Statistical analysis of complex survival data: new contributions in statistical inference, software development and biomedical applications

Gustavo Soutinho
PhD in Applied Mathematics
Faculty of Sciences | University of Porto | Department of Mathematics
2022

Orientador
Prof. Dr. Pedro Oliveira, Institute of Biomedical Sciences Abel Salazar

Coorientador
Prof. Dr. Luís Meira-Machado, University of Minho

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/________

UNIVERSITY OF PORTO

DOCTORAL THESIS

Statistical analysis of complex survival data: new contributions in statistical inference, software development and biomedical applications

Author:
Gustavo SOUTINHO

Supervisor:
Pedro OLIVEIRA

Co-supervisor:
Luís MEIRA-MACHADO

A thesis submitted in fulfilment of the requirements
for the degree of PhD. in Applied Mathematics

at the

Faculty of Sciences
Department of Mathematics

July 1, 2022

Acknowledgements

This thesis is the result of research work developed as a student in the PhD Programme in Applied Mathematics (PDMA). Although being an individual process, during the course of my PhD I had the pleasure to have at my side many people without them this thesis could not be possible.

First of all, I would like to thank my supervisors, Professor Pedro Oliveira and Professor Luís Meira-Machado, for their unconditional support. Each one, in their own way, became a reference for me. In fact, I will never forget their guidance, availability and words of encouragement that enabled me to make my PhD experience so fruitful. In scientific terms, they were rigorous, aware of the more recent contributions in the literature and demonstrated their great knowledge that definitely contributed to the quality of this thesis from which resulted several papers published in indexed journals.

I would also like to express my gratitude to Institute of Public Health of University of Porto (ISPUP) for given me all the condition and support to carry out the research plan of the PhD, and to Foundation for Science and Technology (FCT) for financial support, by assigning an individual doctoral fellowship SFRH/PD/BD/142887/2018. I also thank to all my colleagues and professors who acompanied me at the Faculty of Science to whom I wish the best.

On a personal level, I would like to express my love for my parents and my sister for their care and support in this particular moment in my life. Once again, as always, you are wonderful!

Finally, for all my friends, thank you so much!

To my son Luís

Scientific production

Articles:

Thesis publications:

- Soutinho, G. and Meira-Machado, L. "Nonparametric estimation of the distribution of gap times for recurrent events", Statistical Methods & Applications (2022)
  https://link.springer.com/article/10.1007/s10260-022-00641-6
  DOI:10.1007/s10260-022-00641-6

- Soutinho, G. and Meira-Machado, L. "Parametric Landmark Estimation of the Transition Probabilities in Survival Data with Multiple Events", WSEAS Transactions on Mathematics 21:207-217 (2022)
  https://wseas.com/journals/mathematics/2022/a545106-011(2022).pdf
  DOI:10.37394/23206.2022.21.27

- Soutinho, G. and Meira-Machado, L. "Analysis of Complex Survival Data: a tutorial using the Shiny MSM.app application", arXiv (2022)
  https://arxiv.org/abs/2202.09160
  arXiv:2202.09160v

- Soutinho, G. and Meira-Machado, L. "MSM.App: A Web-Based Tool for the Analysis of Multi-State Survival Data", SSRN Electronic Journal (2021)
  https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3996850
  DOI:10.2139/ssrn.3996850

- Soutinho, G., Sestelo, M. and Meira-Machado, L. "survidm: an R package for Inference and Prediction in an Illness-Death Model", R Journal (2021)
  https://journal.r-project.org/archive/2021/RJ-2021-070/index.html
  DOI: 10.32614/RJ-2021-070

- Soutinho, G. and Meira-Machado, L. "Methods for checking the Markov condition in multi-state survival data", Computational Statistics (2021)
  https://doi.org/10.1007/s00180-021-01139-7
  DOI: 10.1007/s00180-021-01139-7

- Soutinho, G. and Meira-Machado, L. “Estimation of the transition probabilities in multi-state survival data: new developments and practical recommendations”, WSEAS Transactions on Mathematics, 19, 353-366 (2020).
  https://doi.org/10.37394/23206.2020.19.36
  DOI:10.37394/23206.2020.19.36

- Soutinho, G. and Meira-Machado, L. “Some of the most common copulas for simulating complex survival data”, Transactions on Modeling and Computer Simulation, 14, 28-37 (2020).
  https://doi.org/10.46300/9102.2020.14.5
  DOI: 10.46300/9102.2020.14.5

- Soutinho, G., Meira-Machado, L. and Oliveira, P. “A comparison of presmoothing methods in the estimation of transition probabilities”, Communications in Statistics – Simulation and Computation (2020).
  https://doi.org/10.1080/03610918.2020.1762895
  DOI: 10.1080/03610918.2020.1762895

- Soutinho, G. and Meira-Machado, L. “Análise de Dados Multiestado: aplicação a uma base de dados de cancro da mama”, Statistical Portuguese Society (2019).
  https://www.spestatistica.pt/storage/app/uploads/public/
  5ef/5c3/d55/5ef5c3d550134843977065.pdf.

