{h_2}\right) \mathbb{K}\left(\frac{y_2 - \tilde{T}_{2k}}{h_3}\right).$$

In order to select the bandwidth parameters in practice, we propose to maximize the estimated log-likelihood function $\ell_n(h_1, h_2, h_3)$. The bandwidth selection is a problematic issue already in the standard Beran estimator (with one variable of interest and one dimensional covariate). An alternative to the proposed method would be to minimize the integrated square error (w.r.t (t_1, t_2, x)). However, our method is computationally much faster, it does not require numerical integration and it provides an estimate of the corresponding bivariate conditional density as a byproduct. Moreover, it can be easily adapted to multidimensional covariates by applying the so called single-index model (see, e.g., Strzalkowska-Kominiak and Cao (2013) for a single-index model under one-dimensional censoring variable).

6. Simulation study

A simulation study is carried out here to check the performance of the new estimator, previously proposed in (5). The scenarios that we consider are based on Huang, Luo and Follmann (2011). More specifically, we start by generating two exponential random variables, (T_1^0, T_2^0), with unit means, linked by a Gaussian copula. This can be done by following the steps:

- Generating n i.i.d. random variables, $A_i \sim \mathcal{N}(0, \sqrt{\rho})$ and $B_{ki} \sim \mathcal{N}(0, \sqrt{1-\rho})$, for $k = 1, 2$ and $i = 1, ..., n$.
- Setting $T_{ki}^0 = -\ln(1 - \Phi(A_i + B_{ki}))$, where Φ is the cumulative distribution function of a standard normal variable.

Since $(A + B_1, A + B_2)$ follows a standardized bivariate normal variable with correlation ρ, it is easy to prove that

$$\mathbf{F}^0(t_1, t_2) = \mathbb{P}(T_1^0 \leq t_1, T_2^0 \leq t_2) = \Phi_2(\Phi^{-1}(1 - e^{-t_1}), \Phi^{-1}(1 - e^{-t_2})),$$

where Φ_2 denotes the cdf of $(A + B_1, A + B_2)$. Taking into account Sklar's theorem (see Sklar (1959)), it is easy to see that

$$\mathbf{F}^0(t_1, t_2) = \mathbf{C}(F_1^0(t_1), F_2^0(t_2)),$$

where $\mathbf{C}$ refers to the bivariate Gaussian copula with parameter ρ and F_k^0 denotes the cdf of T_k^0, for $k = 1, 2$.

To incorporate the effect of the covariate X in the gap times (T_1, T_2), we define $T_k = \left(T_k^0/a(X)\right)^{1/\kappa}$, for $k = 1,2$, where $a(X) = \exp(\beta X)$. It is easy to check that, conditionally on $X = x$, T_k, for $k = 1,2$, is a Weibull distributed random variable with conditional cdf $F_k(t|x) = 1 - e^{-(t/\lambda)^\kappa}$, for $t > 0$, where the scale parameter equals $\lambda = (1/a(x))^{1/\kappa}$ and the shape parameter equals κ. Observe that the conditional hazard rate of this model is given by $h(t|x) = a(x)\kappa t^{\kappa-1}$ with $a(x) = \exp(\beta x)$. Hence β refers to the parameter of the Cox model with basic hazard defined as $h_0(t) = \kappa t^{\kappa-1}$.

In order to get a copula representation of the conditional cdf of (T_1, T_2) given $X = x$, we use the fact that

$$\begin{aligned}\mathbf{F}(t_1,t_2|x) &= \mathbb{P}\left(T_1^0 \le a(X)t_1^\kappa, T_2^0 \le a(X)t_2^\kappa | X = x\right) = \mathbf{F}^0(a(x)t_1^\kappa, a(x)t_2^\kappa)\\ &= \Phi_2(\Phi^{-1}(1 - e^{-a(x)t_1^\kappa}), \Phi^{-1}(1 - e^{-a(x)t_2^\kappa}))\\ &= \Phi_2(\Phi^{-1}(F_1(t_1|x)), \Phi^{-1}(F_2(t_2|x))),\end{aligned}$$

where $F_k(t|x)$ denotes the cdf of T_k given $X = x$, for $k = 1,2$.

Therefore, this procedure leads us to two Weibull variables, T_1 and T_2, linked through the bivariate Gaussian copula. In order to get more flexibility, other copula functions could be used to link T_1 and T_2. In that case, the conditional cdf of (T_1, T_2) given $X = x$ would be as follows $\mathbf{F}(t_1,t_2|x) = \mathbf{C}(F_1(t_1|x), F_2(t_2|x))$, where $\mathbf{C}$ denotes a bivariate copula function.

In Subsection 6.1 we consider the Gaussian copula with fixed parameter $\rho = 0.5$ and variable parameter $\rho = \rho(x) = x/10$. Furthermore, we study in Subsection 6.2 the Clayton copula with parameters $\theta = 0.5$ and $\theta = 5$. Additionally, we consider a cure rate model in Subsection 6.3, where $\mathbf{C}$ denotes a bivariate Gaussian copula with parameter $\rho = 0.5$ and T_2 given $X = x$ comes from an improper distribution $F_2^0(t_2|x) = pF_2(t_2|x)$ with $p = 0.8$. In order to generate values from these copulas see, for example, Cherubini, Luciano and Vecchiato (2004) and Wu, Valdez and Sherris (2007).

Regarding the marginals, we consider $\beta = 0.3$ and three different values for $\kappa = 1.5, 2, 3$. We investigate the behaviour of the estimator under two different scenarios for the distribution of the covariate, namely $X \sim U(0,10)$ and $X \sim \mathcal{N}(5,1)$. Furthermore, $C \sim U(0, \tau_c)$, where we choose two values for τ_c, such that the proportion of subjects with zero events is approximately 10% and 15%, that is, there are $0.1n$ and $0.15n$ subjects with $\delta_{1i} = 0$ and $\delta_{2i} = 0$. Moreover, for those τ_c we have, respectively, 10% and 15% events for which $\delta_{1i} = 1$ but $\delta_{2i} = 0$. Additionally, for the cure rate model, we choose $\tau_c = 5$ so that for $p = 0.8$, we have 9% $\delta_{1i} = \delta_{2i} = 0$ and 25% events with $\delta_{1i} = 1$ but $\delta_{2i} = 0$. In the following subsections, we present the simulation results for three different models.

To simplify the likelihood based selection of the three dimensional bandwidth vector, we consider that $h_j = c_j h$, for $j = 1,2,3$ and maximize the likelihood function over one parameter h. The most natural choice for the constants c_j is $c_1 = \hat{\sigma}(X)$, $c_2 = \hat{\sigma}(\tilde{T}_1)$ and $c_3 = \hat{\sigma}(\tilde{T}_2)$, where $\hat{\sigma}$ denotes the sample standard deviation.

This simulation study is carried out in the open-source software R and shows the performance of our proposed estimator in terms of bias, variance and mean squared

error (MSE) that are calculated by resampling using 200 trials. Additionally, we check the estimator of the conditional marginal, $F_2(t|x)$, with the Kolmogorov-Smirnov (KS) distance defined in (8) with $\tau_1 = +\infty$.

6.1. Gaussian copula model

In this subsection, we use the Gaussian copula with a constant parameter $\rho = 0.5$. We access the quality of the estimation for different values of the parameter $\kappa \in \{1.5, 2, 3\}$ and $X \sim U(0,10)$ (Tables 1-3). Furthermore, we check the behaviour of our estimation procedure when $X \sim \mathcal{N}(5,1)$ for $\kappa = 2$ (Tables 4-5). For a given x, the estimators are computed in the middle point of the support for $(t_1,t_2) = (E(T_1),E(T_2))$ and in the right side of the support $(t_1,t_2) = (1,1)$ using 200 trials. In the case of $X \sim U(0,10)$ we choose $x = 5$ and when $X \sim \mathcal{N}(5,1)$ we investigate additionally the behaviour in low density regions for $x = 3$. In continuation, we use a dependent Gaussian copula model with parameter $\rho = \rho(X) = X/10$ and $X \sim U(0,10)$ so that not only the marginals depend on the covariate, but also the correlation between T_1 and T_2 changes with x. The estimators are computed in $(t_1,t_2,x) = (E(T_1),E(T_2),5)$ so that $\rho(x) = 0.5$ and in $(t_1,t_2,x) = (E(T_1),E(T_2),8)$ so that $\rho(x) = 0.8$ (see Table 6).

Table 1. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\rho = 0.5$, $\kappa = 1.5$, $X \sim U(0,10)$ *and* $n = 100, 200$.

(t_1,t_2,x)		$(0.39,0.39,5)$ $\mathbf{F}(0.39,0.39\|5) \approx 0.51$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	3.9	0.0077	0.006	0.0061	-0.0186	8e-04	0.0012	0.1375
	2.5	0.0082	0.0062	0.0063	-0.0219	0.0011	0.0016	0.1459
200	3.9	-0.0066	0.0043	0.0043	-0.0173	0.0006	0.0009	0.1088
	2.5	-0.0061	0.0042	0.0042	-0.0157	0.0006	0.0009	0.1131

Table 2. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\rho = 0.5$, $\kappa = 2$, $X \sim U(0,10)$ *and* $n = 100, 200$.

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.45$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	4.6	-0.0010	0.0064	0.0064	-0.0198	0.0009	0.0013	0.1456
	3.1	-0.0138	0.0060	0.0062	-0.0229	0.0010	0.0015	0.1468
200	4.6	-0.0031	0.0037	0.0038	-0.0144	0.0004	0.0006	0.1076
	3.1	-0.0027	0.0037	0.0037	-0.0145	0.0006	0.0008	0.1123

Table 3. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\rho = 0.5$, $\kappa = 3$, $X \sim U(0,10)$ *and* $n = 100, 200$.

(t_1,t_2,x)		$(0.56,0.56,5)$ $\mathbf{F}(0.56,0.56\|5) \approx 0.38$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	5.1	-0.0009	0.0046	0.0046	-0.0212	0.001	0.0014	0.1394
	3.9	-0.0031	0.0057	0.0057	-0.022	0.0011	0.0016	0.1529
200	5.1	0.0014	0.0033	0.0033	-0.0133	0.0005	0.0006	0.1126
	3.9	-0.0084	0.0032	0.0033	-0.0146	0.0006	0.0008	0.1126

Table 4. *Bias, variance and MSE in* (t_1,t_2,x) *with* $x = 5$ *when* $\rho = 0.5$, $\kappa = 2$, $X \sim \mathcal{N}(5,1)$ *and* $n = 100, 200$.

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.45$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	4.6	-0.0150	0.0037	0.0039	-0.0008	0.0004	0.0004	0.1152
	3.1	-0.0112	0.0039	0.0040	-0.0013	0.0005	0.0005	0.1171
200	4.6	-0.0120	0.0024	0.0026	-0.0029	0.0003	0.0003	0.0875
	3.1	-0.0120	0.0020	0.0021	-0.0027	0.0003	0.0003	0.0929

Table 5. *Bias, variance and MSE in* (t_1,t_2,x) *with* $x = 3$ *when* $\rho = 0.5$, $\kappa = 2$, $X \sim \mathcal{N}(5,1)$ *and* $n = 100, 200$.

(t_1,t_2,x)		$(0.46,0.46,3)$ $\mathbf{F}(0.46,0.46\|3) \approx 0.24$			$(1,1,3)$ $\mathbf{F}(1,1\|3) \approx 0.86$			x=3
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	4.6	0.0417	0.0121	0.0138	0.0528	0.0076	0.0104	0.2319
	3.1	0.0225	0.0120	0.0125	0.0414	0.0076	0.0093	0.2395
200	4.6	0.0293	0.0068	0.0076	0.0401	0.0055	0.0071	0.1821
	3.1	0.0450	0.0079	0.0099	0.0454	0.0057	0.0077	0.2009

As can be seen in Tables 1-4, our proposed estimator gives very good results. There is no significant difference in the performance of the estimator for different values of κ nor it changes with change of the distribution of the covariate when $x = 5$ is considered. Introducing a dependent copula model (Table 6) also does not affect negatively the results. As expected the results improve when increasing the sample size and surprisingly they are not very affected by increasing the censoring rate. Additionally, even though the asymptotic properties were proved for compact sets, our new method gives good results even for (t_1,t_2) being near to the right-hand side of the support. The only exception are the low density regions with normal covariate (Table 5), were the quality of estimation

declines. This is, however, not surprising since this type of behaviour we observe also with standard Beran estimator.

Table 6. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\rho = x/10$, $\kappa = 2$, $X \sim U(0,10)$ *and* $n = 100,200$.

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.45$			x=5	$(0.46,0.46,8)$ $\mathbf{F}(0.46,0.46\|8) \approx 0.86$			x=8
n	τ_c	Bias	Variance	MSE	KS	Bias	Variance	MSE	KS
100	4.6	-0.0072	0.0056	0.0056	0.1371	-0.0480	0.0038	0.0061	0.1511
	3.1	-0.0040	0.0066	0.0066	0.1468	-0.0437	0.0035	0.0054	0.1525
200	4.6	-0.0051	0.0034	0.0034	0.1098	-0.0270	0.0022	0.0029	0.1093
	3.1	-0.0102	0.0036	0.0037	0.1115	-0.0339	0.0025	0.0037	0.1119

6.2. Clayton copula model

In this subsection, we use the Clayton copula with a constant parameter $\theta = 0.5$ and $\theta = 5$. As in Subsection 6.1, the estimators are computed in the middle point, $(t_1,t_2,x) = (E(T_1),E(T_2),5)$, where $E(T_1) = E(T_2) = 0.46$ and $\mathbf{F}(0.46,0.46|5) \approx 0.41$, and in the right side of the support $(1,1,5)$, where $\mathbf{F}(1,1|5) \approx 0.98$ (see Tables 7 and 8). From these tables, we observe that the behaviour of our estimator is similar to the case of the Gaussian copula model.

Table 7. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\theta = 0.5$, $\kappa = 2$, $X \sim U(0,10)$ *and* $n = 100,200$.

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.41$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	4.6	0.0038	0.0055	0.0055	-0.0218	0.0009	0.0014	0.1355
	3.1	0.0174	0.0055	0.0058	-0.0250	0.0011	0.0017	0.1420
200	4.6	0.0054	0.0032	0.0033	-0.0176	0.0004	0.0007	0.1085
	3.1	0.0072	0.0030	0.0031	-0.0200	0.0006	0.0010	0.1099

Table 8. *Bias, variance and MSE in* (t_1,t_2,x) *when* $\theta = 5$, $\kappa = 2$, $X \sim U(0,10)$ *and* $n = 100,200$.

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.41$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.98$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	4.6	-0.0259	0.0079	0.0086	-0.0131	0.0010	0.0012	0.1568
	3.1	-0.0237	0.0068	0.0074	-0.0144	0.0009	0.0011	0.1541
200	4.6	-0.0140	0.0043	0.0045	-0.0089	0.0005	0.0006	0.1252
	3.1	-0.0080	0.0055	0.0056	-0.0097	0.0006	0.0007	0.1298

6.3. Cure rate model

In this subsection, we use the Gaussian copula with a constant parameter $\rho = 0.5$. The distribution function of T_2 given $X = x$ is improper, that is $F_2^0(t_2|x) = pF_2(t_2|x)$ with $p = 0.8$. The estimators are computed in the middle point, $(t_1, t_2, x) = (0.46, 0.46, 5)$, where $\mathbf{F}(0.46, 0.46|5) \approx 0.38$ and in the right side of the support $(1,1,5)$, where $\mathbf{F}(1,1|5) \approx 0.79$ (see Table 9).

Table 9. *Bias, variance and MSE in* (t_1, t_2, x) *when* $p = 0.8$, $\theta = 0.5$ *and* $n = 100$

(t_1,t_2,x)		$(0.46,0.46,5)$ $\mathbf{F}(0.46,0.46\|5) \approx 0.38$			$(1,1,5)$ $\mathbf{F}(1,1\|5) \approx 0.79$			x=5
n	τ_c	Bias	Variance	MSE	Bias	Variance	MSE	KS
100	5	-0.0190	0.0029	0.0033	-0.0312	0.0026	0.0036	0.1123

As can be seen in Table 9, when one of the marginal distributions is improper, our new estimator gives very good results already for $n = 100$. However, from a theoretical point of view, those models are out of scope of the present paper.

7. Examples

7.1. Stanford heart transplant data

In this section, we analyze the example of Stanford heart transplant data, previously introduced in Section 1. There are 103 individuals in this dataset, 45 out of them received transplant ($delta = 1$) and died ($status = 1$), 24 received transplant ($delta = 1$) and were still alive at the end of the study ($status = 0$), for 4 patients the study ended before the transplantation ($delta = 0$ and $status = 0$) and the remaining 30 died before receiving the transplant ($delta = 0$ and $status = 1$). In this real example, the death before receiving the transplant can be considered as a semi-competing risk because it is a termination event that can potentially censor the non terminating event of receiving the transplant (see Zhao and Zhou (2010), among others). In fact, this is the case for those 30 patients that have $delta = 0$ and $status = 1$. Considering that T_3 denotes the time from acceptance into the transplantation program to death (in months), it is easy to adapt our model to this situation by replacing C by $C_1 = \min(C, T_3)$. Under this setting, we set $\delta_1 = delta$ and $\delta_2 = status \times delta$. For the sake of illustration of our methodology, we consider the covariate $X = age$.

As in the previous section, we consider that $h_j = c_j h$, for $j = 1,2,3$ and maximize the likelihood function over one parameter h. As before the constants c_j are defined as follows $c_1 = \hat{\sigma}(X)$, $c_2 = \hat{\sigma}(\tilde{T}_1)$ and $c_3 = \hat{\sigma}(\tilde{T}_2)$, where $\hat{\sigma}$ denotes the sample standard deviation. Using this approach, we obtain the following bandwidth $h = (6.29, 1.57, 10.66)$.

A graphical representation of $\mathbf{F}_n$ is given in Figure 2, with $x = 45$ (the sample mean) and $x = 55$, respectively. In Tables 10 and 11, we collect the estimated $\mathbf{F}(t_1, t_2|x)$ for a

fixed x and $(t_1,t_2) \in (1,2,3,4) \times (12,24,36,48,60)$ together with the 95% confidence intervals based on the asymptotic normality proved in Theorem 1, where the bootstrap technique has been used to estimate the standard deviation. In order to show the influence of the covariate X, we also present in Figure 2 the estimator $\mathbf{F}_n^K$ introduced by van Keilegom (2004) which does not take into account the covariate, X. Additionally, for different values of x, Figure 3 shows the estimator of $F_2(t|x)$ derived from $\mathbf{F}_n$ and the standard Beran estimator based on the reduced sample (that is, the individuals for which $\delta_1 = 1$).

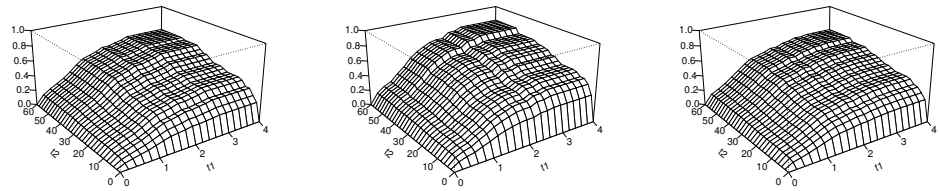

Figure 2. *Estimator of* $\mathbf{F}(t_1,t_2|x)$ *for* $x = 45$ *(left-hand panel) and* $x = 55$ *(middle panel) together with* $\mathbf{F}_n^K(t_1,t_2)$ *(right-hand panel).*

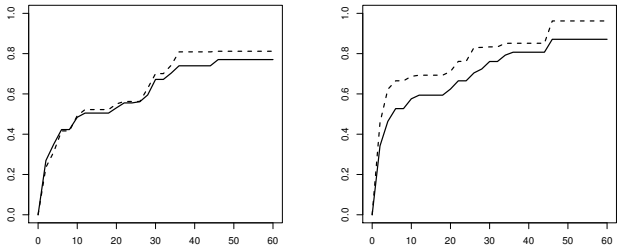

Figure 3. $F_{2n}^{\infty}(t|x)$ *(solid line) and standard Beran estimator on reduced sample (dahed line) where* $x = 45$ *(left-hand panel) and* $x = 55$ *(right-hand panel)*

Table 10. *Estimated* $\mathbf{F}(t_1,t_2|x)$ *for* $x = 45$.

t_1 \ t_2	12	24	36	48	60
1	0.2331	0.2548	0.3314	0.3405	0.3405
	(0.1160,0.3505)	(0.1304,0.3794)	(0.1994,0.4637)	(0.2112,0.4702)	(0.2112,0.4702)
2	0.3847	0.4211	0.5474	0.5628	0.5628
	(0.2516,0.5466)	(0.2847,0.5891)	(0.4302,0.7055)	(0.4592,0.7086)	(0.4592,0.7086)
3	0.4513	0.4944	0.6426	0.6610	0.6610
	(0.2999,0.6029)	(0.3394,0.6497)	(0.5175,0.7679)	(0.5573,0.7648)	(0.5573,0.7648)
4	0.4516	0.4948	0.6431	0.6614	0.6614
	(0.2998,0.6035)	(0.3393,0.6504)	(0.518,0.7683)	(0.5577,0.7653)	(0.5577,0.7653)

Table 11. *Estimated* $\mathbf{F}(t_1,t_2|x)$ *for* $x = 55$.

t_1 \ t_2	12	24	36	48	60
1	0.2837	0.3176	0.3776	0.4111	0.4111
	(0.1553,0.4127)	(0.1796,0.4563)	(0.2284,0.5275)	(0.2683,0.5548)	(0.2683,0.5548)
2	0.4705	0.5277	0.6247	0.6805	0.6805
	(0.3167,0.6414)	(0.3686,0.7057)	(0.4702,0.801)	(0.5541,0.8311)	(0.5541,0.8311)
3	0.5537	0.6215	0.7335	0.7994	0.7994
	(0.3894,0.7192)	(0.4534,0.7909)	(0.5776,0.8908)	(0.6871,0.9133)	(0.6871,0.9133)
4	0.5537	0.6215	0.7335	0.7994	0.7994
	(0.3894,0.7193)	(0.4533,0.791)	(0.5776,0.8908)	(0.6871,0.9133)	(0.6871,0.9133)

We can see, in Figure 2, that the bivariate distribution changes with the covariate, X. Figure 3 shows even more clearly that the age has a big influence on the distribution function. Specifically, the probability to fail increases as the age increases. This is a quite natural effect. Moreover, both from the figures and tables, we observe that the probabilities do not change for $t_2 \in [48,60]$, which is equivalent to the 4th and 5th year after transplantation. Interestingly, based on Figure 3, the probability of death is growing rapidly in the first months after heart transplantation and stabilizes by the time of 4 years after the surgery. Finally, in Figure 3, we can see that when applying the standard Beran estimator on this reduced sample, we get similar results to our marginal estimator $F_{2n}^{\infty}(t|x)$ when $x = 45$ but this Beran estimator seems to overestimate the distribution when $x = 55$.

7.2. Colon cancer data

In this subsection, we analyze the colon cancer data, previously introduced in Section 1. We investigate the effectiveness of treatment with levamisole plus 5-FU versus placebo as well as influence of the age on the survival times. The data set is a part of the data set "colon" in the R survival library. Similarly as Lawless and Yilmaz (2011) we consider those patients treated with Levamisole plus 5-FU and placebo control. The full data set includes a third treatment group (Levamisole). There were 315 patients assigned to the placebo control group and 304 to the treatment group. By the end of the study, 177 patients (56%) in the placebo group had cancer recurrence, among whom 155 died, whereas in the treatment group 119 (39%) patients had cancer recurrence, among whom 108 died. The maximal observed times until recurrence and from recurrence until death are around 9 and 6 years, respectively. We consider age as a covariate. In Figures 4-6 we plot the bivariate and univariate estimators for both groups and 50, 60, and 70 years old patients, being the respective quartiles in the data set. Figure 7 shows the estimator, $F_{2n}^{\infty}(t|x)$, for the treatment group together with bootstrap-based and normal 95% point wise confidence intervals, respectively.

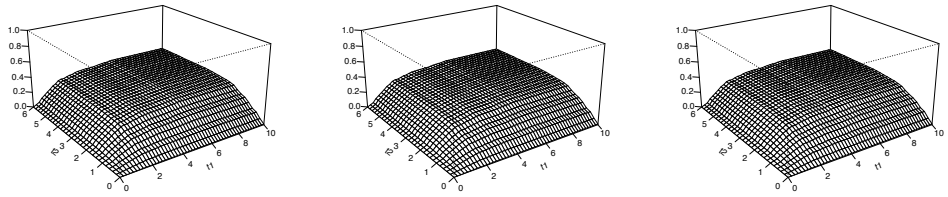

Figure 4. *Estimator of* $\mathbf{F}(t_1,t_2|x)$ *in treatment group for* $x = 50$ *(left-hand panel),* $x = 60$ *(middle panel) and* $x = 70$ *(right-hand panel).*

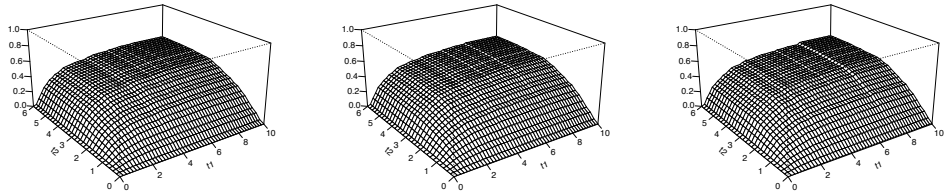

Figure 5. *Estimator of* $\mathbf{F}(t_1,t_2|x)$ *in placebo group for* $x = 50$ *(left-hand panel),* $x = 60$ *(middle panel) and* $x = 70$ *(right-hand panel).*

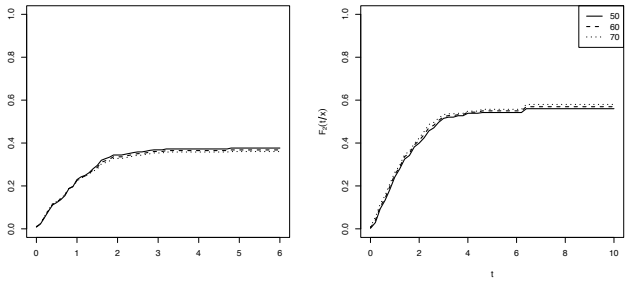

Figure 6. $F_{2n}^{\infty}(t|x)$ *for* $x = 50,60,70$ *for treatment group (left-hand panel) and placebo group (right-hand panel).*

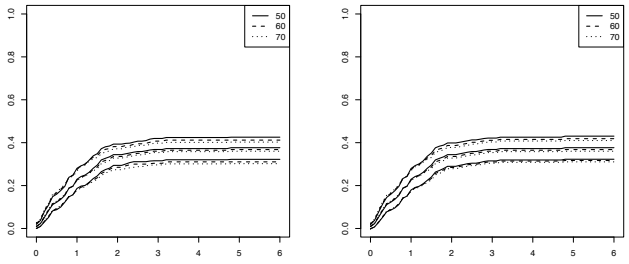

Figure 7. $F_{2n}^{\infty}(t|x)$ *for* $x = 50,60,70$ *for treatment group with bootstrap 95% CIs (left-hand panel) and normal 95% CIs (right-hand panel).*

As expected, treatment seems to have big influence on patients survival, increasing the survival time from recurrence. However, based on the confidence intervals for the treatment group, the influence of the age seems to be non significative although a formal test should be developed to confirm this statement.

8. Discussion and future work

In this paper we have introduced a new method to estimate the bivariate conditional distribution of two consecutive censored gap times and the corresponding marginal distributions. This new methodology is an adaptation and a mixture of the methods proposed by Beran (1981) and van Keilegom (2004). It is worth mentioning that a simpler method could be the extension of the Kaplan-Meier based estimator studied by de Uña-Álvarez and Meira-Machado (2008). Although our method is computationally more intensive compared to the Kaplan-Meier based estimator because it requires bootstrapping to estimate the variance, it has an important advantage. While the Kaplan-Meier based estimator requires the assumption of independence between (T_1, T_2) and C, our method only requires the independence given X (allowing some dependence between C and the explanatory variable X). Based on this weaker assumption, we have proved its asymptotic theoretical properties and studied its finite sample behaviour through a simulation study. Additionally, our approach can be extended to a d-dimensional explanatory variable X by using a single-index model (see Strzalkowska-Kominiak and Cao (2013)) avoiding the curse of dimensionality. However, this issue goes beyond the scope of this paper and will be the basis of our future research.

9. Acknowledgments

E.M. Molanes-López acknowledges support to Grant PID2019-106772RB-I00 from the Spanish Government. E. Strzalkowska-Kominiak acknowledges support to Grant PID2022-138114NB-I00 from the Spanish Government. We thank the editor and two referees for their valuable comments that improved our article.

References

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report*, University California, Berkeley.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). Copula methods in finance. New York: John Wiley & Sons.

de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal* 53, 113–127.

de Uña-Álvarez, J. and Meira-Machado, L.F. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters* 78, 2440–2445.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* 33, 1380–1403.

Fine, J. P., Jiang, H. and Chappell, R. (2001). On semi-competing risks data. *Biometrika* 88(4), 907–919.

Gijbels, I., Veraverbeke, N. and Omelka, M. (2011). Conditional copulas, association measures and their application. *Computational Statistics and Data Analysis* 55, 1919–1932.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Journal of Nonparametric Statistics* 4, 65–78.

Huang, Y. and Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* 85(4), 785–798

Huang, C.-Y., Luo, X. and Follmann, D.A. (2011). A model checking method for the proportional hazards model with recurrent gap time data. *Biostatistics* 12, 535–547.

Iglesias-Pérez, C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics* 10, 213–244.

Kalbfleisch, J.D. and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data, New York: John Wiley & Sons, Inc.

Lawless, J. F. and Yilmaz, Y. E. (2011). Semiparametric estimation in copula models for bivariate sequential survival times. *Biometrical journal* 53(5), 779–796.

Lin, D.Y., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika* 86, 59–70.

Meira-Machado, L. and Roca-Pardiñas, J. (2011). p3state.msm: Analyzing survival data from an illness-death model. *Journal of Statistical Software* 38, Issue 3.

Serrat, C. and Gómez, G. (2007). Nonparametric bivariate estimation for successive survival times. *SORT* 1, 75–96.

Sklar, A. (1959). Fonctions de répartion à n dimensions e leurs marges. *Publications de l'Institut de Statisque de l'Université de Paris* 8, 229–231.

Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Journal of Multivariate Analysis* 114, 74–98.

Strzalkowska-Kominiak, E. and Stute, W. (2010). The statistical analysis of consecutive survival data under serial dependence. *Journal of Nonparametric Statistics* 22, 585–597.

van Keilegom, I. (2004). A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *Journal of Nonparametric Statistics* 16, 659-670.

Visser, M. (1996). Nonparametric estimation of the bivariate survival function with an application to vertically transmitted. *Biometrika* 83, 507–518.

Wang, W. and Wells, M.T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85, 561–572.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65(1), 257–273.

Wu, F., Valdez, E. and Sherris, M. (2007). Simulating from exchangeable Archimedean copulas. *Communications in Statistics – Simulation & Computation* 36, 1019–1034.

Zhao, X. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics* 46, 290–299.

Zhu, H. and Wang, M. C. (2012). Analysing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika* 99(2), 345–361.

Appendix A: Asymptotic properties

Proof of Theorem 1. We have

$$\mathbf{F}_n(y_1,y_2|x)-\mathbf{F}(y_1,y_2|x)=A_n(y_1,y_2|x)+B_n(y_1,y_2|x)+C_n(y_1,y_2|x), \qquad (A.1)$$

where

$$\begin{aligned}
A_n(y_1,y_2|x)&=\int_0^{y_1}(F_{21n}(y_2|t_1,x)-F_{21}(y_2|t_1,x))F_1(dt_1|x)\\
B_n(y_1,y_2|x)&=\int_0^{y_1}F_{21}(y_2|t_1,x)(F_{1n}(dt_1|x)-F_1(dt_1|x))\\
C_n(y_1,y_2|x)&=\int_0^{y_1}(F_{21n}(y_2|t_1,x)-F_{21}(y_2|t_1,x))(F_{1n}(dt_1|x)-F_1(dt_1|x)).
\end{aligned}$$

We now deal with the first term, $A_n(y_1,y_2|x)$, in the right hand side of (A.1). Since, $F_{21n}(y_2|t_1,x)$ is a Beran estimator on the restricted sample, we can use the results from González-Manteiga and Cadarso-Suárez (1994). For $(y_1,y_2)\in A(x)$ and $x\in\{u:f_X(u)>0\}$, we obtain

$$F_{21n}(y_2|t_1,x)-F_{21}(y_2|t_1,x)=\sum_{i=1}^{n}\tilde{B}_{in}(x,t_1)\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)+R_n(y_2|t_1,x),$$

where

$$\frac{\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)}{1-F_{21}(y_2|t_1,x)}=\left[-\int_0^{\tilde{T}_{2i}\wedge y_2}\frac{d\tilde{H}^*(s|t_1,x)}{(1-H^*(s^-|t_1,x))^2}+\frac{1_{\{\tilde{T}_{2i}\le y_2,\delta_{2i}=1\}}}{1-H^*(\tilde{T}_{2i}^-|t_1,x)}\right].$$

To deal with $R_n(y_2|t_1,x)$, we need the properties of the estimators

$$H_n^*(t|t_1,x)=\sum_{i=1}^{n}1_{\{\tilde{T}_{2i}\le t\}}\tilde{B}_{in}(t_1,x),$$

$$\tilde{H}_n^*(t|t_1,x)=\sum_{i=1}^{n}\delta_{2i}1_{\{\tilde{T}_{2i}\le t\}}\tilde{B}_{in}(t_1,x),$$

where

$$\tilde{B}_{in}(t_1,x)=\frac{\frac{1}{nh_1h_2}\delta_{1i}K\left(\frac{x-X_i}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1i}}{h_2}\right)}{\frac{1}{nh_1h_2}\sum_{j=1}^{n}\delta_{1j}K\left(\frac{x-X_j}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1j}}{h_2}\right)}.$$

Remark, that

$$\tilde{B}_{in}(t_1,x)=\tilde{B}^1_{in}(t_1,x)+\tilde{B}^2_{in}(t_1,x), \tag{A.2}$$

where

$$\tilde{B}^1_{in}(t_1,x)=\frac{\frac{1}{nh_1h_2}\delta_{1i}K\left(\frac{x-X_i}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1i}}{h_2}\right)}{\tilde{h}^1(t_1,x)}$$

$$\tilde{B}^2_{in}(t_1,x)=\frac{\tilde{B}_{in}(t_1,x)}{\tilde{h}^1(t_1,x)}\left[\tilde{h}^1(t_1,x)-\frac{1}{nh_1h_2}\sum_{j=1}^{n}\delta_{1j}K\left(\frac{x-X_j}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1j}}{h_2}\right)\right]$$

and $\tilde{h}^1(t_1,x)=\tilde{h}(t_1|x)f_X(x)$ with $\tilde{h}(t_1|x)$ denoting the density of $\tilde{H}(t_1|x)=\mathbb{P}(\tilde{T}_1\leq t_1,\delta_1=1|X=x)$. Hence

$$H^*_n(t|t_1,x)=\sum_{i=1}^{n}1_{\{\tilde{T}_{2i}\leq t\}}\tilde{B}^1_{in}(t_1,x)+\sum_{i=1}^{n}1_{\{\tilde{T}_{2i}\leq t\}}\tilde{B}^2_{in}(t_1,x).$$

Moreover, since $\frac{nh_1h_2}{\log(n)}\to\infty$, using Theorem 1 and 2 together with Remark 8 from Einmahl and Mason (2005) and the Taylor expansion, we obtain

$$\sup_{t\in\mathbb{R},t_1\leq\tau_1(x),x\in I}|\sum_{i=1}^{n}1_{\{\tilde{T}_{2i}\leq t\}}\tilde{B}^1_{in}(t_1,x)-H^*(t|t_1,x)|=O\left(\sqrt{\frac{\log(n)}{nh_1h_2}}\right)+O(h_1^2)+O(h_2^2)$$

almost surely (a.s.) and

$$\sup_{t\in\mathbb{R},t_1\leq\tau_1(x),x\in I}|\sum_{i=1}^{n}1_{\{\tilde{T}_{2i}\leq t\}}\tilde{B}^2_{in}(t_1,x)|=O\left(\left(\frac{\log(n)}{nh_1h_2}\right)^{1/2}\right)+O(h_1^2)+O(h_2^2),\ \text{a.s.}$$

where $I=\{u:f_X(u)>0\}$. Similarly, we deal with $\tilde{H}^*_n(t|t_1,x)$. Finally, following the steps of the proof of Theorem 2.3 in González-Manteiga and Cadarso-Suárez (1994), and since $H^*_n(t|t_1,x)\to H^*(t|t_1,x)$ in probability, we can show that

$$\sup_{(t_1,y_2)\in A(x),x\in I}|R_n(y_2|t_1,x)|=O_{\mathbb{P}}\left(\left(\frac{\log(n)}{nh_1h_2}\right)^{3/4}\right)+O_{\mathbb{P}}(h_1^2)+O_{\mathbb{P}}(h_2^2).$$

Since $nh_1^5\to 0$, $nh_1^3\to\infty$ and $nh_2^5\to c>0$, uniformly in $(y_1,y_2)\in A(x)$ and $x\in\{u:f_X(u)>0\}$, we obtain

$$\int_0^{y_1}R_n(y_2|t_1,x)F_1(dt_1|x)=o_{\mathbb{P}}((nh_1)^{-1/2}).$$

Hence, on the set $(y_1,y_2)\in A(x)$ and $x\in\{u: f_X(u)>0\}$, we have

$$A_n(y_1,y_2|x)=\sum_{i=1}^{n}\int_0^{y_1}\tilde{B}_{in}(t_1,x)\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)F_1(dt_1|x)+o_{\mathbb{P}}((nh_1)^{-1/2}), \quad (A.3)$$

where $\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)$ are i.i.d. and $E(\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)|\tilde{T}_{1i}=t_1,\delta_{1i}=1,X_i=x)=0$.

We now deal with the second term, $B_n(y_1,y_2|x)$, in the right hand side of (A.1). In order to deal with $B_n(y_1,y_2|x)$, we need to define

$$\tilde{H}(t_1|x)=\mathbb{P}(\tilde{T}_1\leq t_1,\delta_1=1|X=x)$$

and its estimator

$$\tilde{H}_n(t_1|x)=\sum_{i=1}^{n}B_{in}(x)1_{\{\tilde{T}_{1i}\leq t_1\}}\delta_{1i}.$$

Additionally, set

$$G_n(t|x)=1-\prod_{i=1}^{n}\left[1-\frac{B_{in}(x)1_{\{\tilde{T}_{1i}\leq t\}}(1-\delta_{1i})}{\sum_{j=1}^{n}1_{\{\tilde{T}_{1j}\geq\tilde{T}_{1i}\}}B_{jn}(x)}\right].$$

Then, if there are no ties and since $\sum_{i=1}^{n}B_{in}(x)=1$, we have that

$$\Delta F_{1n}(\tilde{T}_{1i}|x)=F_{1n}(\tilde{T}_{1i}|x)-F_{1n}(\tilde{T}_{1i}^{-}|x)=\frac{B_{in}(x)\delta_{1i}}{1-G_n(\tilde{T}_{1i}^{-}|x)}.$$

Moreover, for every function $\phi(t_1,x)$ and under A1, we have

$$\int\phi(t_1,x)F_1(dt_1|x)=\int\frac{\phi(t_1,x)}{1-G(t_1^{-}|x)}\tilde{H}(dt_1|x). \quad (A.4)$$

Finally,

$$B_n(y_1,y_2|x)=\int_0^{y_1}F_{21}(y_2|t_1,x)\frac{\tilde{H}_n(dt_1|x)}{1-G_n(t_1^{-}|x)}-\int_0^{y_1}F_{21}(y_2|t_1,x)\frac{\tilde{H}(dt_1|x)}{1-G(t_1^{-}|x)}.$$

Hence

$$B_n(y_1,y_2|x)=B_n^1(y_1,y_2|x)+B_n^2(y_1,y_2|x)+B_n^3(y_1,y_2|x)+B_n^4(y_1,y_2|x),$$

where

$$B_n^1(y_1,y_2|x)=\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{1-G(t_1^{-}|x)}(\tilde{H}_n(dt_1|x)-\tilde{H}(dt_1|x))$$

$$B_n^2(y_1,y_2|x)=\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^{-}|x))^2}[G_n(t_1^{-}|x)-G(t_1^{-}|x)]\tilde{H}(dt_1|x)$$

$$B_n^3(y_1,y_2|x)=\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^{-}|x))^2}[G_n(t_1^{-}|x)-G(t_1^{-}|x)](\tilde{H}_n(dt_1|x)-\tilde{H}(dt_1|x))$$

$$B_n^4(y_1,y_2|x)=\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)[G_n(t_1^{-}|x)-G(t_1^{-}|x)]^2}{(1-G(t_1^{-}|x))^2(1-G_n(t_1^{-}|x))}\tilde{H}_n(dt_1|x).$$

As to $B_n^1(y_1,y_2|x)$, taking into account that $\sum_{i=1}^n B_{in}(x)=1$, we can write it as follows

$$B_n^1(y_1,y_2|x)=\sum_{i=1}^n B_{in}(x)\left(\frac{F_{21}(y_2|\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}^-|x)}\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}-\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{1-G(t_1^-|x)}\tilde{H}(dt_1|x)\right).$$

Hence

$$B_n^1(y_1,y_2|x)=\sum_{i=1}^n B_{in}(x)\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x), \tag{A.5}$$

where

$$\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)=\frac{F_{21}(y_2|\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}^-|x)}\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}-\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{1-G(t_1^-|x)}\tilde{H}(dt_1|x)$$

are i.i.d. and $E(\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)|X_i=x)=0$.
Furthermore, we use again the results from González-Manteiga and Cadarso-Suárez (1994) for the Beran estimator $G_n(t|x)$. Consequently, for $t_1\le\tau_1(x)$, we obtain

$$G_n(t_1^-|x)-G(t_1^-|x)=\sum_{i=1}^n B_{in}(x)\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)+R_{1n}(t_1,x),$$

where

$$\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)=(1-G(t_1^-|x))\left[-\int_0^{\tilde{T}_{1i}\wedge t_1}\frac{d\tilde{\tilde{H}}(s|x)}{(1-H(s^-|x))^2}+\frac{1_{\{\tilde{T}_{1i}\le t_1,\delta_{1i}=0\}}}{1-H(\tilde{T}_{1i}^-|x)}\right],$$

$$\tilde{\tilde{H}}(s|x)=\mathbb{P}(\tilde{T}_1\le s,\delta_1=0|X=x)$$

and

$$\sup_{t_1\le\tau_1(x),x\in I}|R_{1n}(t_1,x)|=O_{\mathbb{P}}\left(\left(\frac{\log(n)}{nh_1}\right)^{3/4}\right)+O_{\mathbb{P}}(h_1^2)$$

Hence, on the set $(y_1,y_2)\in A(x)$ and for $x\in I$, we have

$$\begin{aligned}B_n^2(y_1,y_2|x)=&\sum_{i=1}^n\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^-|x))^2}B_{in}(x)\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)\tilde{H}(dt_1|x)\\&+O_{\mathbb{P}}\left(\left(\frac{\log(n)}{nh_1}\right)^{3/4}\right)+O_{\mathbb{P}}(h_1^2).\end{aligned}$$

Consequently, since $nh_1^5\to 0$ and $nh_1^3\to\infty$, we obtain

$$\begin{aligned}\sqrt{nh_1}B_n^2(y_1,y_2|x)=&\sqrt{nh_1}\sum_{i=1}^n\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^-|x))^2}B_{in}(x)\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)\tilde{H}(dt_1|x)\\&+o_{\mathbb{P}}(1),\end{aligned} \tag{A.6}$$

where $\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)$ are i.i.d. and $E(\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)|X_i=x)=0$.
Hence, it is easy to show that

$$\sqrt{nh_1}(B_n^3(y_1,y_2|x)+B_n^4(y_1,y_2|x))=o_{\mathbb{P}}(1). \tag{A.7}$$

Regarding the third term, $C_n(y_1,y_2|x)$, in the right hand side of (A.1), we have that

$$\sqrt{nh_1}C_n(y_1,y_2|x)=o_{\mathbb{P}}(1). \tag{A.8}$$

Finally, from (A.3)-(A.8), we obtain

$$\begin{aligned}\sqrt{nh_1}(\mathbf{F}_n(y_1,y_2|x)-\mathbf{F}(y_1,y_2|x)) &= \sqrt{nh_1}\sum_{i=1}^n\int_0^{y_1}\tilde{B}_{in}(x,t_1)\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)F_1(dt_1|x)\\ &+\sqrt{nh_1}\sum_{i=1}^n B_{in}(x)\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)\\ &+\sqrt{nh_1}\sum_{i=1}^n\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^-|x))^2}B_{in}(x)\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)\tilde{H}(dt_1|x)+o_{\mathbb{P}}(1).\end{aligned} \tag{A.9}$$