Collaboration in other publication:

- Ferreira J., Carneiro, A., Pedro Cunha, P., Mansilha, A., Vila, I., Cunha, C., Silva, C., Longatto-Filho, A., Correia-Neves, M., Soutinho, G., Meira-Machado, L., Mesquita, A., Cotter, J. “Sarcopenia and Atherosclerotic Occlusive Disease: How Much We Know and What We Need to Know About this Association?”, Artery Research, 26, 86-87 (2020).

Communications:

- “Parametric Landmark estimation of the transition probabilities in survival data with multiple events”, XXV congress of Statistical Portuguese Society (SPE), Évora, October 13-16, 2021 (*Oral Communication*).

- "markovMSM: An R package for checking the Markov condition in multi-state survival data", Open day of Centre Of Mathematics of University of Minho (CMAT), UTAD, October 1, 2021 (*Poster*).

- "Estimation of the transition probabilities conditional on repeated measures in multi-state models", Journeys of Medical Statistical, Bilbao, July 19-23, 2021 (*Poster*).

- "Parametric Landmark estimation of the transition probabilities in survival data with multiple events", 42*nd* Annual Conference of the International Society for Clinical Biostatistics (ISCB 2021), Lyon, July 18-22, 2021 (*Oral communication*).

- "Parametric Landmark estimation of the transition probabilities in survival data with multiple events", 4*th* Annual Meeting CNBIO, University of Vigo, July 1-2, 2021 (*Oral communication*).

- "Estimation of the Transition Probabilities Conditional on Repeated Measures in Multi-state Models", 41*st* Annual Conference of the International Society for Clinical Biostatistics (ISCB 2020), Cracow, August 23-27, 2020 (*Poster*).

- "Estimation of the transition probabilities conditional on repeated measures in multi-state models", Journeys of Medical Statistical, Lisbon, February 12-13, 2020 (*Poster*).

- "Methods for checking the Markovian assumption in Multi-state models", XXIV congress of Statistical Portuguese Society (SPE), Amarante, November 6-9, 2019 (*Oral Communication*).

- "Galician multicenter study of the relationship between anastomotic leak, recurrence and metastasis in rectal cancer", ESCP 14th Scientific and Annual General Meeting, Vienna-Austria, September 25-27, 2019 (*Poster*).

- "Methods for checking the Markovian condition in Multi-state models", 34th International Workshop on Statistical Modelling, July 7-12, 2019, University of Minho (*Poster*).

- "Methods for checking the Markovian assumption in Multi-state models", XVII Spanish Biometric Conference and the VII Ibero-American Biometric Meeting - CEB-EIB 2019, June 18-21, 2019, University of Valencia – Spain (*Oral communication*).

- “A Comparison of Presmoothing Methods in the Estimation of Transition Probabilities”, III Luso-Galician Meeting of Biometry, June 28-30, 2018, University of Aveiro (*Oral communication*).

Invited talks at seminars:

- “Analysis of Multistate Survival Data: a Review of Recent Contributions in Statistical Inference and Software Development”, University of Minho, February 24, 2022
- “Multi-state models”, University of Aveiro, November 27, 2021
- “Inference in multi-state models and application to real data sets using R”, University of Aveiro, April 17, 2020.

Short Duration Courses:

- “Joint Models under the Bayesian approach”, by Dimitris Rizopoulos & Tropical Medicine), 42*nd* Annual Conference of the International Society for Clinical Biostatistics (ISCB 2021), July 18–22, 2021.
- “Statistical modelling with missing data: challenges and practical solutions”, by James Carpenter (London School of Hygiene & Tropical Medicine), 34*th* International Workshop on Statistical Modelling, July 7, 2019.
- “Scientific and Reproducible Programming in R”, University of Valencia by Virgilio Gómez Rubio (Department of Mathematics, University of Castilla-La Macha), June 18, 2019.

UNIVERSITY OF PORTO

Abstract

Faculty of Sciences

Department of Mathematics

PhD. in Applied Mathematics

Statistical analysis of complex survival data: new contributions in statistical inference, software development and biomedical applications

by Gustavo SOUTINHO

Multi-state models are a useful way of describing complex processes in which the individual moves through a number of finite states in continuous time. Since simulation studies play an important role in the evaluation of the performance of a variety of statistical methods, in this dissertation, we present a collection of practical algorithms for simulating multivariate data from a wide class of copulas and survival data in a variety of scenarios of multi-state models.

One other major goal in clinical applications is the estimation of transition probabilities because they allow to long term predictions of the disease progression of a patient. These quantities are usually estimated by the Aalen-Johansen estimator, which assumes the process to be Markovian. The consistency of this estimator is not guaranteed when the process is non-Markovian leading in these cases to biased estimators. To tackle this, we also review the most important nonparametric methods for the estimation of transition probabilities and introduce a new proposal for these quantities in multi-state settings that are not necessarily Markovian, in a form of counting process.