The Equation (A.9) is not yet the i.i.d. representation which we aim at. To deal with its first term, using (A.2), we obtain

$$\sqrt{nh_1}\sum_{i=1}^n\int_0^{y_1}\tilde{B}_{in}(x,t_1)\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)F_1(dt_1|x)=T_{1n}+T_{2n},$$

where

$$T_{1n}=\frac{1}{\sqrt{nh_1}h_2}\sum_{i=1}^n\int_0^{y_1}\frac{\delta_{1i}K\left(\frac{x-X_i}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1i}}{h_2}\right)}{\tilde{h}^1(t_1,x)}\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)F_1(dt_1|x)$$

and

$$\begin{aligned}T_{2n}=\sqrt{nh_1}\sum_{i=1}^n\int_0^{y_1}\frac{\tilde{B}_{in}(t_1,x)}{\tilde{h}^1(t_1,x)}&\left[\tilde{h}^1(t_1,x)-\frac{1}{nh_1h_2}\sum_{j=1}^n\delta_{1j}K\left(\frac{x-X_j}{h_1}\right)K\left(\frac{t_1-\tilde{T}_{1j}}{h_2}\right)\right]\\ &\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)F_1(dt_1|x).\end{aligned}$$

As to T_{2n}, recall that $E(\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)|\tilde{T}_{1i}=t_1,\delta_{1i}=1,X_i=x)=0$. Moreover, using (A.2) again, it is easy to show that $T_{2n}=o_{\mathbb{P}}(1)$. As to T_{1n}, by (A.4), change of variables and Taylor expansion, we obtain

$$T_{1n}=\frac{1}{\sqrt{nh_1}}\frac{1}{f_X(x)}\sum_{i=1}^n\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}K\left(\frac{x-X_i}{h_1}\right)\frac{\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}^-|x)}+O_{\mathbb{P}}(n^{1/2}h_1^{1/2}h_2^2).$$

The properties of the second and third term in (A.9) based on

$$B_{in}(x)=\frac{\frac{1}{nh_1}K\left(\frac{x-X_i}{h_1}\right)}{f_X(x)}+\frac{B_{in}(x)}{f_X(x)}\left[f_X(x)-\frac{1}{nh_1}\sum_{j=1}^n K\left(\frac{x-X_j}{h_1}\right)\right].$$

Finally,

$$\sqrt{nh_1}(\mathbf{F}_n(y_1,y_2|x)-\mathbf{F}(y_1,y_2|x))=\frac{1}{f_X(x)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi_{ni}(y_1,y_2,x)+o_{\mathbb{P}}(1), \quad \text{(A.10)}$$

where

$$\begin{aligned}\Phi_{ni}(y_1,y_2,x)&=\frac{1}{\sqrt{h_1}}\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}K\left(\frac{x-X_i}{h_1}\right)\frac{\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}-|x)}\\&+\frac{1}{\sqrt{h_1}}K\left(\frac{x-X_i}{h_1}\right)\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)\\&+\frac{1}{\sqrt{h_1}}\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^-|x))^2}K\left(\frac{x-X_i}{h_1}\right)\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)\tilde{H}(dt_1|x),\end{aligned}$$

$$\frac{\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,t_1,x)}{1-F_{21}(y_2|t_1,x)}=\left[-\int_0^{\tilde{T}_{2i}\wedge y_2}\frac{d\tilde{H}^*(s|t_1,x)}{(1-H^*(s^-|t_1,x))^2}+\frac{1_{\{\tilde{T}_{2i}\le y_2,\delta_{2i}=1\}}}{1-H^*(\tilde{T}_{2i}^-|t_1,x)}\right],$$

$$\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)=\frac{F_{21}(y_2|\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}^-|x)}\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}-\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{1-G(t_1^-|x)}\tilde{H}(dt_1|x)$$

and

$$\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)=(1-G(t_1^-|x))\left[-\int_0^{\tilde{T}_{1i}\wedge t_1}\frac{d\tilde{\tilde{H}}(s|x)}{(1-H(s^-|x))^2}+\frac{1_{\{\tilde{T}_{1i}\le t_1,\delta_{1i}=0\}}}{1-H(\tilde{T}_{1i}^-|x)}\right]$$

Since, for $i=1,..,n$, $\Phi_{ni}(y_1,y_2,x)$ are i.i.d. and $E\Phi_{ni}(y_1,y_2,x)=O(h_1^{5/2})$, the right hand side of (A.10) is a sum of i.i.d. random variables plus remainder of order $o_{\mathbb{P}}(1)$. Hence we obtain

$$\sqrt{nh_1}(\mathbf{F}_n(y_1,y_2|x)-\mathbf{F}(y_1,y_2|x))\to\mathcal{N}(0,\sigma_1^2(y_1,y_2|x)), \quad \text{(A.11)}$$

where

$$\sigma_1^2(y_1,y_2|x)=\frac{\int K^2(t)dt}{f_X(x)}\rho_1^2(y_1,y_2,x),$$

and

$$\begin{aligned}\rho_1^2(y_1,y_2,x)=Var\Bigg(&\delta_{1i}1_{\{\tilde{T}_{1i}\le y_1\}}\frac{\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,\tilde{T}_{1i},x)}{1-G(\tilde{T}_{1i}^-|x)}+\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)\\&+\int_0^{y_1}\frac{F_{21}(y_2|t_1,x)}{(1-G(t_1^-|x))^2}\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)\tilde{H}(dt_1|x)|X_i=x\Bigg) \quad \text{(A.12)}\end{aligned}$$

and the proof is completed.

Remark that, if there is no censoring, $C\equiv\infty$, it is easy to show that

$$\xi_1(\tilde{T}_{2i},\delta_{2i},y_2,\tilde{T}_{1i},x)=1_{\{T_{2i}\le y_2\}}-F_{21}(y_2|\tilde{T}_{1i},x),$$

$$\xi_2(\tilde{T}_{1i},\delta_{1i},y_1,y_2,x)=F_{21}(y_2|\tilde{T}_{1i},x)1_{\{T_{1i}\le y_1\}}-\mathbf{F}(y_1,y_2|x)$$

and

$$\xi_3(\tilde{T}_{1i},\delta_{1i},t_1,x)=0.$$

Hence $\rho_1^2(y_1,y_2,x)=\mathbf{F}(y_1,y_2|x)(1-\mathbf{F}(y_1,y_2|x))$ as desired.

SORT 48 (2) July-December 2024, 209-234 DOI: 10.57645/20.8080.02.19

Second-order Markov multistate models

Mireia Besalú[1] and Guadalupe Gómez Melis[2]

Abstract

Multistate models are well developed for continuous and discrete times under a first-order Markov assumption. Motivated by a cohort of COVID-19 patients, a multistate model was designed based on 14 transitions among 7 states of a patient. Since a preliminary analysis showed that the first-order Markov condition was not met for some transitions, we have developed a second-order Markov model where the future evolution not only depends on the state at the current time but also on the state at the preceding time. Under a discrete time analysis, assuming homogeneity and that past information is restricted to two consecutive times, we expanded the transition probability matrix and proposed an extension of the Chapman-Kolmogorov equations. We propose two estimators for the second-order transition probabilities and illustrate them within the cohort of COVID-19 patients.

MSC: *62M09, 62N02, 60J10.*

Keywords: *Multistate models, Non-Markov, COVID-19.*

1. Introduction

Multistate models (MSM) provide a very convenient methodology to describe the life history of an individual which at any time occupies one of a few possible states. In particular, they are appropriate to describe the clinical course of a disease and are routinely used in research to model the progression of patients among different states.

MSM theoretical justification is based on the theory of stochastic processes, that is, on sets of random variables representing the evolution of a process over time. The time can be chosen to be discrete or continuous; while discrete times assume a stepwise process where the fixed time between successive steps is not part of the model, continuous

[1] Departament Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain.

[2] Departament d'Estadística i Investigació operativa, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain.

Received: March 2023

Accepted: February 2024

time models allow changes of the states at any time. This class of models allows for an extremely flexible approach that can model almost any kind of longitudinal failure time data. This is particularly relevant for modeling different events, which have an event-related dependence, like the occurrence of a disease changing the risk of death (Hougaard, 1999).

The first-order Markov condition establishes that the future evolution of the stochastic process only depends on the current state and is frequently assumed in multistate models. However, this condition might often be not too realistic to describe clinical situations. To test it, Titman and Putter (2020) develop general log-rank tests that can be applied to general multistate models under right-censoring.

To circumvent the fact that the first-order Markov condition does not hold, the state space could have been extended with new states formed by two adjacent states of the original model. In this case, a first-order Markov condition would probably be satisfied, but the resulting model would be more complex and more difficult to interpret. In addition, the extended model would require more data to be estimated. See, for example, COVID model proposed by Mody et al. (2020), instead of one state of death, they have three depending on the history of the patient.

Another plausible approach to lessen the first-order Markov assumption is to consider a higher-order Markov process. A Markov process of order k is such that the dependence of the process on the whole history is only through the k states previously occupied. Although it is often observed that higher-order Markov processes can model the data better, models for Markov processes of higher order are scarcely used in practice because they depend on a very large number of parameters, leading to computational difficulties (Ching, Fung and Ng, 2003; Logan, 1981). Most instances of higher-order Markov models, which have been used so far, involved discrete time models (known as Markov chains). Tong (1975) defines a k order Markov chain $\{X_1, \cdots, X_n, \cdots\}$ as the one such that the conditional probabilities satisfy

$$P(X_n|X_{n-1}, X_{n-2}, \cdots) = P(X_n|X_{n-1}, X_{n-2}, \cdots, X_{n-k}) \tag{1}$$

for all n, where $k > 0$ is the smallest integer holding the above condition.

In this paper, we propose second-order Markov multistate models as a way of enriching the pathway information and still control the number of parameters while keeping the interpretability of the transition probabilities. Analysis using second-order Markov models are scarce. Among them, Shorrocks (1976) investigated the Markovian assumption in modelling income mobility and concluded that transition rates should depend on both current income and immediate past history, hence, a second-order Markov model was implemented. Shamshad et al. (2005) uses a second-order Markov model for synthetic generation of wind speed time series data.

Second-order Markov models assume that the progression of the individuals not only depends on the current state but also on the state at the preceding time. Second-order Markov multistate models are characterized by means of a $M \times M \times M$ tensor, where M is the number of states. In this work, we define an extended transition probability

matrix as M different matrices of order $M \times M$. To be able to compute n-step transition probabilities, we extend the first-order Chapman-Kolmogorov equations. We propose two different estimators for the second-order transition probabilities. We continue the paper with an illustration consisting of a cohort of more than 2000 COVID-19 patients from five hospitals in the Barcelona metropolitan area who were hospitalized during the first wave of the coronavirus pandemic (March-April 2020). For this data we have built a multistate model based on 14 possible transitions among the seven states where a patient can be in after his/her admission. We have adopted a second-order Markov based on plausible medical interpretations and in view of our available data.

We estimate the second-order transition probabilities and based on those we compute, among others, the transition probability from one state to another, after a given number of hospitalized days, and differentiating between patients that arrive to the hospital with severe pneumonia from those who arrive to the hospital with mild pneumonia. The paper ends with a discussion on shortcomings while setting the path for future research.

2. Characterization of first-order Markov multistate processes

A multistate process is a continuous (or discrete)-time stochastic process $X = \{X_t,\ t \geq 0\}$ taking values in a discrete state space $\mathcal{S} = \{1, \cdots, M\}$. We denote by $\mathcal{F}_t := \sigma\{X_s : s \leq t\}$ a σ-algebra consisting on the observation of the process over the interval $[0,t]$ and we refer to it as a filtration. We can think of a filtration as the history of the process up to time t containing the information on the previous occupied states up until time t.

The law of a multistate process is defined by its finite-dimensional distribution and is fully characterized through either one of the following three transition functions: transition probabilities, transition intensities or cumulative transition intensities. The *transition probability* between states h and j for times s and t, $s < t$ is defined by:

$$P_{hj}(s,t;\mathcal{F}_{s-}) = P(X_t = j \mid X_s = h; \mathcal{F}_{s-}) \qquad \text{for} \qquad h, j \in \mathcal{S} = \{1, \cdots, M\}$$

and denotes the probability of the process being at state j at time t knowing that it has been at state h at time s as well as knowing all the previous trajectory before s. The *transition intensity* between states h and j, for time t is defined by:

$$\alpha_{hj}(t; \mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P_{hj}(t, t+\Delta t; \mathcal{F}_{t-}) \qquad \text{for} \qquad h, j \in \mathcal{S} = \{1, \cdots, M\}$$

and denotes the instantaneous probability to change from state h to state j at time t. The *cumulative (integrated) transition intensity* between states h and j at time t is defined by:

$$A_{hj}(t; \mathcal{F}_{t-}) = \int_0^t \alpha_{hj}(u; \mathcal{F}_{u-}) du \qquad \text{for} \qquad h, j \in \mathcal{S} = \{1, \cdots, M\}$$

Transition probabilities, transition intensities and cumulative transition intensities are summarized by means of $M \times M$ matrices. In particular, for every trajectory collected

in $\mathcal{F}_{t-}$ and for every s,t such that $s<t$, we denote by $\mathbf{P}$ the transition probability matrix

$$\mathbf{P}(s,t;\mathcal{F}_{s-})=\{P_{hj}(s,t;\mathcal{F}_{s-});h,j\in\mathcal{S}=\{1,\cdots,M\}\}.$$

2.1. First-order Markov and homogeneity assumptions

It is clear that some restrictions have to be made in order to estimate the transition probabilities $P_{hj}(s,t;\mathcal{F}_{s-})$ for every pair of states h and j, for every pair of times s and t and for all the possible trajectories before s. The Markov and the homogeneity assumptions are key to make inferences feasible.

Definition 2.1. *A multistate process satisfies the first-order Markov assumption if for all $h,j\in\mathcal{S}=\{1,\cdots,M\}$ and s,t such that $s<t$*

$$P_{hj}(s,t;\mathcal{F}_{s-})=P(X_t=j\mid X_s=h;\mathcal{F}_{s-})=P(X_t=j\mid X_s=h)=P_{hj}(s,t). \qquad (2)$$

That is, under the first-order Markov assumption, different trajectories before s will not change the transition probabilities. Under the first-order Markov assumption, an $M\times M$ matrix, $P(s,t)$, is needed for every (s,t)

$$\mathbf{P}(s,t)=\{P_{hj}(s,t);\ h,j\in\mathcal{S}=\{1,\cdots,M\}\}$$

Definition 2.2. *A first-order Markov multistate process is said to be homogeneous if the transition probability between any states at given times t,s $(s<t)$ depends only on the difference between these two times $(t-s)$, that is,*

$$P_{hj}(s,t)=P_{hj}(0,t-s)=P_{hj}(t-s).$$

In this case only a $M\times M$ matrix $\mathbf{P}(t)$ for every time t is needed.

2.2. Markov Test

We should validate the (first-order) Markov condition if we want to proceed analysing the data under this assumption. One choice would be to include the time of entry into each state as a covariate within a Cox model and test its significance through a likelihood ratio test (Kay, 1986). A second possibility would be to use the stratified version of the Commenges-Andersen's test to detect a shared frailty. Other authors (Rodríguez-Girondo and de Uña Álvarez, 2012) have developed local and global tests for the Markov conditions based upon the observed Kendall's τ for the progressive three-state illness-death model.

In this paper, and in the subsequent COVID-19 analysis, we will validate the Markov assumption for each transition by means of Titman and Putter (2020)'s test that we briefly describe. The main idea of this test is that under the first-order Markov assumption, the

rate of transitions at time $t > s$ will not be affected by the state occupied at time s. If we want to check the Markov assumption for the transition between the states l and m ($l, m \in \mathcal{S}$), we divide the subjects into two different groups: the ones that at time s are in a fixed state $j \in \mathcal{S}$ and the ones who are not there. Then, for each transition ($l \to m$) the null hypothesis for a fixed state j and fixed time s ($s \in [t_0, t_{\max}] \subset [0, \tau]$, τ total follow-up) is stated as:

$$\mathrm{H}_{0s}^{(j)}(l,m) : \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \text{for any} \quad t \in [s, \tau]$$

and can be tested with the log-rank statistic

$$U_s^{(j)}(l,m) = \sum_{i=1}^{n} \int_s^{\tau} \left\{ \delta_i^{(j)}(s) - \frac{\sum_{k=1}^{n} \delta_k^{(j)}(s) Y_{kl}(t)}{\sum_{k=1}^{n} Y_{kl}(t)} \right\} dN_i^{(lm)}(t),$$

where $\delta_i^{(j)}(s) = \mathbb{1}\{X_i(s) = j\}$ denotes whether individual i has been in state j for time s, $Y_i(t)$ is the at risk indicator for the process $X_i(t)$, $Y_{il}(t) = \mathbb{1}\{X_i(t^-) = l\} Y_i(t)$ is the at risk indicator of transition $l \to m$ for subject i and $N_i^{(lm)}(t)$ is the counting process reporting the number of times of the transition $l \to m$ up to time t.

The standardized statistics

$$\overline{U}_s^{(j)}(l,m) = \frac{U_s^{(j)}(l,m)}{\sqrt{\widehat{Var}(U_s^{(j)})(l,m)}}.$$

can be compared to a $N(0,1)$. Moreover, $\left\{ \overline{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}] \right\}$ converges to a zero mean Gaussian process with a covariance function that can be consistently estimated.

Given the null hypothesis for a fixed state j

$$\mathrm{H}_0^{(j)}(l,m) : \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \forall s \in [t_0, t_{\max}] \subset [0, \tau] \text{ and } t \in [s, \tau],$$

a global test statistic can be defined based on summary statistics of $\{\overline{U}_s^{(j)}(l,m), s \in [t_0, t_{\max}]\}$ such as $\int_{t_0}^{t_{\max}} \left| \overline{U}_s^{(j)}(l,m) \right| ds$, $\sup_{s \in [t_0, t_{\max}]} |\overline{U}_s^{(j)}(l,m)|$ or $\int_{t_0}^{t_{\max}} w(s) |\overline{U}_s^{(j)}(l,m)| ds$ for some weight function $w(s)$.

Finally, an overall test statistic for the null hypothesis for any possible j and for all $s \in [t_0, t_{\max}] \subset [0, \tau]$ and $t \in [s, \tau]$

$$\mathrm{H}_0(l,m) : \alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \forall j, \forall s \in [t_0, t_{\max}] \subset [0, \tau] \text{ and } \forall t \in [s, \tau]$$

can be defined from the global test statistics, for instance as the mean, the maximum or weighted mean of them. These tests are implemented in `R` with the function `MarkovTest` of the package `mstate` of de Wreede, Fiocco and Putter (2011). Details of how are implemented are postponed to the illustration in Section 6.3.

3. Characterization of second-order Markov transition probabilities

To address the limitations of the first-order Markov assumption, we introduce a second-order Markov assumption, which acknowledges that the future evolution of the stochastic process depends not only on its current state but also on the state it occupied in the preceding time. We begin by defining the second-order Markov transition probabilities and describing how they can be summarized into a set of as many matrices as states.

Definition 3.1. *For times (s,t,u), $s<t<u$ and states h,j,k, the probability $P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j; \mathcal{F}_{s-})$ satisfies a second-order Markov assumption if and only if*

$$\begin{aligned} P_{hjk}(s,t,u;\mathcal{F}_{s-}) &= P(X_u = k \mid X_s = h, X_t = j | \mathcal{F}_{s-}) \\ &= P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u). \end{aligned}$$

Under the second-order Markov assumption the transition probabilities are summarized, for every three times (s,t,u), $s<t<u$, by an $M \times M \times M$ tensor $\mathbf{P}(s,t,u)$

$$\mathbf{P}(s,t,u) = \{P_{hjk}(s,t,u); h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

In order to have a more manageable mathematical object we denote, for each state $h \in \mathcal{S}$, a matrix of dimension M, $\mathbf{P}_{(\mathbf{h})}(s,t,u)$ as follows:

$$\mathbf{P}_{(\mathbf{h})}(s,t,u) = \left(P_{hjk}(s,t,u)\right)_{j,k\in\mathcal{S}},$$

hence, the tensor $\mathbf{P}(s,t,u)$ of transition probabilities can be equivalently represented as M matrices of order M for each $s<t<u$.

Remark 3.2. *The matrices $\mathbf{P}_{(\mathbf{h})}(s,t,u)$ are not always stochastic matrices because*

$$\sum_{k\in\mathcal{S}} P_{hjk}(s,t,u) = \begin{cases} 0 & \text{if } \forall j,k, \text{ the transitions } h \to j \text{ or } j \to k \text{ are not possible} \\ 1 & \text{otherwise.} \end{cases}$$

For example, if h is an absorbent state all the matrices will be 0 except for the element $P_{hhh} = 1$.

Definition 3.3. *A second-order Markov multistate process is said to be homogeneous if the transition probability between any three states at given times (s,t,u), $s<t<u$, depends only on the differences $t-s$ and $u-t$ between the two consecutive times that is,*

$$P_{hjk}(s,t,u) = P_{hjk}(t-s,u-t)$$

In this case only a $M \times M \times M$ tensor $\mathbf{P}(s,t)$ for every pair of times (s,t) $(s<t)$ is needed

$$\mathbf{P}(s,t) = \{P_{hjk}(s,t); h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

Using the previous notation, denote as $\mathbf{P}_{(\mathbf{h})}(\mathbf{s},\mathbf{t})$ the matrix of dimension M for each state $h \in \mathcal{S}$ and for every pair of times (s,t) $(s<t)$, that is,

$$\mathbf{P}_{(\mathbf{h})}(s,t) = \left(P_{hjk}(s,t)\right)_{j,k\in\mathcal{S}}.$$

Note that the tensor $\mathbf{P}(s,t)$ of transition probabilities under homogeneity can be equivalently represented as M matrices of dimension M for each two times $(s,t); s<t$ where s stands for the time from h to j and t stands for the time from j to k.

4. Computation of transition probabilities

Given that clinical outcomes are often collected in days and aiming to compute the probability of being in a given state after a certain number of days, we consider in this section a discrete-time multistate process instead of a continuous-time stochastic process defined for $t \in [0,T]$. Other instances of discrete-time multistate process have been used to model COVID-19 disease progression and clinical outcomes (Chakladar et al., 2022).

In this section we provide the expressions to compute general probabilities, such as

$$P_{hjl}(s,s+n,s+n+m) = P(X_{s+n+m} = l \mid X_{s+n} = h, X_s = j). \tag{3}$$

for any three states j,h,l at any three times $s,s+n,s+n+m$. In order to get there we start extending the Chapman-Kolmogorov equations from first to second-order Markov chains based on 2-step second-order transition probabilities, that is, on $P(X_{s+2} = l \mid X_{s-1} = h, X_s = j)$. In Subsection 4.3, probabilities (3) are written as a function of the 1-step first-order transition probabilities, $P(X_{s+1} = h | X_s = j)$, and 1-step second-order transition probabilities for consecutive times, $P(X_{s+n+1} = l \mid X_{s+n} = h, X_{s+n-1} = j)$.

Remark 4.1. *We use the term n-step **second-order** transition probabilities to refer to n-step transition probabilities conditioned to 2 consecutive times, that is, $P(X_{s+n} = l \mid X_{s-1} = h, X_s = j)$, for $n \geq 1$. We also define the n-step **first-order** transition probabilities as follows $P(X_{s+n} = l \mid X_s = h)$ for $n \geq 1$.*

4.1. Chapman-Kolmogorov equations for first-order Markov chains

A first-order discrete-time multistate models, known as Markov chain, taking values in a discrete state space $\mathcal{S} = \{1,\cdots,M\}$ is the discrete version of a first-order continuous-time Markov process. Hence, a Markov chain is a stochastic model describing a sequence of possible events happening on discrete times in which the probability of each event depends only on the state attained in the previous event. The Chapman-Kolmogorov

relation is an important result in the theory of (discrete) Markov chains as it provides a method for calculating the n-step transition probabilities.

The Chapman-Kolmogorov equations, for any $s,t,u \in \mathbb{N}$ $(s < u < t)$ and any two states h, j are given by

$$P_{hj}(s,t) = \sum_{l=1}^{m} P_{hl}(s,u)\, P_{lj}(u,t), \tag{4}$$

where $P_{hj}(s,t) = P(X_t = j \mid X_s = h)$ is the transition probability defined in (2). Chapman-Kolmogorov equations follow as a consequence of the Markov condition. Chapman-Kolmogorov equations allow to reduce the general computation of $P_{hj}(s,t)$, for any $s < t$, $(s,t \in \mathbb{N})$ to the computation of 1-step first-order transition probabilities, $P_{hj}(s,s+1)$, that is,

$$P_{hj}(s,s+n) = \sum_{l=1}^{M} P_{hl}(s,s+1)\, P_{lj}(s+1,s+n), \tag{5}$$

Denote by $\mathbf{P}(s)$ the one-time step transition probability matrix under the Markov assumption, that is,

$$\mathbf{P}(s) = \{P_{hj}(s);\, h, j \in \mathcal{S} = \{1, \cdots, M\}\}$$

where $P_{hj}(s)$ stands for $P_{hj}(s,s+1)$. The collection of matrices $\mathbf{P}(s)$ is reduced to the transition probability matrix $\mathbf{P}$ given by

$$\mathbf{P} = \{P_{hj} = P_{hj}(1);\, h, j \in \mathcal{S} = \{1, \cdots, M\}\}$$

under the homogeneity assumption (see Definition 2.2). Hence, to study the evolution of the process for more than one time step, and thanks to the Chapman-Kolmogorov equations (4), it is only necessary to calculate the one-time initial transition probabilities. In the next section we develop an extension of this result for second-order Markov chains.

4.2. Chapman-Kolmogorov equations for second-order Markov chains

Under second-order Markov and homogeneity assumptions the transition probability matrices defined in Section 3 satisfy, for times $s < t < u$ and states h, j, k:

$$P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u) = P_{hjk}(t-s, u-t)$$

In particular, for any $s \in \mathbb{N}$, $s > 1$ and consecutive times $s, s+1, s+2$, computation of the probabilities $P_{hjk}(s,s+1,s+2;\mathcal{F}_{s-})$ is reduced to the computation of 1-step second-order transition probabilities, that is,

$$\begin{aligned} P_{hjk}(s,s+1,s+2;\mathcal{F}_{s-}) &= P(X_{s+2} = k \mid X_s = h, X_{s+1} = j) \\ &= P_{hjk}(s,s+1,s+2) = P_{hjk}(1,1). \end{aligned}$$

Next theorem presents the equations to compute n-step second-order transition probabilities such as

$$P(X_{s+n+1} = l | X_{s+1} = j, X_s = h) = P_{hjl}(1,n), \tag{6}$$

for $s, n \in \mathbb{N}, n > 1$ and $l, j, k \in \mathcal{S}$ using only the initial transition probabilities $P(X_3 = l|X_2 = j, X_1 = h) = P_{hjl}(1,1) = P_{hjl}$.

Since our desired probabilities only depend on n, without lost of generality we can assume $s = 1$ and the probabilities at (6) can be equivalently written as

$$P(X_{n+2} = l|X_2 = j, X_1 = h), \quad \text{for} \quad n \in \mathbb{N},\ l, j, k \in \mathcal{S}.$$

Recall that these transition probabilities can be summarized into M matrices of dimension M for each $n \in \mathbb{N}$. So for each state $h \in \mathcal{S}$ and $n > 1$

$$\mathbf{P}_{(\mathbf{h})}(1,2,n+2) = \left(P_{hjl}(1,n)\right)_{j,k\in\mathcal{S}} = \left(P_{hjl}(n)\right)_{j,l\in\mathcal{S}},$$

we will omit n of the previous notation when $n = 1$.

Notation 4.2. *Previous to the main result, we will present the matrix notation used in order to simplify the reading.*

- *Row j of matrix h:* $P_{j\cdot(h)} = (P_{hjk})_{k\in\mathcal{S}}$.
- *Column k of matrix h:* $P_{\cdot k(h)} = (P_{hjk})_{j\in\mathcal{S}}$.
- *We will use the $*$ symbol to denote that the elements of the row multiply each row of the matrix. For example $P_{hj\cdot} * P_{(h)}$ means that element P_{hjk} multiplies all the elements of the row k of $P_{(h)}$.*
- $\mathbf{P}^{(l)}$ *is the matrix composed with the l column of each of the P_h matrices $h \in \mathcal{S}$.*
- $\mathrm{Tr}()$ *will denote the trace of a matrix.*

Theorem 4.3. *Assume $(X_n)_{n\in\mathbb{N}}$ is an homogeneous second-order Markov chain. For any states $h, j, l \in \mathcal{S}$ and the notation defined in Notation 4.2, we have (where* Tr *denotes trace)*

$$P(X_4 = l \mid X_2 = j, X_1 = h) = \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \tag{7}$$

$$P(X_5 = l \mid X_2 = j, X_1 = h) = \mathrm{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)}\right) \tag{8}$$

$$P(X_6 = l \mid X_2 = j, X_1 = h) = \sum_{k_1=1}^{M} P_{hjk_1} \cdot \mathrm{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{9}$$

$$P(X_7 = l \mid X_2 = j, X_1 = h) = \sum_{k_2=1}^{M}\sum_{k_1=1}^{M} P_{hjk_2}P_{jk_2k_1} \cdot \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{10}$$

General case $n \geq 7$

$$P(X_{n+1} = l \mid X_2 = j, X_1 = h) = \sum_{k_{n-4}=1}^{M} \cdots \sum_{k_2=1}^{M}\sum_{k_1=1}^{M} P_{hjk_{n-4}}P_{jk_{n-4}k_{n-3}} \ldots P_{k_3k_2k_1} \times \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right)$$

Proof. The proof will be divided into three steps.

Step 1. We are proving the first case (7).

Using the total probabilities Theorem and the second-order Markov property,

$$\begin{aligned}
P(X_4 = l \mid X_2 = j, X_1 = h) &= \sum_{k=1}^{M} P(X_4 = l, X_3 = k \mid X_2 = j, X_1 = h) \\
&= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j, X_1 = h) \\
&= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j) \\
&= \sum_{k=1}^{M} P_{hjk} \cdot P_{jkl} \qquad (11)
\end{aligned}$$

We recall that $P(X_4 = l \mid X_3 = k, X_2 = j) = P_{jkl}(1,1) = P_{jkl}$ since for all the assumptions only the difference between times determine the transition probabilities. From here we can write it in matricial form as in (7).

Step 2. Now we focus in the second case (8).

Using the same argument of the previous case and also the result obtained (11)

$$\begin{aligned}
P(X_5 = l \mid X_2 = j, X_1 = h) &= \sum_{k_1=1}^{M} P(X_5 = l, X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j, X_1 = i) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} \left(\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2} \right) \cdot P_{h j k_1}
\end{aligned}$$

Now, if we want to write it in a matricial way we can observe that $\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2}$ corresponds to the product of the l column of each matrix by the matrix $\mathbf{P}_{(j)}$ and then each row of this matrix product is multiplied by the probabilities $P_{h j k_1}$ which are the elements of row j of matrix $\mathbf{P}_{(h)}$. From here we obtain the formula in (8).

Step 3. We follow by proving (9).

We repeat here the arguments in the previous steps and also we apply the previous results.

$$
\begin{aligned}
P(X_6 = l \mid X_2 = j, X_1 = h) &= \sum_{k_3=1}^{M} P(X_6 = l, X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j, X_1 = h) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P_{hjk_3} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right) \\
&= P_{hj1} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right) + \ldots + P_{hjM} \left(\sum_{k_1=1}^{M} P_{jk_3k_1} \sum_{k_2=1}^{M} P_{k_1k_2l} \cdot P_{k_3k_1k_2} \right)
\end{aligned}
$$

We observe that we obtain a similar expression as the previous step but multiplied by the probabilities $P_{hj\cdot}$ that correspond to the row j of matrix $\mathbf{P}_{(h)}$.Thus, the matricial expression for this case is immediate.

From here, in order to prove Equation (10) and the general case we can just repeat the same arguments as in this last case to easily obtain the general formula by induction. In these two cases the principal modifications of the matricial form will focus in adding one sum for each step. ■

Corollary 4.3.1. *Assume $(X_n)_{n\in\mathbb{N}}$ is an homogeneous second-order Markov chain. For time $s > 0$, any states $h, j, l \in \mathcal{S}$ and the notation defined in Notation 4.2, we have*

$$
\begin{aligned}
P(X_{s+3} = l \mid X_{s+1} = j, X_s = h) &= \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \\
P(X_{s+4} = l \mid X_{s+1} = j, X_s = h) &= \mathrm{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)} \right) \\
P(X_{s+5} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_1=1}^{M} P_{hjk_1} \cdot \mathrm{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right) \\
P(X_{s+6} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_2} P_{jk_2k_1} \cdot \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right)
\end{aligned}
$$

General case $n \geq 7$

$$
\begin{aligned}
P(X_{s+n} = l \mid X_{s+1} = j, X_s = h) = \sum_{k_{n-4}=1}^{M} \ldots \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_{n-4}} P_{jk_{n-4}k_{n-3}} \ldots P_{k_3k_2k_1} \\
\times \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)} \right) \quad (12)
\end{aligned}
$$

Remark 4.4. *We observe that the extended Chapman-Kolmogorov equations only consider the case where the two past times are consecutive.*

4.3. Computation of arbitrary transition probabilities

So far, the extended Chapman-Kolmogorov equations have only considered those cases where the two past times were consecutive. For some models and specific cases with non return states it is possible to compute these probabilities.

In this section we prove that for any 3 times $s, s+n, s+n+m$ ($s>0$, $n,m>1$) transition probabilities defined as follows:

$$P_{hjl}(s,s+n,s+n+m) = P(X_{s+n+m}=l \mid X_{s+n}=h, X_s=j)$$

when the two past times are not consecutive, can be written as a function of the 1-step first-order transition probabilities and the 1-step second-order transition probabilities for consecutive times. The tools presented in the previous subsections will be now crucial to obtain the desired probability.

Theorem 4.5. *Assume $(X_n)_{n\in\mathbb{N}}$ is an homogeneous second-order Markov chain. For any states $h,j,l \in \mathcal{S}$ and time $s>0$ and $n,m>1$ we have*

$$P(X_{s+n+m}=l \mid X_{s+n}=h, X_s=j) = \frac{\sum_{e_1=1}^{M}\cdots\sum_{e_{n-1}=1}^{M} P(X_{s+n+m}=l \mid X_{s+n}=h, X_{s+n-1}=e_{n-1}) P_{e_{n-1}e_{n-2}h}(s+n)\ldots P_{je_1e_2}(s+2) P(X_{s+1}=e_1 \mid X_s=j)}{P(X_{s+n}=h \mid X_s=j)}$$

where $P(X_{s+n+m}=l \mid X_{s+n}=h, X_{s+n-1}=e_{n-1})$ can be written in terms of 1-step second-order transition probabilities for consecutive times as it is shown in (12) *and $P(X_{s+n}=h \mid X_s=j)$ can be written in terms of 1-step first-order transition probabilities as it is shown in* (5).

Proof. Indeed,

$$\begin{aligned} P_{hjl}(s,s+n,s+n+m) &= P(X_{s+n+m}=l \mid X_{s+n}=h, X_s=j) \\ &= \frac{P(X_{s+n+m}=l, X_{s+n}=h, X_s=j)}{P(X_{s+n}=h, X_s=j)} \end{aligned}$$

We can write the numerator in the following way

$$\begin{aligned} &P(X_{s+n+m}=l, X_{s+n}=h, X_s=j) \\ &\quad = \sum_{e_1=1}^{M}\cdots\sum_{e_{n-1}=1}^{M} P(X_{s+n+m}=l, X_{s+n}=h, X_{s+n-1}=e_{n-1},\ldots,X_{s+1}=e_1, X_s=j) \\ &\quad = \sum_{e_1=1}^{M}\cdots\sum_{e_{n-1}=1}^{M} P(X_{s+n+m}=l \mid X_{s+n}=h, X_{s+n-1}=e_{n-1},\ldots,X_{s+1}=e_1, X_s=j) \\ &\qquad\qquad \times P(X_{s+n}=h, X_{s+n-1}=e_{n-1},\ldots,X_{s+1}=e_1, X_s=j) \end{aligned}$$

We can iterate the process and apply the second-order Markov hypothesis and finally obtain

$$\begin{aligned}
&P(X_{s+n+m}=l,X_{s+n}=h,X_s=j)\\
&\quad=\sum_{e_1=1}^{M}\cdots\sum_{e_{n-1}=1}^{M}P(X_{s+n+m}=l\mid X_{s+n}=h,X_{s+n-1}=e_{n-1})\\
&\qquad\times P(X_{s+n}=h\mid X_{s+n-1}=e_{n-1},X_{s+n-2}=e_{n-2})\times\ldots\\
&\qquad\times P(X_{s+2}=e_2\mid X_{s+1}=e_1,X_s=j)\times P(X_{s+1}=e_1|X_s=j)P(X_s=j)
\end{aligned}$$

For the denominator we just recall the conditional probability definition

$$P(X_{s+n}=h,X_s=j)=P(X_{s+n}=h|X_s=j)P(X_s=j).$$

Now we have already prove the result since $P(X_s=j)$ is in both numerator and denominator and we can simplify it. ■

Remark 4.6. *With the result of this Theorem, the Chapman-Kolmogorov equations for first and second-order and the one-step transition probabilities also for first and second-order we can now compute any transition probability for any triplet of times.*

5. Estimation and inference under second-order Markov assumption

Given three different states h, j, $l\in\mathcal{S}$ such that h, j are not absorbent, the purpose of this Section is to estimate the r-step transition probabilities $P(X_{s+r}=l\mid X_{s-1}=j,X_{s-2}=h)$ for any $s,r\in\mathbb{N}$, $s,r>1$. Under the homogeneity assumption we have that $P(X_s=l\mid X_{s-1}=j,X_{s-2}=h)=P_{hjl}(1,1)=P_{hjl}$ and, following Corollary 4.3.1, in order to estimate $P(X_{s+r}=l\mid X_{s-1}=j,X_{s-2}=h)$ is enough to estimate the initial transition probabilities $P_{hjl}(1,1)=P_{hjl}$.

We assume that individuals are followed until a maximum of T units of time (days as in the illustration). Let $\{X_s^i,\ s=0,1,\cdots,T\}$ denote the non-reversible multistate process for subject $i=1,\ldots,n$, where $X_s^i\in\mathcal{S}$. For $i=1,\ldots,n$, $s=2,\cdots,T$ and h, j, $l\in\mathcal{S}$ we define the counting processes

$$N_{hjl}^i(s)=\mathbb{1}\{X_{s-2}^i=h,X_{s-1}^i=j,X_s^i=l\}$$

counting 1 if subject i has transit from state h to state j and to state l at times $s-2$, $s-1$, s, respectively; and 0 otherwise. The total number of individuals who have followed the path $h\to j\to l$ at times $s-2$, $s-1$, s is given by the sum $\widetilde{N}_{hjl}(s)=\sum_{i=1}^n N_{hjl}^i(s)$. $\widetilde{N}_{hjl}(s)$ is a binomial random variable with parameters $(n,\pi_{hjl}(s))$ where the probability

$\pi_{hjl}(s)$ corresponds to $\pi_{hjl}(s) = P(X_{s-2} = h, X_{s-1} = j, X_s = l)$. We also define the at-risk process of subject i corresponding to states h and j at times $s = 2, \cdots, T$,

$$Y_{hj}^i(s-1) = \mathbb{1}\{X_{s-2}^i = h, X_{s-1}^i = j\}$$

counting 1 if subject i was at risk of moving to adjacent states to j or stay at j given that he/she was in states h and j at times $s-2$ and $s-1$, respectively. The total number of individuals at risk at time s is given by $\widetilde{Y}_{hj}(s-1) = \sum_{i=1}^n Y_{hj}^i(s-1)$ and corresponds to a binomial random variable with parameters $(n, \pi_{hj}(s-1))$ where $\pi_{hj}(s-1) = P(X_{s-2} = h, X_{s-1} = j)$.

Regarding the estimation of the transition probability $P_{hjl}(s) = P(X_s = l \mid X_{s-1} = j, X_{s-2} = h)$ for a given $s \geq 2$, we will proceed in two different ways. The first one takes advantage of the ratio of the two probabilities $P_{hjl}(s) = \pi_{hjl}(s)/\pi_{hj}(s-1)$ while the second one exploits directly the estimation of the conditional probability $P_{hjl}(s)$.

From a practical point of view we will have to guarantee that the number of individuals at risk $\widetilde{Y}_{hj}(s-1)$ is large enough for a meaningful estimation of P_{hjl}.

5.1. Estimation of P_{hjl} via the Bernoulli probabilities π_{hjl} and π_{hj}

Given that for all $s = 2, \cdots, T$

$$P_{hjl} = P_{hjl}(s) = P(X_s = l \mid X_{s-1} = j, X_{s-2} = h) = \frac{P(X_{s-2} = h, X_{s-1} = j, X_s = l)}{P(X_{s-2} = h, X_{s-1} = j)}$$
$$= \frac{\pi_{hjl}(s)}{\pi_{hj}(s-1)}$$

a natural estimator for P_{hjl} can be built estimating separately both numerator and denominator by $\sum_{s=2}^T \widetilde{N}_{hjl}(s)/n$ and $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)/n$, respectively. Observe that $\sum_{s=2}^T \widetilde{N}_{hjl}(s)$ corresponds to the total number of individuals that have followed the path $h \to j \to l$ at any three times $(s-2, s-1, s)$ and $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)$ is the total number of individuals that have followed the path $h \to j$ consecutively at any two times $(s-2, s-1)$.

Definition 5.1. *For given states (h, j, l), the statistic*

$$\widetilde{P}_{hjl} = \frac{\sum_{s=2}^T \widetilde{N}_{hjl}(s)}{\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)} \tag{13}$$

estimates P_{hjl}. Whenever the denominator $\sum_{s=2}^T \widetilde{Y}_{hj}(s-1)$ is equal to 0, meaning that no individuals have contributed to the path $h \to j$, we will take $\widetilde{P}_{hjl} = 0$.

Theorem 5.2. *For given states (h, j, l), the statistic $\widetilde{P}_{hjl}$ defined in* (13) *is a consistent estimator of P_{hjl}.*

Proof. Note that by the Law of Large Numbers we have convergence in probability of the following two estimators:

$$\frac{1}{n}\sum_{s=2}^{T}\widetilde{N}_{hjl}(s)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=2}^{T}N_{hjl}^{i}(s)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=2}^{T}N_{hjl}^{i}(s)\right)=\sum_{s=2}^{T}E\left(N_{jhl}^{i}(s)\right)$$
$$=\sum_{s=2}^{T}\pi_{hjl}(s)$$
$$\frac{1}{n}\sum_{s=2}^{T}\widetilde{Y}_{hj}(s-1)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=2}^{T}Y_{hj}^{i}(s-1)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=2}^{T}Y_{hj}^{i}(s-1)\right)$$
$$=\sum_{s=2}^{T}E\left(Y_{jh}^{i}(s-1)\right)=\sum_{s=2}^{T}\pi_{hj}(s-1)$$

Second-order homogeneity implies that

$$\begin{aligned}\pi_{hjl}(s)&=P\{X_{s-2}=h,X_{s-1}=j,X_s=l\}\\&=P\{X_s=l\mid X_{s-2}=h,X_{s-1}=j\}P\{X_{s-1}=j,X_{s-2}=h\}\\&=P_{hjl}(1,1)P\{X_{s-1}=j,X_{s-2}=h\}=P_{hjl}(1,1)\pi_{hj}(s-1),\end{aligned}$$

and we conclude that $\widetilde{P_{jhl}}$ converges in probability to P_{hjl}:

$$\widetilde{P}_{jhl}=\frac{\sum_{s=2}^{T}\tilde{N}_{hjl}(s)}{\sum_{s=2}^{T}\tilde{Y}_{hj}(s-1)}=\frac{\sum_{s=2}^{T}\tilde{N}_{hjl}(s)/n}{\sum_{s=2}^{T}\tilde{Y}_{hj}(s-1)/n}\xrightarrow[n\to\infty]{\mathrm{P}}\frac{\sum_{s=2}^{T}\pi_{hjl}(s)}{\sum_{s=2}^{T}\pi_{hj}(s-1)}$$
$$=\frac{\sum_{s=2}^{T}P_{hjl}(1,1)\pi_{hj}(s-1)}{\sum_{s=2}^{T}\pi_{hj}(s-1)}=P_{hjl}$$

■

5.2. Estimation of P_{hjl} via the conditional probability

For every $s\geq 2$, the relative frequency given by the ratio $\widetilde{N}_{hjl}(s)/\widetilde{Y}_{hj}(s-1)$ is, whenever $\widetilde{Y}_{hj}(s-1)>0$, an straightforward estimator of $P_{hjl}(s)$.

Because the homogeneity assumption, we have for all $s=2,\cdots,T$, $P_{hjl}(s)=P_{hjl}$ and hence, an estimator for P_{hjl} can be obtained as the average of $\widehat{P}_{hjl}(s)$.

Definition 5.3. *For given states (h,j,l) and for every $s\geq 2$, define the statistic*

$$\widehat{P}_{hjl}(s)=J_{hj}(s-1)\frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}$$

where $J_{hj}(s-1) = \mathbb{1}\{\widetilde{Y}_{hj}(s-1) > 0\}$. *To estimate* P_{hjl} *we define an alternative estimator as follows:*

$$\widehat{P}_{hjl} = \frac{1}{t_{hj}} \sum_{s=2}^{T} \widehat{P}_{hjl}(s). \tag{14}$$

where $t_{hj} = \sum_{s=2}^{T} J_{hj}(s-1)$ *counts the number of times where at least there is an individual experiencing the path* $j \to h$.

Theorem 5.4. *For given states* (h, j, l) *the statistic* $\widehat{P}_{hjl}$ *given in* (14) *is an unbiased estimator of* P_{hjl}.

Proof. We assume that t_{hj}, the number of times where at least there is an individual at risk, is fixed. Then,

$$\begin{aligned}
\mathrm{E}[\widehat{P}_{hjl}] &= \mathrm{E}\left[\frac{1}{t_{hj}} \sum_{s=2}^{T} J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}\right] = \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[\mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)} \,\middle|\, \tilde{Y}_{hj}(s-1)\right]\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\mathrm{E}\left[\widetilde{N}_{hjl}(s) \,\middle|\, \tilde{Y}_{hj}(s-1)\right]}{\widetilde{Y}_{hj}(s-1)}\right] \\
&= \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) \frac{\widetilde{Y}_{hj}(s-1) p_{hjl}}{\tilde{Y}_{hj}(s-1)}\right] = \frac{1}{t_{hj}} \sum_{s=2}^{T} \mathrm{E}\left[J_{hj}(s-1) p_{hjl}\right] \\
&= p_{hjl} \mathrm{E}\left[\frac{1}{t_{hj}} \sum_{s=2}^{T} J_{hj}(s-1)\right] = p_{hjl}
\end{aligned}$$

where we have used that $\tilde{N}_{hjl}(s) | \tilde{Y}_{hj}(s-1) \sim \mathrm{Bin}(\tilde{Y}_{hj}(s-1), p_{hjl})$ ■

6. DIVINE model

6.1. Description

The dataset we use as illustration corresponds to a cohort of 2076 COVID-19 hospitalised patients (during the first wave of the pandemic, March-April 2020) in five hospitals located in the southern Barcelona metropolitan area (Spain). Since all the patients were monitored until discharge from hospital or death, the transition times (in days) are known exactly for all subjects and there are not incomplete data due to lost to follow-up.

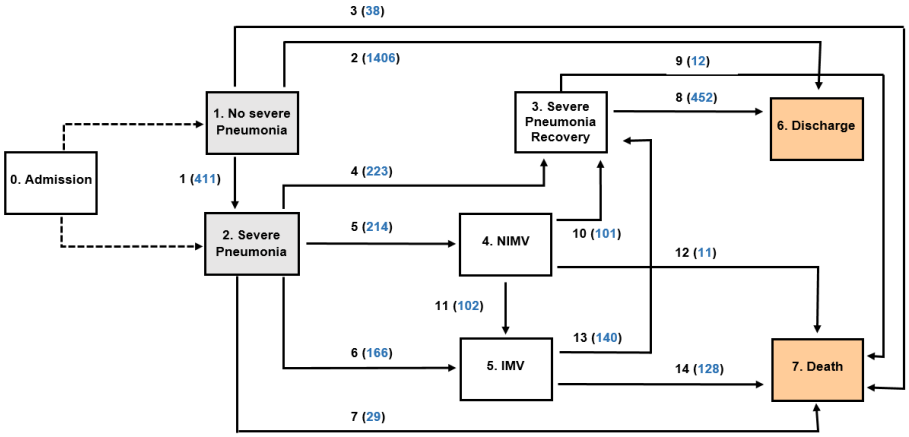

Figure 1. *Graphical representation of the multistate model for modelling the trajectory of hospitalized COVID-19 patients. Seven states are considered and 14 possible transitions (in parentheses the sample size of each transition). NIMV: Non-Invasive mechanical ventilation, IMV: Invasive mechanical ventilation.*

This data is part of the DIVINE project (https://grbio.upc.edu/en/research/highlighted-projects) for which a multidisciplinary research team integrated by researchers from the GRBIO (UPC-UB), Bellvitge University Hospital, and Bellvitge Biomedical Research Institute has collaborated to define a statistical framework with a clear clinician focus on achieving deeper understanding of the severe form of the disease caused by the SARS-CoV-2 virus. Based on the team cooperative knowledge a multistate model with seven states and 14 transitions has been built (see Figure 1 where the numbers in parentheses denote the patients doing that transition). As seen in Figure 1, 7 states are considered: (1) No Severe Pneumonia (NSP), (2) Severe Pneumonia (SP), (3) Severe Pneumonia Recovery (Recov), (4) Non invasive mechanical ventilation (NIMV), (5) Invasive mechanical ventilation (IMV), (6) Discharge (Disch) and (7) Death (Death).

The following considerations are in place:

1. Once a patient has been admitted (state 0), he/she is immediately assigned to one of the two initial states: No Severe Pneumonia and Severe Pneumonia. It is assumed that the process starts at time $t = 0$ in one of these two states.

2. Discharge and Death are absorbing states implying that once a patient has been discharged or has died he/she cannot re-enter to be hospitalised again.

3. The time scale used in this model is days since the hospital admission. For all the transitions, the transition times (in days) are exactly known.

4. Patients can only jump to a neighboring state in a single day.

For more details on the data and the clinical patient characteristics see Pallarès et al. (2023); Garmendia, Cortés and Gómez Melis (2023); Piulachs et al. (2023).

The main goal with this illustration is to study the evolution of the patients without the restriction of a first-order Markov assumption. To do so we start validating for which transitions of the previous multistate model the Markov assumption holds. Next, we will estimate the transition probabilities between two states taking into account that they might depend as well on the immediate previous state. Finally, we will compare the evolution of those patients admitted with No Severe Pneumonia versus those admitted with Severe Pneumonia.

6.2. Description of direct and two-step transitions

Table 1 summarises the number of patients for each direct transition and the number of patients for the corresponding related 2-step transitions (consecutive states but not necessarily consecutive times). For instance, individuals doing the direct transition Recov $\rightarrow$ Disch might arrive from SP : SP $\rightarrow$ Recov $\rightarrow$ Disch from NIMV : NIMV $\rightarrow$ Recov $\rightarrow$ Disch or from IMV : IMV $\rightarrow$ Recov $\rightarrow$ Disch. Note that we are only considering those direct transitions $j \rightarrow l$ for which there exists, at least, a state k adjacent to j ($k \rightarrow j \rightarrow l$).

Table 1. *Aggregation of the 2-step paths for each direct transition taking into account the previous immediate state.*

Direct transition	Sample size	2-step transition	Sample size	Percent.
SP $\rightarrow$ Recov	223	NSP $\rightarrow$ SP $\rightarrow$ Recov	171	76.68
		SP $\rightarrow$ Recov	52	23.32
SP $\rightarrow$ NIMV	214	NSP $\rightarrow$ SP $\rightarrow$ NIMV	134	62.62
		SP $\rightarrow$ NIMV	80	37.38
SP $\rightarrow$ IMV	166	NSP $\rightarrow$ SP $\rightarrow$ IMV	92	55.42
		SP $\rightarrow$ IMV	74	44.58
SP $\rightarrow$ Death	29	NSP $\rightarrow$ SP $\rightarrow$ Death	14	48.26
		SP $\rightarrow$ Death	15	51.72
Recov $\rightarrow$ Disch	452	SP $\rightarrow$ Recov $\rightarrow$ Disch	223	49.34
		NIMV $\rightarrow$ Recov $\rightarrow$ Disch	96	21.24
		IMV $\rightarrow$ Recov $\rightarrow$ Disch	133	29.42
Recov $\rightarrow$ Death	12	SP $\rightarrow$ Recov $\rightarrow$ Death	0	0
		NIMV $\rightarrow$ Recov $\rightarrow$ Death	5	41.67
		IMV $\rightarrow$ Recov $\rightarrow$ Death	7	58.33
IMV $\rightarrow$ Death	128	SP $\rightarrow$ IMV $\rightarrow$ Death	71	57.26
		NIMV $\rightarrow$ IMV $\rightarrow$ Death	57	45.97
IMV $\rightarrow$ Recov	140	SP $\rightarrow$ IMV $\rightarrow$ Recov	95	67.86
		NIMV $\rightarrow$ IMV $\rightarrow$ Recov	45	32.14

States: NSP: No Severe Pneunomia, SP: Severe pneumonia, Recov: Severe Pneumonia recovery, NIMV: Non-Invasive Mechanical Ventilation, IMV: Invasive Mechanical Ventilation, Disch: Discharge, Death: Death

Relating the sample size of the direct transitions appearing in Table 1 with the notation introduced in Section 5, we see that the sample size of the direct transition $h \to j$ corresponds to the number of patients at-risk: $\sum_{s=2}^{T} \tilde{Y}_{hj}$, while the sample size of the 2-step transition $h \to j \to l$ does not coincide with $\sum_{s=2}^{T} \tilde{N}_{hjl}$ since we have considered all patients doing this path either in consecutive times or not.

Table 1 reveals that the proportion of patients for a given transition (e.g, IMV$\to$ Recov) drastically differs whether the patients were before in SP (68%) or in NIMV (32%). We can also examine the transition SP$\to$Recov, if we separate the patients between those with NSP in the admission (76.68%) and those with SP in the admission (23.32%) we also observe important differences. Similar interpretation is in place with transition SP$\to$NIMV. This suggests that the model may not fulfill the Markov assumption and that it may be important to take the two previous states into account when calculating the transition probabilities.

6.3. Testing the Markov assumption

In order to check which 2-step transitions are not first-order Markovian, we use the Markov test described in Section 2.2 and follow Titman and Putter (2020) guidelines with respect to the time intervals $[t_0, t_{\max}]$ where the test can be conducted. Basically, the comparison is restricted to windows of time with enough individuals and to direct transitions that have an immediate previous state.

To evaluate the logrank test we compute the statistics for an equally 0.5-day spaced grid in the interval $[1, 11]$ for all the transitions except for transitions 7 (SP $\to$ Death) and 12 (NIMV $\to$ Death) in which the interval is $[1, 7]$ and transition 14 (IMV $\to$ Death) with the interval $[1, 16]$.

Table 2 summarizes the p-values of the log-rank tests obtained from 5000 wild bootstrap resamples (Lin, Wei and Ying, 1993) and considering the three possible summary statistics: weighted mean, mean, and supremum described in Section 2.2. For each transition (rows), we have carried out the test for all the possible previous states as well as the overall chi-squared test. The partial p-values are the ones corresponding to the global test. For each transition, the transition intensity compares the subjects who were previously at fixed state j (in columns) versus the ones who were not there.

Considering the overall p-values, transitions 4 (SP $\to$ Recov), 5 (SP $\to$ NIMV), 6 (SP $\to$ IMV) and 8 (Recov $\to$ Disch) show clear departures from the Markov assumption, while transition 13 (IMV $\to$ Recov) is marginally significant. Furthermore, any one of the summary tests rejects the Markovianity in transitions 4, 5 and 6. The supremum statistic would not reject Markovianity from states NSP and Recov in transition 8. Finally, the global p-value of 0.059 in transition 13 is mainly due to the non Markovianity coming from states SP and IMV.

These findings suggest that once a patient is critically ill, for instance, in states NIMV and IMV, the future clinical evolution is independent of whether he/she was diagnosed with NSP or SP when hospitalized. However, the clinical evolution to NIMV or IMV will be different for those patients initially diagnosed with NSP versus those diagnosed with

Table 2. *p-values obtained from the computation of the Markov test for each transition and each previous state. Three different summary statistics have been computed: unweighted mean (UM), weighted mean (WM) and supremum (S). In bold the transitions that are statistically significant at 0.05.*

Transitions		NSP	SP	Recov	NIMV	IMV	overall
4 (SP→ Recov)	UM	0.005	0.005				**0.0042**
	WM	0.006	0.006				
	S	0.029	0.029				
5 (SP→ NIMV)	UM	$< 10^{-16}$	$< 10^{-16}$				**0.0018**
	WM	$< 10^{-16}$	$< 10^{-16}$				
	S	0.026	0.026				
6 (SP→ IMV)	UM	0.009	0.009				**0.016**
	WM	0.002	0.002				
	S	0.042	0.042				
7 (SP→ Death)	UM	0.106	0.106				0.196
	WM	0.120	0.120				
	S	0.340	0.340				
8 (Recov→ Disch)	UM	0.007	$< 10^{-5}$	0.165	$< 10^{-5}$	$< 10^{-5}$	$< \mathbf{10^{-16}}$
	WM	0.010	$< 10^{-5}$	0.186	$< 10^{-5}$	$< 10^{-5}$	
	S	0.104	$< 10^{-5}$	0.388	$< 10^{-5}$	$< 10^{-5}$	
9 (Recov→ Death)	UM	0.652	0.298	0.145	0.495	0.143	0.357
	WM	0.644	0.273	0.144	0.464	0.151	
	S	0.656	0.313	0.353	0.639	0.309	
10 (NIMV→ Recov)	UM	0.594	0.190		0.694		0.609
	WM	0.588	0.183		0.717		
	S	0.831	0.432		0.892		
11 (NIMV→IMV)	UM	0.514	0.819		0.728		0.807
	WM	0.501	0.858		0.765		
	S	0.432	0.311		0.304		
12 (NIMV→ Death)	UM	0.348	0.218		0.649		0.437
	WM	0.378	0.253		0.619		
	S	0.338	0.342		0.719		
13 (IMV→ Recov)	UM	0.564	$< 10^{-3}$		0.514	0.005	0.059
	WM	0.531	$< 10^{-3}$		0.456	0.005	
	S	0.780	0.034		0.376	0.037	
14 (IMV→ Death)	UM	0.296	0.663		0.318	0.099	0.305
	WM	0.296	0.674		0.269	0.100	
	S	0.471	0.369		0.264	0.205	

SP when hospitalized. These results lead us to consider second-order Markov multistate models in order to study the evolution of the hospitalized COVID-19 patients during the fist wave of the pandemia.

6.4. Estimation of the transition probability matrices

We will now estimate the seven possible transition probability matrices using the estimators presented in Section 5. As we have mentioned before, we know the exact transition times, so we can easily estimate the transition probability by taking into account the number of patients who are at risk of the transition and the patients who finally have done the transition. For each row of each matrix the number of patients at risk will be different.

We present here the estimation of the matrices $\mathbf{P}_{(\mathbf{1})}$ and $\mathbf{P}_{(\mathbf{2})}$. The estimation of the rest of the matrices is similar, except for matrices $\mathbf{P}_{(\mathbf{6})}$ and $\mathbf{P}_{(\mathbf{7})}$ (6 and 7 are absorbent states) which are null matrices except for the elements $(6,6)$ and $(7,7)$ which are equal to 1.

In order to estimate the matrix $\mathbf{P}_{(\mathbf{1})}$ we start from all patients who have been hospitalized with entry in state NSP. The day after a patient has been hospitalized he/she can still be at NSP or can move to SP, Discharge or Death. So rows 1, 2, 6 and 7 are the only ones with probabilities different from 0. In order to estimate the probability cells in row 1 of $\mathbf{P}_{(\mathbf{1})}$ we consider, for each time s, all the patients who have been at least two consecutive days in NSP, that is, $\sum_{s=2}^{43} Y_{11}(s-1) = 10577$ (note here that 43 is the maximum number of days a patient has been two consecutive days in NSP). From those 10577 patients at risk, the number of patients who have stayed in NSP the next day is $\sum_{s=2}^{43} N_{111}(s) = 8919$, while $\sum_{s=2}^{43} N_{112}(s) = 257$ have transited to SP, $\sum_{s=2}^{43} N_{116}(s) = 1369$ have been discharged and, finally, $\sum_{s=2}^{43} N_{117}(s) = 32$ have died. We proceed analogously for the estimation of the probability cells in row 2 of $\mathbf{P}_{(\mathbf{1})}$ starting with those $\sum_{s=2}^{36} Y_{12}(s-1) = 411$ patients who have moved to SP from NSP the next day.

$$\widehat{\mathbf{P}}_{(\mathbf{1})} = \begin{pmatrix} \frac{8919}{10577} & \frac{257}{10577} & 0 & 0 & 0 & \frac{1369}{10577} & \frac{32}{10577} \\ 0 & \frac{253}{411} & \frac{3}{411} & \frac{92}{411} & \frac{62}{411} & 0 & \frac{1}{411} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In a similar way we estimate $\mathbf{P}_{(\mathbf{2})}$. In this case we start with all patients who have been at state SP at any time. The next day these patients can still be at SP or can move to Recovery, NIMV, IMV or Death. So rows 1 and 6 are 0 because there is no direct transition from SP to NSP nor to Discharge. For row 2, the number of patients at risk, that is, the number of patients spending two consecutive times in state SP is $\sum_{s=2}^{47} Y_{22}(s-1) = 2668$. Row 3 starts with those patients who have moved from SP to Recovery,

a total of $\sum_{s=2}^{47} Y_{23}(s-1) = 223$. Analogously, for rows 4 and 5, $\sum_{s=2}^{14} Y_{24}(s-1) = 214$ patients have transited immediately from SP to NIMV while $\sum_{s=2}^{23} Y_{25}(s-1) = 166$ patients moved from SP to IMV. Probability matrices $\mathbf{P}_{(3)}$, $\mathbf{P}_{(4)}$ and $\mathbf{P}_{(5)}$ are estimated proceeding in an analogous manner, each one starting from patients in states Recov, NIMV and IMV, respectively.

$$\widehat{\mathbf{P}}_{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2307}{2668} & \frac{220}{2668} & \frac{68}{2668} & \frac{49}{2668} & 0 & \frac{24}{2668} \\ 0 & 0 & \frac{207}{223} & 0 & 0 & \frac{16}{223} & 0 \\ 0 & 0 & \frac{6}{214} & \frac{159}{214} & \frac{45}{214} & 0 & \frac{4}{214} \\ 0 & 0 & \frac{3}{166} & 0 & \frac{160}{166} & 0 & \frac{3}{166} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

6.5. Prediction via Chapman-Kolmogorov equations

The Markov test computed in Section 2.2 rejects the first-order Markov assumption for three of the four transitions from Severe Pneumonia (SP): to Non Invasive Mechanical Ventilation (NIMV), Invasive Mechanical Ventilation (IMV) and Recovery (Recov), indicating that whether or not the patient was diagnosed with Non Severe Pneumonia (NSP) marks a difference in his/her prognosis. A second-order model allows the prediction of the time to future events as a function of the diagnostic when they were hospitalized. Chapman-Kolomogorov extension in Theorem 4 is the key to the corresponding probabilities.

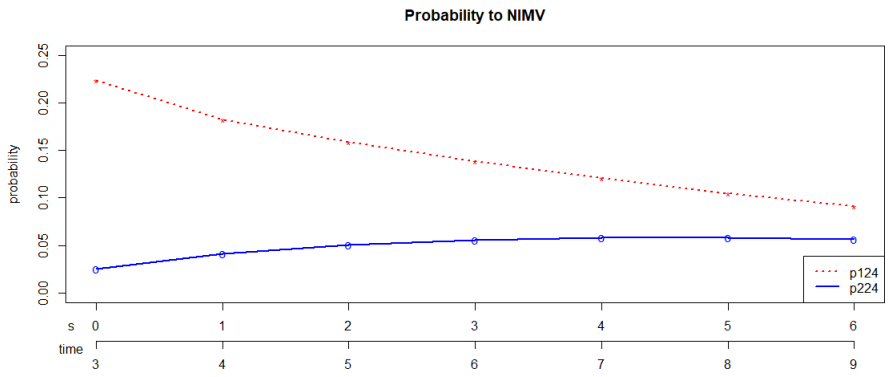

Figure 2. *Probabilities from SP to NIMV for patients who had been two consecutive days in SP (line) compared with patients who have been one day in NSP and one day in SP (dots).* $p_{224} = P(X_{3+s} = 4|X_2 = 2, X_1 = 2)$ *vs* $p_{124} = P(X_{3+s} = 4|X_2 = 2, X_1 = 1)$

For the transition from SP $\rightarrow$ NIMV, we will compute for $s \in \{0, \ldots, 6\}$ the probabilities $p_{224} = P(X_{3+s} = 4|X_2 = 2, X_1 = 2)$, that is, the probability to NIMV for patients who

arrive at the hospital with a SP diagnosis and they still were in SP the second day. And also $p_{124} = P(X_{3+s} = 4|X_2 = 2, X_1 = 1)$ the probability to NIMV for patients with NSP at admission who had moved to SP the second day. We have plotted these probabilities in Figure 2, where we can we see how important is the initial state for the initial times. The probability of moving to NIMV of patients initially diagnosed with SP ($X_1 = 2, X_2 = 2$) is much smaller than the probability of moving to NIMV of patients initially diagnosed with NSP ($X_1 = 1, X_2 = 2$). These two probabilities close the gap as days go by.

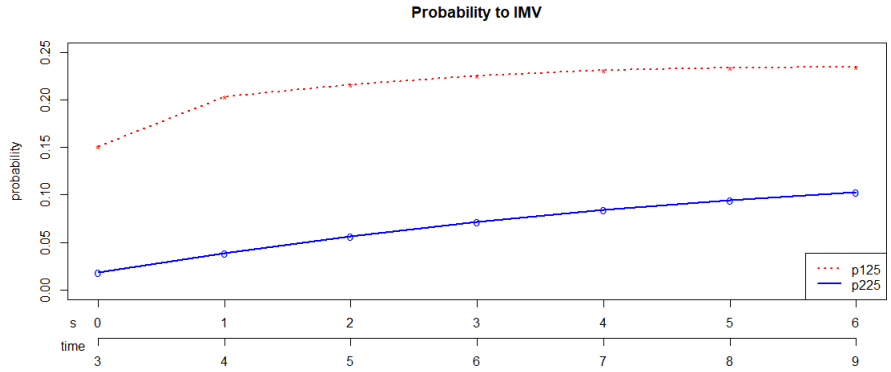

Figure 3. *Probabilities from SP to IMV for patients who had been two consecutive days in SP (line) compared with patients who have been one day in NSP and one day in SP (dots).* $p_{225} = P(X_{3+s} = 5|X_2 = 2, X_1 = 2)$ *vs* $p_{125} = P(X_{3+s} = 5|X_2 = 2, X_1 = 1)$.

The same type of plot is depicted in Figure 3 to study the transition SP $\rightarrow$ IMV. In this case the patients are also splitted based on their initial state: NSP or SP. As in Figure 2, the probability of moving to IMV of patients initially diagnosed with SP ($X_1 = 2, X_2 = 2$) is much smaller than the probability of moving to IMV of patients initially diagnosed with NSP ($X_1 = 1, X_2 = 2$). However, both probabilities increase over the time and their difference is kept along the next days. This reveals the different prognosis for needing respiratory mechanical ventilation (IMV) among those patients initially diagnosed with NSP versus being diagnosed with SP.

7. Discussion

In this paper we have introduced a second-order Markov multistate model and we have developed an extension of the Chapman-Kolmogorov equations to compute r-step transition probabilities. We have used the DIVINE COVID-19 data to estimate the transition probabilities and to predict probabilities to NIMV and IMV in terms of the states where a patient was during the first 2 days of his/her hospitalization.

As we briefly mention in the introduction, a second-order Markov model could had been transformed into a first-order Markov model by redefining the state space. This would be possible creating extra states formed by direct 1-step transitions. For instance,

in the DIVINE data case instead of one unique Death state we could have defined 3 new states formed by those patients arriving to Death from NIMV, IMV or Recovery. The advantage of these new states is clear because we would be able to apply all the knowledge on first-order Markov models. However, the number of states and transitions of the new model will increase substantially and the interpretation will become cumbersome. Furthermore, since the number of parameters to estimate will increase and, so does the needed sample sample size to estimate all of them, the second-order Markov approach is preferable.

It should be mentioned that Chapman-Kolmogorov extension is based on a discretization of the time scale and is only computed conditionally to two-consecutive times. But, we prove that one-step transition probabilities and one-step second-order transition probabilities together with the extended Chapman-Kolmogorov equations are enough to compute the transition probabilities if the previous two times are not consecutives.

In this paper we sketch two different ways to estimate the transition probabilities. The first one using the Bernoulli probabilities, which is the one used to compute the transition probabilities in the COVID illustration example and the second using conditional probabilities. Since the data from the DIVINE project was collected one year after the end of the first wave, we have complete registries and, for this reason, we have so far only developed both methods for complete (uncensored) data. Nevertheless, it is indeed relevant to extend these estimators to account for right-censored data. The second method of estimation presented in Subsection 5.2 gives a clue of how we could proceed to account for right-censored data. This estimator, an average of the ratios, for each time, of those subjects doing an specific transition among the number of subjects at risk, has an analogy to the Nelson-Aalen estimator for the cumulative hazard function. For a thorough statistical analysis, the derivation of the variance of these estimators as well as of their asymptotic distribution is needed. Furthermore, estimators for the state occupation probabilities and for the transition intensities for complete and right-censored data are as well a topic of interest. All these ideas remain open for our future research.

Acknowledgements

This research has been funded by the Ministerio de Ciencia e Innovación (Spain) [PID 2019-104830RB-I00/ DOI(AEI): 10.13039/501100011033] and by Generalitat de Catalunya through the projects 2020PANDE00148 and 01421 SGR-Cat 2021. We are indebted to our colleagues in the DIVINE group for their clever contributions and dedicated time.

References

Chakladar, S., R. Liao, W. Landau, M. Gamalo, and Y. Wang (2022). Discrete Time Multistate Model With Regime Switching for Modeling COVID-19 Disease Progression and Clinical Outcomes. *Statistics in Biopharmaceutical Research*, 14, 52–66.

Ching, W. K., Fung, E. S. and Ng, M. K. (2003). A higher-order Markov model for the Newsboy's problem. *Journal of the Operational Research Society*, 54, 291–298.

de Wreede, L., Fiocco, M. and Putter, H. (2011). `mstate`: An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38, 1–30.

Garmendia, L., Cortés, J. and Gómez Melis, G. (2023). MSMpred: Interactive modelling and prediction of individual evolution via multistate models. *BMC Medical research methodology*, 23.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Anal.*, 5, 239–264.

Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics*, 42, 855–865.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–572.

Logan, J. A. (1981). A structural model of the higher-order Markov process incorporating reversion effects. *The Journal of Mathematical Sociology*, 8, 75–89.

Mody, A., Lyons, P. G., Vazquez Guillamet, C., Michelson, A., Yu, S., Namwase, A. S., Sinha, P., Powderly, W. G., Woeltje, K. and Geng, E. H. (2020). The Clinical Course of Coronavirus Disease 2019 in a US Hospital System: A Multistate Analysis. *American Journal of Epidemiology*, 190, 539–552.

Pallarès, N., Tebé, C., Abelenda-Alonso, G., Rombauts, A., Oriol, I., Simonetti, A. F., Rodríguez-Molinero, A., Izquierdo, E., Díaz-Brito, V., Molist, G., Gómez Melis, G., Carratalà, J., Videla, S. and study groups, M. (2023). Characteristics and Outcomes by Ceiling of Care of Subjects Hospitalized with COVID-19 During Four Waves of the Pandemic in a Metropolitan Area: A Multicenter Cohort Study. *Infectious diseases and therapy*, 12, 273–289.

Piulachs, X., Langhor, K., Besalú, M., Pallarès, N., Carratalà, J., Tebé, C. and Gómez Melis, G. (2023). Semi-Markov multistate approaches for multicohort event history data. *(submitted)*.

Rodríguez-Girondo, M. and de Uña Álvarez, J. (2012). A nonparametric test for Markovianity in the illness-death model. *Statistics in Medicine*, 31, 4416–4427.

Shamshad, A., Bawadi, M., Wan Hussin, W., Majid, T. and Sanusi, S. (2005). First and second order markov chain models for synthetic generation of wind speed time series. *Energy*, 30, 693–708.

Shorrocks, A. F. (1976). Income Mobility and the Markov Assumption. *The Economic Journal*, 86, 566–578.

Titman, A. C. and Putter, H. (2020). General tests of the Markov property in multi-state models. *Biostatistics*, 23, 380–396.

Tong, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12, 488–497.

SORT 48 (2) July-December 2024, 235-258 DOI: 10.57645/20.8080.02.20

Conditional likelihood based inference on single-index models for motor insurance claim severity

Catalina Bolancé[1], Ricardo Cao[2] and Montserrat Guillen[1]

Abstract

Prediction of a traffic accident cost is one of the major problems in motor insurance. To identify the factors that influence costs is one of the main challenges of actuarial modelling. Telematics data about individual driving patterns could help calculating the expected claim severity in motor insurance. We propose using single-index models to assess the marginal effects of covariates on the claim severity conditional distribution. Thus, drivers with a claim cost distribution that has a long tail can be identified. These are risky drivers, who should pay a higher insurance premium and for whom preventative actions can be designed. A new kernel approach to estimate the covariance matrix of coefficients' estimator is outlined. Its statistical properties are described and an application to an innovative data set containing information on driving styles is presented. The method provides good results when the response variable is skewed.

MSC: *62G05, 62P20, 91G70.*

Keywords: *covariance matrix of estimator, kernel estimator, marginal effects, telematics covariates, right-skewed cost variable.*

1. Introduction

We analyse costs of claims in a motor insurance data set. Because higher costs occur much less frequently than lower costs of claims, the dependent variable here is right-skewed. Specifically, we are interested in modelling the distribution of costs of claims conditional on the values of covariates that reflect driving habits. We focus on the whole conditional distribution rather than on the conditional expectation to measure the influence of covariates on different quantiles, specifically on the costly claims, i.e., the right

[1] Department of Econometrics, RISKcenter-IREA, Universitat de Barcelona (UB).

[2] Research Group MODES, Department of Mathematics, CITIC, Universidade da Coruña and ITMATI.

Received: April 2023

Accepted: January 2024

tail of the severity distribution. This problem could be addressed by quantile regression, for fixed quantile levels, but this could potentially lead to contradictory results for close quantiles. Modelling the cost of claims conditional on covariate information has remained a bottleneck for insurance companies, as a result of which average costs are used in practice worldwide. We address this problem also considering data on driving patterns and driving conditions, a type of information that is available through sensor data regularly collected by insurtech firms. Some new motor insurance rate making schemes are based on near-miss telematics information which measures the propensity of risky events that do not always lead to an accident (see Guillen et al., 2019, 2020 and Guillen, Nielsen and Pérez-Marín, 2021). Risk scores such as the ones obtained with index-models can be combined with the evaluation of near-miss information to improve the performance of predictive modelling in motor insurance pricing.

Single-index regression models are semiparametric methods for generalising linear regression. They specify the dependence between a random variable Y (here the cost of a traffic accident, or claim severity) and a d-dimensional vector X as follows (see Härdle et al., 1993):

$$Y = m\left(\theta^\top X\right) + \varepsilon, \tag{1}$$

where θ is a vector of unknown parameters, m is an unknown smooth function, and ε is a random variable with zero-mean conditional on X.

Traditional approaches for estimating the linear predictor coefficients θ and the function m are based on the conditional expectation rather than on the whole conditional distribution and, as a consequence, they are vulnerable to the presence of extremes, heavy tails or strong asymmetry, as in many applications. Our contribution is to extend the maximum likelihood estimation of (1) and, in so doing, to open the door to single-index conditional distribution modelling which has enormous potential for a range of applications.

In order to estimate the vector θ, Härdle, Hall and Ichimura (1993) proposed the direct minimisation of the residual sum of squares, so their estimator is

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} \left[Y_i - \hat{m}_i\left(\theta^\top X_i\right)\right]^2,$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid observations of the covariates and the dependent variable and $\hat{m}_i$ indicates the leave-one-out kernel estimator of m. Alternatively, Hristache, Juditsky and Spokoiny (2001) analysed the average derivative estimator of the vector of parameters in the index model, introduced by Stoker (1986) and as subsequently employed by Powell, Stock and Stoker (1989). Hristache et al. (2001) presented the method for estimating the vector of coefficients, θ, by minimising an $M-$function, with a score function ψ, that again compares Y_i with a nonparametric estimator $\hat{m}(\cdot)$, i.e., $\arg\min_\theta \sum_{i=1}^{n} \psi\left[Y_i, \hat{m}\left(\theta^\top X_i\right)\right]$. All these methods ignore the shape of the conditional distribution because they are based on fitting the conditional expectation.

Delecroix, Härdle and Hristache (2003) investigated the pseudo-maximum likelihood estimation of θ in (1). They proposed starting from a preliminary $\sqrt{n}$-consistent

estimator and, subsequently, correcting it with the gradient and the Hessian of the log-likelihood function. They showed that the corrected estimator is efficient. Previously, Klein and Spady (1993) had analysed the maximum likelihood estimation of θ but only for a binary response dependent variable. In the context of survival data with censored observations, Strzalkowska-Kominiak and Cao (2013) investigated maximum likelihood alternatives based on the kernel estimation of the conditional distribution and showed that previous methods for censored data could be improved.

Nonparametric regression is more general than the single-index model specified in (1). Indeed, it emanates from a more general specification $Y = m(X) + \varepsilon$, where the aim is to estimate the regression curve $m(x) = E(Y|X = x)$; Härdle (1990). However, in practice, nonparametric regression presents two considerable challenges. First, estimation becomes increasingly difficult as the number of covariates rises (the curse of dimensionality). The second challenge is that any interpretation of the effects of the explanatory variables cannot be carried out directly and it is necessary to plot the different relations to explore these effects. Another alternative to the single-index model is the generalised additive model (see Hastie and Tibshirani, 1990); however, it faces the same challenges as those described for nonparametric regression.

Here, a new maximum likelihood estimator of θ in (1) is proposed, inspired by the work of Strzalkowska-Kominiak and Cao (2013) with right-censored data. As these authors proposed we use two different smoothing parameters: one associated with the distribution of Y and the other one associated with the distribution of the index $\theta^\top X$. The new theoretical results that we present in Section 2 for uncensored data do not follow directly as a particular case of Strzalkowska-Kominiak and Cao (2013), since some assumptions of the censored data case can be relaxed or dropped. In this paper, we deduce the covariance matrix that can be easily estimated using a kernel estimator. We evaluate the inference power of the statistical test for the covariate effects deduced from our maximum likelihood estimator. Details on the method, some results of the simulation study and proofs are available in the Supplementary Material.

We show the superiority of our estimator, in particular, when there are extreme values, like in our application where we observe only a few severe accidents. Additionally, we show that the results of the estimated index model are easily interpretable from different points of view, for example, for the prediction of conditional mean, quantiles and marginal effects.

We analyse a data set obtained from a specific portfolio from an insurance company in Spain. The portfolio is made up of a small group of policyholders under 35 years of age, who have underwritten a new insurance contract that requires a telematics device to be installed in their vehicle. The data set contains information on mean yearly claim cost per policy and on telematic and non-telematic characteristics. Our aim is to find the influence of telematic information on pricing compared to a traditional approach with only classical non-telematic variables. The data set is available at http://www.ub.edu/rfa/R/SORT-BCG/. We observe how the mean yearly claim cost per policy does not change with a linear index; however, the shape of the distribution depends on a linear index, something that could be considered when calculating the premium.

In a simulation study presented in Section 3, the finite-sample properties of our proposal are compared with several alternative methods for different distributions with heterogeneity in the location and in the scale parameters. We also carry out basic inference about the estimators. In addition, we evaluate how the results are affected when the covariates are correlated and binary explanatory variables are included. Note that Hall and Yao (2005) and Newey and Stoker (1993) only consider continuous covariates; indeed, not many papers to date have dealt with discrete covariates in single-index models. One exception is Horowitz and Härdle (1996), who focused on analysing a direct estimator for the effect of the discrete covariates. Elsewhere, methods such as those proposed by Härdle et al. (1993), Hristache et al. (2001) and Delecroix et al. (2003), while allowing dummy (binary) variables to be incorporated, do not consider the consequences of their inclusion.

2. Methods

Let us denote the vector of covariates $X = (X_1, \ldots, X_d)^\top$ and let $f(\cdot|\mathbf{x})$ be the density function of Y given $X = \mathbf{x}$, where $\mathbf{x} = (x_1, \ldots, x_d)$ is a fixed vector where $f(y|\mathbf{x}) = f_{\theta_0}(y|\theta_0^\top \mathbf{x})$, where $f_{\theta_0}(\cdot|\theta_0^\top \mathbf{x})$ is the conditional density of Y given $\theta_0^\top X = \theta_0^\top \mathbf{x}$ and θ_0 is the parameter vector to be estimated. Furthermore, we assume that $F(y|\mathbf{x}) = F_{\theta_0}(y|\theta_0^\top \mathbf{x})$ is its conditional cumulative distribution function. For any θ_0 and any nonzero real number λ, then vector θ_0 can be replaced by $\lambda\theta_0$. This means that the conditional distribution of the response given $X = \mathbf{x}$ only depends on this covariate vector via the linear combination $t = \theta_0^\top \mathbf{x}$. If we choose any nonzero real number λ, then, since there is a one-to-one correspondence between t and λt, it is also true that the conditional distribution only depends on the covariate vector via the linear combination $\lambda\theta_0^\top \mathbf{x}$. Consequently, infinitely multiple choices exist for the single-index parameter vector θ_0. The usual way to solve this identification problem is to introduce a scale constraint, for example $||\theta_0|| = 1$ or fixing one component of θ_0 to be equal to one. In practice, the identification problem implies that the signs of the effects of the covariates on the dependent variable are not identified but are comparable, i.e., two parameters with different sign indicate opposite effects and, if variables are measured in the same scale, then their corresponding parameter estimates can be compared directly.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of the dependent variable and the covariates, where $X_i = (X_{i1}, \ldots, X_{id})^\top$ and it is assumed that at least one covariate is continuous. Let K be a nonnegative kernel and h_1, h_2 two positive bandwidths. In line with Bashtannyk and Hyndman (2001), the kernel conditional density estimator is:

$$\hat{f}_\theta(y|t) = \frac{\hat{r}(t,y)}{\hat{s}(t)}, \tag{2}$$

where

$$\hat{s}(t) = \hat{s}_{h_1}(t) = \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{t - \theta^\top X_i}{h_1}\right) \tag{3}$$

and the product bivariate kernel density estimator is used for $\hat{r}(t,y)$; see Chapter 6 of Scott (2015). The product kernel is just a simple way to smooth using multiplicative weights, so:

$$\hat{r}(t,y) = \hat{r}_{h_1,h_2}(t,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) K\left(\frac{y-Y_i}{h_2}\right). \tag{4}$$

We use a Gaussian kernel, and the smoothing parameters are calculated using alternative criteria considering the estimator type, i.e., the parameter vector, the conditional density, the conditional distribution or the conditional mean.

In line with Hall, Wolff and Yao (1999), the kernel estimator of the conditional distribution function is:

$$\widehat{F}_\theta(y|t) = \frac{\widehat{R}(t,y)}{\hat{s}(t)},$$

where

$$\widehat{R}(t,y) = \widehat{R}_{h_1,h_2}(t,y) = \frac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) \mathbf{K}\left(\frac{y-Y_i}{h_2}\right)$$

and $\mathbf{K}$ is the kernel distribution function.

2.1. Maximum conditional likelihood estimation

If we know F_θ except for the value of the index vector θ (a highly unrealistic assumption), then we can define the following theoretical conditional likelihood function:

$$\tilde{L}_n(\theta) = \prod_{i=1}^{n} f_\theta(Y_i|\theta^\top X_i).$$

Maximising this function is equivalent to maximising its logarithm:

$$\tilde{\ell}_n(\theta) = \frac{1}{n}\log\left(\tilde{L}_n(\theta)\right) = \frac{1}{n}\sum_{i=1}^{n}\log f_\theta(Y_i|\theta^\top X_i). \tag{5}$$

Here, the ideal estimator should maximise the theoretical log-likelihood

$$\tilde{\theta}_n = \arg\max_\theta \tilde{\ell}_n(\theta).$$

In practice, f_θ (or F_θ) is unknown and so, we need to estimate it and plug it into the logarithm of the theoretical conditional likelihood function.

We propose to maximise the kernel estimation of the log-likelihood function defined in (5) with respect to θ and to the two smoothing parameters, h_1 and h_2. At this point, we note that, in the kernel estimation, when a smoothing parameter selector is obtained by optimising some criteria, such as the integrated square error or the likelihood function, which required computing a kernel estimator; using the whole observed data

set, $(X_1,Y_1),\ldots,(X_n,Y_n)$, produces undersmoothing of the optimal smoothing parameter values; see Silverman (1986). As a consequence, we need to modify the estimated likelihood with a leaving-one-out procedure so as not to pick artificially small bandwidths. Let $\hat{f}_\theta^{-i}(Y_i|\theta^\top X_i)$ be the estimator defined in (2), where the sum in (3) and (4) runs over $j \neq i$. Then, we define the leaving-one-out estimated conditional log-likelihood:

$$\hat{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_\theta^{-i}(Y_i|\theta^\top X_i). \tag{6}$$

Given h_1 and h_2, the final maximum conditional likelihood estimator is defined as

$$\hat{\theta}_n = \arg\max_\theta \hat{\ell}_n(\theta).$$

The estimation procedure including the two smoothing parameters h_1 and h_2 will be described in sub-section 2.3. A similar procedure based on the leave-one-out estimator of the hazard rate model was proposed by van den Berg et al. (2021). We point out that it can be difficult to avoid local optima in the maximisation of the log-likelihood in (6). Considering the existence of local optima, in the described estimation procedure we checked how initial values for the smoothing parameters affect the final estimation. We have observed that the final estimation is practically not affected by the initial values of the covariate coefficients.

2.2. Properties

In this sub-section we study the properties of $\hat{\theta}_n$. Let the score function be defined as the expected log-likelihood:

$$\ell(\theta) = E(\tilde{\ell}_n(\theta)).$$

We start by proving that the true parameter vector, θ_0, can be characterised as the maximiser of the score function. The existence of that function is the only condition required:

A1: $E(\log f_\theta(Y_i|\theta^\top X_i)) < \infty$ for any θ

Theorem 1. *The true single-index parameter, θ_0, is the maximiser of the score function, i.e., $\theta_0 = \arg\max_\theta \ell(\theta)$.*

To establish the main results for the estimator, we need to assume some further conditions:

A2: $E(X|\theta_0^\top X, Y) = E(X|\theta_0^\top X)$

A3: $E(XX^\top) < \infty$ componentwise.

Condition A2 is a technical one needed to prove our theoretical results. It essentially means that all the information needed to predict the values of the explanatory variables

given the index and the response variable is contained just in the index. Assumption A2 also implies exogeneity of the explanatory variables, i.e., covariates are known previous to the response.

The two bandwidths h_1, h_2 should fulfill the following conditions

A4: $\sqrt{n}h_1^4 \to 0$, $\sqrt{n}h_2^2 \to 0$, $nh_1^6 \to \infty$ and $h_1, h_2 \to 0$ when $n \to \infty$.

Consider f_{θ_0} the bivariate joint density function of $(\theta_0^\top X, Y)$ and $f_{\theta_0^\top X}$ the marginal density function of $\theta_0^\top X$. Finally, let $\ell^{[1]}(\theta_0) = \nabla_\theta \ell(\theta)|_{\theta=\theta_0}$ denote the gradient of $\ell(\theta)$ over θ evaluated in θ_0. Further, let $\ell^{[2]}(\theta)$ denote the Hessian matrix of $\ell(\theta)$. The following regularity conditions are also assumed.

A5: The derivatives $\frac{\partial^j}{\partial^j u}\frac{\partial^k}{\partial^k v} f_{\theta_0}(u,v)$, $\frac{d^j}{d^j u} f_{\theta_0^\top X}(u)$ and $\frac{d^j}{d^j u} E(X|\theta_0^\top X = u)$ exist for $j = 1,2,3$ and $k = 1,2$.

A6: The function $h(\mathbf{x}, y) = \frac{\partial}{\partial \theta_j} f_\theta(\theta^\top \mathbf{x}, y)_{\theta=\theta_0}$ is continuous and $\frac{\partial^2}{\partial^2 \theta_j} f_\theta(\theta_0^\top \mathbf{x}, y)_{\theta=\theta_0}$ exists.

A7: The Hessian matrix $\ell^{[2]}(\theta^*)$ is positive definite for θ^* belonging to a neighbourhood of θ_0.

Now we can state the first result for the proposed estimator.

Lemma 1. *Under A1, A4 and A6 we have* $\hat{\theta}_n - \theta_0 = -\left[\hat{\ell}_n^{[2]}(\hat{\theta}_n^*)\right]^{-1} (\hat{\ell}_n^{[1]}(\theta_0) - \ell^{[1]}(\theta_0))$, *where* $\hat{\theta}_n^*$ *is between* $\hat{\theta}_n$ *and* θ_0.

Theorem 2. *Under A1-A7, we have* $\hat{\theta}_n \to \theta_0$ *in probability.*

Theorem 3. *Let us assume conditions A1-A7. Then, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, \Sigma), \tag{7}$$

where

$$\Sigma = \Sigma_2 \Sigma_1 \Sigma_2^\top, \tag{8}$$

$$\Sigma_2 = \left[\ell^{[2]}(\theta_0)\right]^{-1}$$

and

$$\Sigma_1 = E\left[\nabla_\theta \log(f_\theta(Y|\theta^\top X))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(Y|\theta^\top X))_{\theta=\theta_0})^\top\right]$$
$$= \int (\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0})^\top f(\mathbf{x}, y) d\mathbf{x} dy.$$

All the proofs can be found in the Supplementary Material.

The asymptotic variance-covariance matrix in (8) is different from the one obtained by Delecroix et al. (2003). These authors obtained this matrix from $\tilde{\ell}_n(\theta)$ defined in (5)

and took into account the almost sure convergence of the parameter estimator and the weak convergence of $\hat{\ell}_n(\theta)$, defined in (6), and some of its partial derivatives. Instead, to obtain the asymptotic variance-covariance matrix, we take into account that θ_0 is estimated by maximising the kernel estimator of the conditional likelihood function $\hat{\ell}_n(\theta)$ defined in (6).

2.3. Estimation procedure

To obtain $\hat{\theta}_n$, h_1 and h_2 we have used an algorithm in two steps. The first step aims to obtain $\hat{\theta}_n$ by maximising the likelihood function in (6) given fixed values for the smoothing parameters h_1 and h_2. In the second step the smoothing parameters are recalculated by maximising the same likelihood function given the values of $\hat{\theta}_n$ obtained in the previous step. Both steps are repeated until convergence. In the first step the initial values of the smoothing parameters are given by $h_1 = a\hat{\sigma}_{\theta^\top X} n^{-2/13}$ and $h_2 = a\hat{\sigma}_Y n^{-4/13}$, where $a > 0$ and $\hat{\sigma}_{\theta^\top X}$ and $\hat{\sigma}_Y$ are the empirical standard deviations (see Silverman (1986) for rule-of-thumb smoothing parameters in kernel density estimation). The sample size orders, $n^{-2/13}$ and $n^{-4/13}$, respectively for the two bandwidths, are chosen in order to fulfill the asymptotic assumptions for the bandwidths needed for Condition A4. We have observed that initial values of the smoothing parameters considerably affect the final estimation. Initially we used $a = 1$ but it is recommended to consider a grid of values around 1. The initial values of the covariate coefficients hardly affect the results, so to start the algorithm we set all these coefficients equal to 1. To maximise the likelihood function in the first step, we use the function "optim()" with the default optimization method ("Nelder-Mead") of the "stats" R package. In the second step, to recalculate the values h_1 and h_2 we also use function "optim()" but with optimization method "L-BFGS-B". We need to define limits for the smoothing parameters because it is known that $\hat{\ell}_n(\theta) \to \infty$ as $h_1, h_2 \to 0$. The limits are defined as $(c_1^{(1)}\hat{\sigma}_{\theta^\top X} n^{-2/13}, c_2^{(1)}\hat{\sigma}_{\theta^\top X} n^{-2/13})$ for h_1 and $(c_1^{(2)}\hat{\sigma}_Y n^{-4/13}, c_2^{(2)}\hat{\sigma}_Y n^{-4/13})$ for h_2, for some $c_1^{(j)} < c_2^{(j)},\ j = 1,2$.

Our two-step algorithm is designed to guarantee the conditions established in the theoretical properties shown in the previous sub-section. In practice, we are selecting the best estimation in a set of pre-fixed smoothing parameters which are calculated taking into account the sample size and the scale of the dependent variable and the index.

To estimate the variance-covariance matrix in (8) we calculate the corresponding derivatives of the leave-one-out kernel estimation of conditional log-likelihood defined in (6). Asymptotic normality inference, based on (7), is carried out using the estimated variance-covariance matrix, replacing theoretical derivatives by estimated ones (kernel estimator of the gradient $\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0}$ is direct). For kernel estimator of $\ell^{[2]}(\theta_0)$ see Lemma 9 in the Supplementary Material.

2.4. Marginal effects estimation

For a given $\theta = \theta_0$, using the conditional distribution function we can obtain the p-th conditional quantile: $Q_\theta(p|\theta^\top \mathbf{x}) = F_\theta^{-1}(p|\theta^\top \mathbf{x})$, i.e., $F_\theta(y_p|\theta^\top \mathbf{x}) = p$ where $p \in$

$(0,1)$. As in any generalised linear model, comparing marginal effects is equivalent to comparing parameters, i.e., for two covariates X_k and $X_{k'}$, with $k \neq k'$, we obtain:

$$\frac{\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_k}}{\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_{k'}}} = \frac{\theta_k}{\theta_{k'}},$$

where:

$$\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_k} = -\frac{\frac{\partial F_\theta(Q_\theta(p|\theta^\top \mathbf{x})|t)}{\partial t}_{t=\theta^\top \mathbf{x}} \cdot \theta_k}{f_\theta(Q_\theta(p|\theta^\top \mathbf{x})|\theta^\top \mathbf{x})}. \tag{9}$$

For estimating the marginal effects we will use kernel estimators for $f_\theta(y|\theta^\top \mathbf{x})$, $F_\theta(y|\theta^\top \mathbf{x})$ and their derivatives, as shown below.

The kernel estimator of the index marginal effects on the conditional distribution function is:

$$\frac{\partial \widehat{F}_\theta(y|t=\theta^\top \mathbf{x})}{\partial t} = \left[\frac{\hat{R}'_{h_1,h_2}(\theta^\top \mathbf{x}, y)}{\hat{s}_{h_1}(\theta^\top \mathbf{x})} - \widehat{F}_\theta(y|\theta^\top \mathbf{x})\frac{\hat{s}'_{h_1}(\theta^\top \mathbf{x})}{\hat{s}_{h_1}(\theta^\top \mathbf{x})}\right],$$

where

$$\hat{R}'_{h_1,h_2}(t,y) = \frac{1}{nh_1^2h_2}\sum_{i=1}^{n} K'\left(\frac{t-\theta^\top X_i}{h_1}\right)\mathbf{K}\left(\frac{y-Y_i}{h_2}\right)$$

and

$$\hat{s}'_{h_1}(t) = \frac{1}{nh_1^2}\sum_{i=1}^{n} K'\left(\frac{t-\theta^\top X_i}{h_1}\right),$$

where K' is the first derivative of the kernel.

In this paper, we obtained the marginal effects using kernels estimators of the different functions that appear in the expression (9). The smoothing parameters of the kernel estimator of conditional density can be calculated using the sample size orders of reference rules obtained in Bashtannyk and Hyndman (2001). The kernel estimator of the conditional distribution and its derivatives are obtained directly from the estimated conditional density. Considering that in this paper the aim of estimating marginal effects is purely descriptive, we have obtained the values of smoothing parameters subjectively from graphic visualization. However, a double-cross-validation approach as suggested van den Berg et al. (2021) can be used.

2.5. Scoring rules for prediction

To evaluate the goodness of fit and the predictive capacity of the single-index model, a variety of measures is available. Gneiting and Raftery (2007) present an exhaustive review of different families of scoring rules for moments, density and distributional forecasts. We use three types of score described in Gneiting and Raftery (2007).

The predictive model choice criterion (PMCC) selects the best model based on the first two moments of the predicted values, i.e., the mean and the variance, as follows

$$PMCC = -\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}\left(\theta^\top X_i\right)\right]^2 - \hat{\sigma}^2\left(\theta^\top X_i\right), \tag{10}$$

where $\hat{m}\left(\theta^\top X_i\right)$ is the kernel estimator of the conditional expectation $E\left(Y_i|\theta^\top X_i\right)$ and $\hat{\sigma}^2\left(\theta^\top X_i\right)$ is estimated with the kernel estimates of both expectations as follows:

$$\hat{\sigma}^2\left(\theta^\top X_i\right) = \hat{E}\left(Y_i^2|\theta^\top X_i\right) - \left[\hat{E}\left(Y_i|\theta^\top X_i\right)\right]^2,$$

where

$$\hat{E}\left(Y_i|\theta^\top X_i\right) = \hat{m}\left(\theta^\top X_i\right) = \frac{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right)}$$

and

$$\hat{E}\left(Y_i^2|\theta^\top X_i\right) = \frac{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) Y_i^2}{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right)}.$$

Here h_1 is calculated using the optimal sample size order ($n^{-1/5}$) to estimate the conditional expectation and considering the scale of the dependent variables.

The logarithmic scoring rule is calculated as

$$\hat{\ell}(\theta) = \sum_{i=1}^{n} \log\left[\hat{f}\left(Y_i|\theta^\top X_i\right)\right]. \tag{11}$$

From $\hat{l}(\theta)$ other widely used criteria such as the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) can be obtained.

For the p-quantile prediction of the dependent variable, Y, the goodness of fit criterion proposed by Koenker and Bassett (1978) for quantile regression is:

$$\begin{aligned} QE_p(\theta) = &\frac{1}{n}\sum_{i=1, Y_i > \hat{Q}_\theta(p|\theta^\top X_i)}^{n} p\ \left|Y_i - \hat{Q}_\theta(p|\theta^\top X_i)\right| \\ &+ \frac{1}{n}\sum_{i=1, Y_i \le \hat{Q}_\theta(p|\theta^\top X_i)}^{n} (1-p)\ \left|Y_i - \hat{Q}_\theta(p|\theta^\top X_i)\right|\ , \end{aligned} \tag{12}$$

where $\hat{Q}_\theta(p|\theta^\top X_i)$ is the kernel conditional quantile estimator based on the kernel estimator of the conditional distribution function. For a set of probabilities $p_1, \ldots, p_k$, we define $QE = \frac{1}{k}\sum_{j=1}^{k} QE_{p_j}(\theta)$ and its corresponding weighted version, $WQE = \frac{1}{k}\sum_{j=1}^{k} p_j QE_{p_j}(\theta)$.

3. Simulation study

We carry out a simulation study, the aim being to evaluate the finite-sample properties of our estimator. The properties of the parameter estimator, $\hat{\theta}$, are summarised in the Supplementary Material and the basic inferences about the value of these parameters are presented in this section. The results are obtained using a Gaussian kernel.

We compare the variance, the bias and the mean square error (MSE) of the estimated parameters in the vector $\hat{\theta}$, using our flexible maximum conditional likelihood (FMCL) estimator and three alternatives. The first is based on fitting the single-index model to individual conditional expected values as proposed by Härdle et al. (1993) (hereinafter, HHI). The second alternative is based on Delecroix et al. (2003) (hereinafter, DHH), where we use as our initial parameters those obtained with the HHI method which are $\sqrt{n}$-consistent. The third is the direct method proposed by Hristache et al. (2001) (hereinafter, HJS).

We analyse six different conditional distributions for the dependent variable Y, two symmetric distributions (zero skewness) and four right-skewed distributions. The conditional distributions are shown in Table 1.

Table 1. *Conditional distributions for dependent variable as a function of the linear index $\theta^\top \mathbf{x}$ for the simulation study.*

Skewness	Distribution	Parameters	Density
Zero	normal	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\sqrt{2\pi\lvert\theta^\top \mathbf{x}\rvert^2}} \exp\left(-\frac{(y-\theta^\top \mathbf{x})^2}{2\lvert\theta^\top \mathbf{x}\rvert^2}\right)$
	logistic	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\lvert\theta^\top \mathbf{x}\rvert} \frac{\exp\left(\frac{(y-\theta^\top \mathbf{x})}{\lvert\theta^\top \mathbf{x}\rvert}\right)}{1+\exp\left(\frac{(y-\theta^\top \mathbf{x})}{\lvert\theta^\top \mathbf{x}\rvert}\right)}$
Positive	lognormal	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{y\sqrt{2\pi\lvert\theta^\top \mathbf{x}\rvert^2}} \exp\left(-\frac{(ln(y)-\theta^\top \mathbf{x})^2}{2\lvert\theta^\top \mathbf{x}\rvert^2}\right)$
	Weibull	$(\alpha = 1, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\lvert\theta^\top \mathbf{x}\rvert} \exp\left(-\frac{y}{\lvert\theta^\top \mathbf{x}\rvert}\right)$
	Champernowne	$(\alpha = 1, M = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{\lvert\theta^\top \mathbf{x}\rvert}{(y+\lvert\theta^\top \mathbf{x}\rvert)^2}$
		$(\alpha = 2, M = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{2\lvert\theta^\top \mathbf{x}\rvert^2 y}{(y^2+\lvert\theta^\top \mathbf{x}\rvert^2)^2}$

For our two choices of symmetric distribution, the logistic distribution has more kurtosis and heavier tails than the normal distribution. If we compare our selection of right-skewed distributions, we find that the Champernowne or log-logistic has a heavier tail than the lognormal and the Weibull; see Buch-Larsen et al. (2005) for a description of the Champernowne distribution.

In our simulation study, we use six vectors of covariates X that we identify as vectors V1, V2, V3, V4, V5 and V6. For the first three $\theta^\top = (1, 1.3, 0.5)$ and for the fourth $\theta^\top = (1, 1.3, 0.5, 0.8)$. The values in vector V1 are generated from three uncorrelated standard normal distributions. The vectors V2 and V3 are trivariate normal distributions with correlated marginals. For V2 the components are three standard normal distributions whose covariances are $cov(X_k, X_{k'}) = 0.3$ for $k \neq k'$ and $k, k' = 1, 2, 3$. The same holds for V3 but with covariances $cov(X_1, X_2) = cov(X_2, X_3) = 0.7$ and $cov(X_1, X_3) = 0.5$. Vector V4 consists of V1 and a binary variable whose values are generated from a Bernoulli distribution with probability 0.4, independent of the three components of V1. Furthermore, the number of categorical covariates is usually greater than one. We have carried out an alternative simulation study using two new vectors of covariates V5 and V6, with $\theta^\top = (1, 1.3, 0.5, 0.8)$. Vector V5 consists of two independent standard normal variables and two binary variables whose values are generated from two Bernoulli distributions with probabilities 0.4 and 0.7, respectively. The covariate vector V6 includes the same two binary variables, one lognormal with mean 0 and σ equal to 0.5 and one with a standard normal distribution.

We generate 500 samples of size $n = 100$, 500 and $2,000$ and calculate the bias, the standard deviation (STD) and the MSE of the estimators using each method, FMCL, HHI, DHH and HJS. The results of the simulation study show that the proposed FMCL estimator is the most suitable when the conditional distribution is right-skewed and also when the tail of the conditional distribution is heavy. Moreover, the FMCL is more robust to multicollinearity and to the presence of binary and asymmetric covariates.

3.1. Basic inference

Power analysis of hypothesis tests is fundamental to determining whether the effect of a covariate is significantly different from zero. The null hypothesis for each parameter is $H_0: \theta_k = 0$, $k = 1, \ldots, d$ and as an alternative hypothesis we assume that the sign of the parameter is known, i.e., $H_1: \theta_k > 0$, $k = 1, \ldots, d$. The statistic test is $Z = \hat{\theta}_j / se(\hat{\theta}_j)$, where *se* indicates the standard error. The statistic Z asymptotically follows a $N(0,1)$ distribution. To obtain the power of the test we calculate the proportion of times that we reject the null hypothesis in the 500 samples obtained from each analysed conditional distribution and sample size. Alternatively, we also analyse the power of the test when the null hypothesis is $H_0: \theta_2 = \theta_3$ and the alternative hypothesis $H_1: \theta_2 > \theta_3$. Again, we know that the alternative hypothesis is true. The statistic for this test is $Z = (\hat{\theta}_2 - \hat{\theta}_3) / se(\hat{\theta}_2 - \hat{\theta}_3)$.

Table 2. *Power of the test for skewed distributions. The values are calculated using the 500 samples for each skewed distribution in Table 1.*

		Lognormal		Weibull		Champernowne $\alpha = 1$		Champernowne $\alpha = 2$	
	H_0	$n = 500$	$n = 2,000$	$n = 500$	$n = 2,000$	$n = 500$	$n = 2,000$	$n = 500$	$n = 2,000$
V1	$\theta_2 = 0$	1.000	1.000	0.864	0.996	0.722	0.970	0.984	0.998
	$\theta_3 = 0$	1.000	1.000	0.876	0.998	0.702	0.972	0.992	1.000
V4	$\theta_2 = 0$	1.000	1.000	0.856	1.000	0.636	0.908	1.000	1.000
	$\theta_3 = 0$	1.000	1.000	0.828	1.000	0.622	0.902	1.000	1.000
	$\theta_4 = 0$	1.000	1.000	0.770	0.984	0.584	0.862	0.996	1.000
V1	$\theta_2 = \theta_3$	1.000	1.000	0.882	0.996	0.730	0.976	0.988	1.000
V4	$\theta_2 = \theta_3$	1.000	1.000	0.662	1.000	0.598	0.880	0.998	1.000

Table 3. *Percent of no-rejection of null hypothesis. The values are calculated using the 200 samples for each distribution in Table 1.*

		Normal		Logistic		Lognormal	
	H_0	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
V1	$\theta_4 = 0$	0.848	0.955	0.942	0.985	0.696	0.850
V4	$\theta_5 = 0$	0.828	0.980	0.992	0.980	0.345	0.890
		Weibull		Champernowne $\alpha = 1$		Champernowne $\alpha = 2$	
	H_0	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
V1	$\theta_4 = 0$	0.850	0.965	0.530	0.570	0.752	0.720
V4	$\theta_5 = 0$	0.924	0.965	0.478	0.795	0.720	0.770

The results for symmetric distributions have a power about 100% for almost all tests when $n \geq 500$, these results are shown in the Supplementary Material. Here we focus on the results for the power of tests for skewed distributions.

Table 2 shows the powers of the two tests proposed for skewed distributions. Both tests are at the 95% confidence level. These results indicate that when $n = 500$ the power decreases considerably for the Weibull and the Champernowne distribution with $\alpha = 1$, compared to a larger sample size, $n = 2,000$.

To analyse the percent of times the null hypothesis that the parameter is equal to zero is not rejected, we have designed an alternative reduced simulation study that consists of adding a new covariate with associated parameter equal zero in the estimation procedure; this implies to re-estimate the parameters. To reduce the computation time, instead of 500 replicates, we use 200 replicates of sizes $n = 500$ and $n = 2,000$. The null hypothesis is $H_0 : \theta_j = 0,\ j = 4,5$, and the results of the percent of no-rejection of the null hypothesis, for the models described in Table 1 and using extended covariate vectors, are shown in Table 3. For $n = 500$ the results for skewed distributions are poorer than those obtained for a symmetric distributions. For $n = 2,000$, in general, the results improve compared to a smaller sample size, except for the Champernowne distribution, which is heavy tailed. These results suggest that if the dependent variable is asymmetric, a transformation to achieve a symmetric distribution should be suitable.

4. Data analysis and model estimations of automobile claim costs

In this section we analyse the effect of risk factors on the distribution of the cost per automobile claim in a real case study. We show that single-index models constitute a new tool for identifying the influence of some of those covariates that are known to the insurer at the beginning of the contract or during the coverage period. We estimate the single-index model coefficients with the FMLC method. The results are obtained using a Gaussian kernel. Some parametric models based on Weibull, gamma, log-normal and log-logistic distributions, which are not reported here, produced poor fits. Furthermore, significant effects of the covariates were not found.

We analyse a data set obtained from a Spanish insurance company. The original portfolio consists of policyholders between age 18 and 35, who underwrote a motor insurance policy and accepted a telematics engine that allows the company to gather data on the policyholder's driving behaviour. In the available data set, all claims are settled. In the original data set, a few claims result from no fault agreements between insurers, in these cases the amount recorded is equal to the legally established cost. Claims regulated by a no-fault agreement were excluded from our analysis. Hence, our data are not censored. Those in the no-fault agreement had to be removed because there was no information on the true cost of the claim, which could be lower or higher than the amount established by the agreement. To estimate the proposed single-index model, we have selected a sample of $n = 489$ car insurance policyholders who reported at least one claim in 2011. Furthermore, we have also selected another sample of 100 policyholders to carry out a predictive analysis. The claims correspond to third-party liability accident. For each policyholder in the sample, the total incurred losses and the number of claims along the year is known, the ratio between both values is equal to the yearly mean claim cost per policy. The cost refers to incurred and paid losses.

For each policyholder, we have information about the following covariates (labels in parentheses): cost per policyholder in thousands of euros (cost), age in years (age), number of years holding a driving licence (agelic), age of car in years (agecar), a binary indicator equal to 1 if car is parked in a garage overnight and 0 otherwise (parking), annual distance driven in thousands of kilometres (tkm), percentage of kilometres driven at night (nightkm), percentage of kilometres driven on urban roads (urbankm) and percentage of kilometres driven above the speed limit (speedkm). These data correspond to a sample of insureds for whom the company collected driving behaviour information employing a telematics device installed in their vehicle. Thus, "tkm", "urbankm", "nightkm" and "speedkm" correspond to the so-called "telematics covariates" that capture policyholders' driving style and driving patterns. We do not include the gender variable in the model because European Union regulations prohibit discrimination between men and women in the field of insurance premiums; for more information on these data, see Guillen et al. (2019).

Table 4 shows our descriptive statistics for the cost per policyholder variable in the original scale, transformed into logarithmic form (log(cost)), and information on the

covariates. We show that our data set contains one extreme observation for the response variable corresponding to a claim that exceeded €130,000 (natural logarithm close to 5).

Table 4. *Descriptive statistics of the variables in the claim costs dataset.*

	Mean	Std.	Min.	Q25	Median	Q75	Max.
cost	1.810	6.191	0.018	0.417	0.818	1.878	130.870
log(cost)	-0.145	1.128	-4.031	-0.874	-0.201	0.630	4.874
age	27.009	3.246	20.586	24.496	26.820	29.886	34.067
agelic	6.429	2.833	2.001	4.337	5.864	7.992	14.686
agecar	8.916	4.162	2.111	5.777	7.943	11.370	20.468
parking	0.763	0.426	0.000	1.000	1.000	1.000	1.000
tkm	8.356	4.530	1.220	5.174	7.549	10.635	35.105
nightkm	7.514	6.504	0.044	2.979	5.841	9.954	42.830
urbankm	27.127	14.163	3.810	16.565	24.401	35.245	80.659
speedkm	7.203	7.100	0.122	2.286	4.969	9.403	48.002

Q25 and Q75 are the first and third quartiles.

log(•) denotes natural logarithm.

The single-index models that we estimate in this section are fitted using "log(cost)" as the dependent variable. Table 5 shows the results of the estimated parameters ($\hat{\theta}$) of the single-index models when using our FMCL method and three different covariate vectors, that is, all the explanatory variables, only the telematics variables and only the traditional rating factors, i.e., the non-telematics covariates. Note that the smoothing parameters h_1 and h_2 obtained for each estimated parameter vector are the same. This is just a coincidence which does not occur for other analyses. We establish "speedkm" as the variable with the constrained coefficient $\theta_1 = 1$ while for the model with the non-telematics variables we use "age". This is convenient because the nature of these covariates makes interpretation straightforward in this context. The reason we opt to fix the effect of the speed variable is that we believe that high speeds result in a greater risk of being involved in severe accidents, which in turn are more costly than minor accidents. For each estimated parameter $\hat{\theta}_j,\ j = 2,\ldots,8$, we test individual significance.

We also estimated the parameters with methods HHI and DHH, that are shown in the Supplementary Material. These results indicate that the estimated parameters with the HHI method have larger standard errors than those obtained with our method. In general, HHI and DHH are more sensitive than FMCL to the selection of the covariate vector.

Table 5 shows that the sign of the effect of the telematics variables is unchanged when comparing the model with all the variables and that one using only the telematics variables. This indicates that, in the single-index model of the logarithm of the cost per policyholder, driving patterns matter. For example, in the model with all variables, as the effect of "speedkm" is the reference and the effects of "age" and "agelic" are positive and

we conclude that the longer the driving experience is, the greater the risk is; however, driving experience is associated with "tkm', the effect of which is negative.

As shown above in Section 2, given that we assume $\theta_1 = 1$, even when the signs of the coefficients of the explanatory variables are not identified, we are still able to analyse the relation between these effects. For example, in Table 5, we observe, on the one hand, that "tkm" has an opposite effect to "speedkm", i.e., excess speed can be offset by the amount of time spent driving (measured here in terms of total distance driven). On the other hand, the coefficients of "nightkm" and "urbankm" present the same sign as the "speedkm" coefficient. Thus, if a higher percentage of driving at speeds above the limit implies higher values of the index, then the same is true as night-time and/or urban driving increase.

Table 5. *Estimated parameters and their significance (p-value in parentheses) for the single-index model in the claim cost data set.*

	Model		
	All variables	Only telematics	Only non-telematics
speedkm	1.000	1.000	–
age	0.153 (<0.0001)	–	1.000
agelic	0.097 (0.0034)	–	-0.246 (<0.0001)
agecar	-0.107 (<0.0001)	–	0.074 (<0.0001)
parking	-0.162 (0.2570)	–	-0.655 (<0.0001)
tkm	-0.044 (0.0004)	-0.423 (<0.0001)	–
nightkm	0.117 (<0.0001)	0.089 (0.0005)	–
urbankm	0.141 (<0.0001)	0.080 (<0.0001)	–

$h_1 = 0.3857$ and $h_2 = 0.1488$

The first coefficient of each model is fixed and equal to 1.000.

The computational times are 4.28 minutes with all variables,

35.27 seconds with telematics and 21.49 seconds with no telematics

The values of the index do not have a direct interpretation. These values allow us to analyse how the shape of the conditional distribution and the marginal effects change. To analyse these results in greater detail, we use plots, shown now in the original scale of the cost per claim as opposed to their log-transformation. In Figures 1 and 2, we plot the index against the fitted mean of the model with all variables and with non-telematics variables only, the median and p-th quantiles with $p = 0.90$, $p = 0.95$ and $p = 0.99$ (the plot with only telematics variables is similar to Figure 1). The mean curve is estimated using the Nadaraya-Watson estimator of the regression function between the dependent variable and the estimated linear index. The median and the higher quantiles are estimated from the inverse of the estimated conditional distribution function. The smoothing parameters are calculated specifically for each estimated curve, i.e., for the kernel regression the order is $n^{-1/5}$ and for the quantile it is $n^{-1/3}$. The main result is that

the cost distribution conditional on the value of the index is not constant. Furthermore quantiles are not monotonic in the index. This is evidence that some combinations of the covariates lead to a conditional cost distribution with a longer tail than others. For example, Figure 1 shows that when the index takes values around 22.5 and around 31 the conditional distribution has a heavier tail than for the rest of the index domain. We have calculated the mean of the covariates for the policyholders with index values between 22 and 23 and the results indicate that these individuals tend to use night parking and the means for the telematics covariates (tkm, nightkm, urbankm and speedkm) are higher than the means for the whole sample. A second group of policyholders with heavier tail takes index values around 31. The means of the covariates for policyholders with index values between 30 and 32 indicate that these individuals also use parking and drive more than 20% of total kilometres above the speed limit. These features are not captured by the mean curve, which is flat; thus, we can conclude that using a single-index conditional distribution model prediction is helpful to insurance companies when setting up wider margins that correspond to the values of those predicted in the intervals where the conditional distribution presents a remarkable heavy tail.

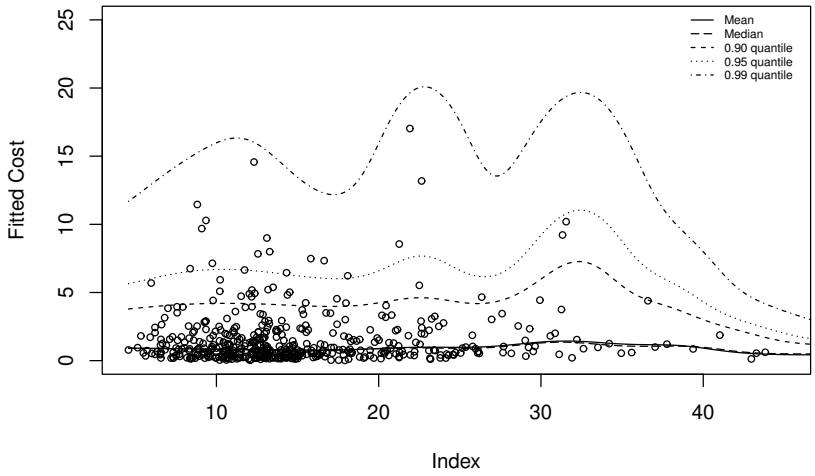

Figure 1. *Fitted values of the conditional mean (solid line) and quantiles (dotted and dashed lines) with all covariables in the model.*

When comparing the plot of the model with all the variables (Figure 1) with the plot of that with only the traditional rating variables (Figure 2), the benefits of including the telematics regressors become evident. By doing so, the intervals of the index corresponding to a conditional distribution with a longer tail are easily identified and, as a result, in such cases the insurance company estimates a slight increase in the median cost and a marked increase in the upper quantiles.

The single-index value provides a one-dimensional summary of the characteristics that discriminate between the policyholders in terms of the conditional cost distribution.

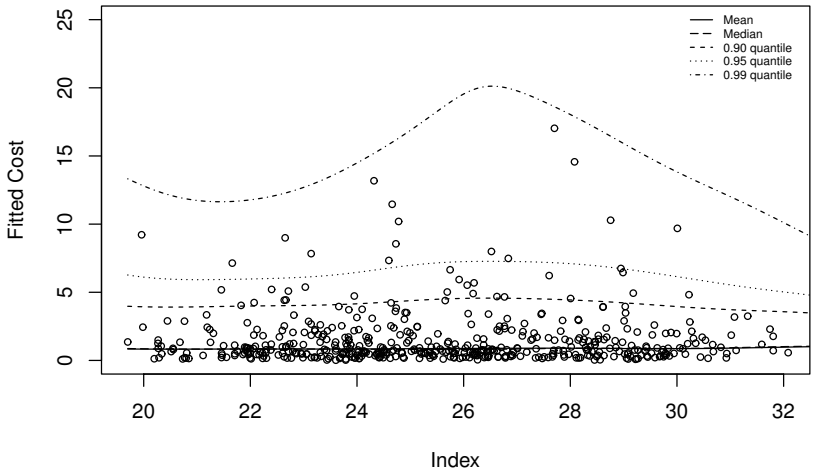

Figure 2. *Fitted values of the conditional mean (solid line) and quantiles (dotted and dashed lines) with only the non-telematics covariables in the model.*

4.1. Marginal effects on extreme quantiles

The study of marginal effects on extreme quantiles of the cost per policyholder, obtained from the derivative of the estimated inverse conditional distribution function, provides us with information about the changes of the risk of high losses when explanatory variables increase or decrease. Furthermore, we analysed to what extent the effects of the variables are different in the extremes and in the central values of the variable cost per policyholder. The results of the marginal effects have been obtained from the kernel estimates described in Subsection 2.4, using the significant parameters. These results are purely descriptive and show the flexibility of our proposal.

As we explained in Section 2, for a given vector of values of the covariates $\mathbf{x} = (x_1, \ldots, x_d)$, we estimated the marginal effects along a grid of values of the covariate x_k. We focused this analysis on telematics variables and studied some examples for certain policyholders. Specifically, once the grid for each telematics variable was fixed, we estimated the marginal effects for the younger and the older individuals, when the rest of the variables take their minimum values. Figures 3 and 4 show the marginal effects for variables "speedkm", and "nightkm", respectively, similar results for "tkm" and "ubankm" are shown in Supplementary Material. In general, we observe that marginal effects of telematics variables are different for the median and for the 0.95 quantile of the cost per policyholder. Furthermore, the marginal effects for the younger and older policyholders exchange their position when they are calculated at the median or when they are calculated at the 0.95 quantile. For example, in Figure 3 we see that the impact on the severity of a claim of exceeding approximately 10% or more kilometers over the posted speed limit is higher for older than for younger drivers at the 95% quantile (right plot) while it is lower at the median (left plot). Note that a negative marginal effect is possible because drivers that exceed speed limits by more that 10% could be more skilled than the rest

and for them the median cost may be lower (left plot) while the extreme quantile may be higher (right plot).

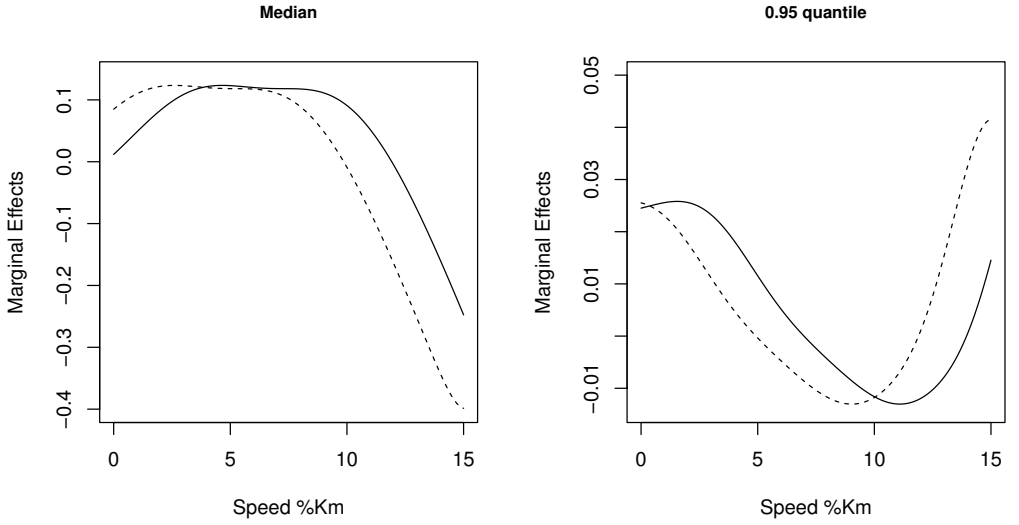

Figure 3. *Marginal effects on the median (left plot) and on the 0.95 quantile (right plot) of the cost per policyholder vs the percentage of kilometers with excess speed. Younger policyholder (solid line) and older policyholder (dashed line). The rest of covariates take their minimum values.*

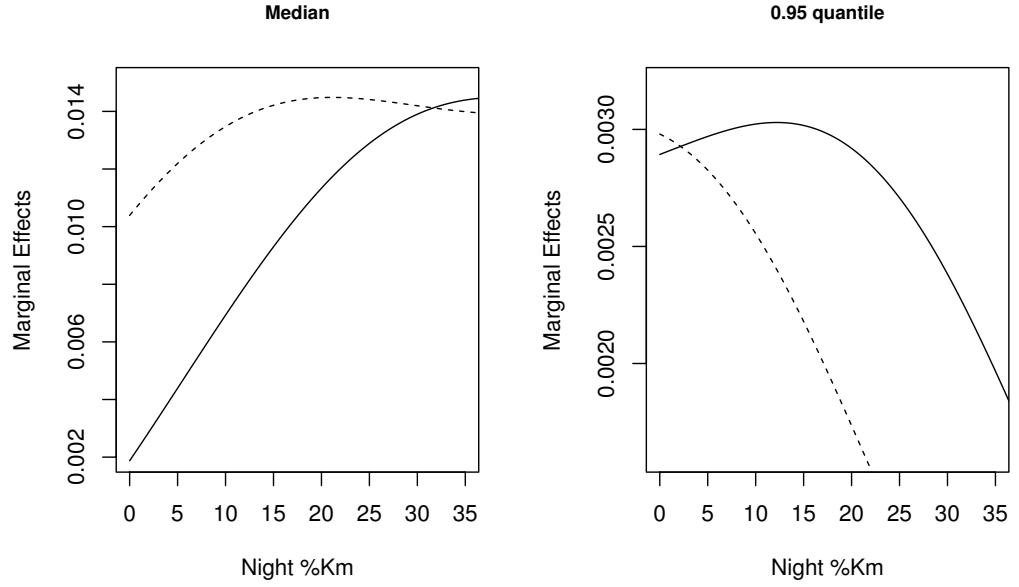

Figure 4. *Marginal effects on the median (left plot) and on the 0.95 quantile (right plot) of the cost per policyholder vs the percentage of kilometers at night. Younger policyholder (solid line) and older policyholder (dashed line). The rest of covariates take their minimum values.*

4.2. Predictive analysis of automobile red claim costs

To analyse the predictive capacity of the estimated single-index models with the three estimators FMCL, HHI and DHH, we used a sample of 100 cases, which were not included in the sample used to estimate the models in the previous subsection. The results using HHI and DHH are practically the same, only some differences are observed from the fourth decimal, for this reason we only analyse the results for HHI. In Table 6 we present the descriptive statistics of this new sample.

Table 6. *Descriptive statistics of the variables in the claim costs sample for prediction analysis.*

	Mean	Std.	Min.	Q25	Median	Q75	Max.
cost	1.557	2.080	0.030	0.438	0.820	1.880	13.584
Log(cost)	-0.138	1.107	-3.518	-0.827	-0.199	0.631	2.609
age	26.438	2.763	20.594	24.410	26.418	28.461	32.769
agelic	5.952	2.670	1.859	3.908	5.458	7.592	14.628
agecar	8.349	3.875	2.283	5.175	7.943	10.665	20.468
parking	0.790	0.409	0.000	1.000	1.000	1.000	1.000
tkm	7.644	4.006	0.560	5.018	7.267	9.876	23.336
nightkm	8.022	6.794	0.462	3.235	5.889	12.115	40.694
urbankm	29.232	14.522	10.266	17.386	27.811	36.246	85.553
speedkm	6.638	6.369	0.155	1.911	4.582	9.103	29.420

Q25 and Q75 are the first and third quartiles.

Table 7. *Criteria for the scoring rules for prediction. QE and WQE are multiplied by* 10*.*

Methods	FMCL	HHI
MSE	1.303	1.514
PMCC	-2.529	-2.662
$\hat{l}(\theta)$	-149.592	-153.595
AIC	313.184	321.191
BIC	331.420	339.427
$10 \times QE$	11.495	11.531
$10 \times WQE$	8.388	8.436

For each of the indices that were estimated with the alternative methods, all the criteria defined in Section 2.5 are calculated and presented in Table 7. In addition, the mean squared error (MSE) associated with the first order predicted conditional moment corresponding to the first sum of the PMCC criterion is also calculated. All these criteria are obtained with the estimated parameters of the single-index models that include all the covariates. These parameters are applied to the sample described in Table 6 to obtain the values of the index, i.e., the parametric part of the single-index model. To obtain the nonparametric estimation of the conditional functions used in the evaluated criteria, we

need to calculate the appropriate degree of smoothing in each case, which depends on the type of function, on the sample size and on the scale of the variable. So, the optimal bandwidth for the kernel estimation of the density function is of order $n^{-1/5}$ and for the distribution function and quantile it is of order $n^{-1/3}$.

Scoring rules are shown in Table 7, QE and WQE are calculated for a sequence of values for p between 0.5 and 0.999 in intervals of 0.001 units. The results show that the best fit is provided by the FMCL method for prediction in all cases.

5. Conclusion

The method proposed herein provides a full specification of the conditional distribution, while preserving the flexible nature of the single-index. Contrary to this principle, one limitation of the traditional approach to generalised linear modelling is the fact that the linear predictor is linked to the mean which, in general, is related to the location parameter of a given distribution that is assumed to be true.

In many contexts, heterogeneity is likely to be more closely associated with the shape of the distribution and not so much with location. This is precisely the case of the application presented as a case study herein. The use of a single-index model allows us to analyse all the components of the motor insurance claims cost distribution: that is, its mathematical expectation, its median, its quantiles and the marginal effects of the covariates at their different values.

Here, we have developed an estimator for the conditional distribution single-index model based on maximisation of the estimated conditional likelihood. We have used this approach to estimate the conditional distribution and, more specifically, its quantiles. This, today, is fundamental in data analysis, given that in certain applications a knowledge of the mean is not as interesting as a knowledge of other characteristics of the distribution. In our application, the estimation of the probability of a severe accident given some covariates, i.e., a cost larger than a fixed value, is a measure of the risk of driving unsafely.

From the expression of the marginal effect of a covariate on a given quantile, we have developed a way to interpret the estimated parameters of the index. Furthermore, we can also interpret the specific marginal effects for each insured individual.

Our main theoretical results demonstrate the asymptotic properties of the estimator for a vector of parameters in a conditional distribution single-index model and provide an expression for its covariance matrix. Likewise, the simulation study conducted herein demonstrates the power of the inference using the kernel estimator of the covariance matrix. These results are fundamental in situations in which the analyst does not have any prior knowledge for identifying the variables that are actually responsible for changes in the distribution of the dependent variable. The estimation of the variance-covariance matrix considering the possible censored data in line with what is described by Laudagé, Desmettre and Wenzel (2019) is a future goal.

Moreover, the simulation study shows how our method is, in fact, an improvement with respect to the finite-sample properties of certain known alternative methods, especially when the conditional distribution is skewed and has a long right tail. This is frequent in economic variables measuring revenues and expenses. The estimator proposed is a considerable improvement on the alternatives analysed, showing robustness in the presence of extreme values. However, for a distribution of this shape using a sample of small size, the results are still not especially good, but they can be improved with the use of a logarithmic transformation.

In the application described here, the observed characteristics of the insured drivers can be usefully employed to understand the distribution of claims cost. Additionally, if single-index models were implemented in practice, they would enable insurers to combine the cost per policyholder distribution with predictions about the expected number of claims, which is currently the baseline for premium calculation dependent on such covariates as age, number of years holding a driving licence, power of the vehicle, age of the vehicle, and so on. Moreover, when we include driving behaviour information in the model (that is, variables such as distance driven and a range of driving habits), our approach allows us to identify the values of the single-index that correspond to a long-tailed cost distribution and, therefore, to detect situations in which the probability of observing a large claim increases. In addition, our proposal presents better predictive scores and, therefore, more adjusted predictions than other existing alternatives.

SUPPLEMENTARY MATERIAL

SM: The file contains: 1. The proofs of the theoretical results in Section 2.2. 2. The results of simulation study related to the properties (bias, variance and MSE) of the alternative methods and the inference power of FMLC for symmetric distributions. 3. The results of application using HHI and HHD methods and additional plots marginal effects using FMLC method and are available in http://www.ub.edu/rfa/R/SORT-BCG/.

DS: Data set and R program used in the illustration of FMCL method in Section 4 are available in https://data.mendeley.com/datasets/py3kb2hn2b/1 and http://www.ub.edu/rfa/R/SORT-BCG/

Acknowledgements

This article is part of the I+D+i projects PID2019-105986GB-C21 and grant TED2021-130187B-I00, financed by MCIN/ AEI/10.13039/501100011033. MG thanks ICREA Academia.

References

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36, 503–518.

Buch-Larsen, T., Guillen, M., Nielsen, J. P. and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39, 503–518.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86, 213–226.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Guillen, M., Nielsen, J. P., Ayuso, M. and Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39, 662–672.

Guillen, M., Nielsen, J. P. and Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, 88, 569–589.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M. and Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, 24, 141–152.

Hall, P., Wolff, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94, 154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution function using dimension reduction. *The Annals of Statistics*, 33, 1404–1421.

Härdle, W. (1990). *Applied Nonparametric Regression*. UK: Cambridge University Press.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21, 157–178.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall/CRC.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91, 1632–1640.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 29, 595–623.

Klein, R. W. and Spady, R. H. (1993). Efficient semiparametric estimator for binary response models. *Econometrica*, 61, 387–421.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Laudagé, C., Desmettre, S. and Wenzel, J. (2019). Severity modeling of extreme insurance claims for tariffication. *Insurance: Mathematics and Economics*, 88, 77–92.

Newey, W. K. and Stoker, T. M. (1993). Efficient of weighted average derivatives estimators and index models. *Econometrica*, 61, 1199–1223.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57, 1403–1430.

Scott, D. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New Jersey: John Wiley & Sons.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54, 1461–1481.

Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Journal of Multivariate Analysis*, 114, 74–98.

van den Berg, G. J., Janys, L., Mammen, E. and Nielsen, J. P. (2021). A general semi-parametric approach to inference with marker-dependent hazard rate models. *Journal of Econometrics*, 221, 43–67.

Information for authors

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process. The article must also use inclusive language, that is, language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as “man”, “mankind”, and masculine pronouns, the use of which might be considered to exclude women. The article should also inform about whether the original data of the research takes gender into account, in order to allow the identification of possible differences.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating’s discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

The articles are licensed under a Creative Commons Atribution-NonCommercial-NoDerivs 3.0 Spain License.

You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work), you may not use the work for commercial purposes and you may not alter, transform, or build upon the work.

Published articles represent the author's opinions; the journal SORT-Statistics and Operations Research Transactions does not necessarily agree with the opinions expressed in the published articles.

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.