Recently, alternative estimators were introduced in the literature based on subsampling (also known as landmarking) that are consistent regardless the Markov assumption. The computation of their estimators is performed in small sample sizes providing large standard errors in some cases. To avoid this issue, we propose estimators based on presmoothing which are obtained by replacing the censoring indicator variables in the classical definitions by values of regression estimator.

We also introduce feasible estimation methods for the transition probabilities conditionally on covariates observed with repeated measures. To this regard, we use the landmark methodology and existing methods for joint modeling of longitudinal and survival data. This way, we can take into account the effect of the longitudinal marker and not only a single value of the covariates as occurs using the standard Breslow's method.

Once the checking of the Markovian assumption is a relevant issue for the inference in multi-state models, we also present new tests based on measuring the discrepancy of the Aalen-Johansen estimator and recent approaches that do not rely on this assumption, and compared them with the other existing approaches in the literature.

The validity and behavior of the proposed methods were evaluated through simulation studies and illustrated using data sets as examples of application. We have also developed several R packages covering all the methods described in this dissertation, as well as an interactive web application to be used by any user to perform a dynamic analysis independently of their knowledge of informatics.

UNIVERSITY OF PORTO

Resumo

Faculty of Sciences
Department of Mathematics

PhD. in Applied Mathematics

Análise estatística de dados de sobrevivência complexos: novos contributos em inferência estatística, desenvolvimento de software e aplicações biomédicas

por Gustavo SOUTINHO

Os modelos multiestado são uma forma útil de descrever processos complexos nos quais os indivíduos se podem mover entre um número finito de estados ao longo do tempo. Uma vez que os estudos de simulação desempenham um importante papel na avaliação da qualidade de uma variedade de métodos estatísticos, nesta dissertação apresentamos um conjunto de algoritmos para a realização de simulações, a partir de uma vasta classe de cópulas e de dados multivariados e de sobrevivência.

Ao nível das aplicações clínicas, no que se refere a modelos multiestado, a estimação de probabilidades de transição reveste-se da maior importância. uma vez que as mesmas possibilitam previsões a longo prazo da progressão das doenças. Estas quantidades são habitualmente estimadas através do estimador de Aalen-Johansen que assume que o processo é Markoviano. No entanto, a consistência deste estimador não é garantida para casos em que processo é não-Markoviano o que origina um natural enviesamento das estimativas. Nesse sentido, ao longo da tese, procedemos a uma revisão dos mais importantes métodos não-paramétricos para a estimação de probabilidades de transição e introduzimos uma nova proposta de estimação, em modelos multiestado não necessariamente Markovianos, obtida através de processos de contagem.

Recentemente, surgiram na literatura estimadores alternativos para probabilidades de transição baseamos em amostras (também designados por *landmarking*) que são consistentes mesmo em situações em que não se verifica o prossuposto de Markov. Uma vez que em muitas situações as estimativas são obtidas a partir de amostras de pequenas dimensões este facto implica erros-padrão grandes. De forma a evitar esta situação, nesta dissertação propomos estimadores baseados em pressuavização, os quais são obtidos substituindo a variável indicadora nos estimadores clássicos por valores obtidos através de regressão.

Em termos das probabilidades de transição, é, igualmente, proposto um método de estimação condicional a covariáveis representando medidas repetidas. Nesse sentido, é utilizada a abordagem landmark e a adaptação da modelação conjunta de dados longitudinais e de sobrevivência. Deste modo, é possível ter em consideração no processo de estimação o efeito de marcadores longitudinais e não apenas um único valor como ocorre através do habitual método de Breslow.

Considerando a importância do pressuposto de Markov para a inferência em modelos multiestado, nesta dissertação são apresentados novos testes, baseados na quantificação das discrepâncias entre as estimativas obtidas usando o estimador de Aalen-Johansen e as recentes abordagens que não se sustentam na Markovianidade do processo.

A validade e o comportamento dos métodos propostos foram avaliados através de estudos de simulação, bem como exemplificados a partir de bases de dados. Foram, igualmente, elaboradas diversas bibliotecas usando a linguagem R, que abordam os métodos descritos nesta dissertação, assim como uma aplicação web que permite que qualquer utilizador, independentemente dos seus conhecimentos informáticos, realizar uma análise interativa de dados envolvendo a sobrevivência e modelos multiestado.

Keywords

Kaplan-Meier, Multi-state model, Nonparametric estimation, Presmoothing, ROC curves, Survival Analysis.

Contents

List of Figures

List of Tables

Glossary

AFT	Accelerated Failure Time Model
AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
AJ	Aalen-Johansen
AMH	Ali, M.M., Mikhail and Haq
ANOVA	Analysis of Variance
APP	Aplication
AUC	Area Under the Curve
BRES	Breslow
CD4	Cluster of differentiation 4
CIF	Cumulative Incidence Function
COPD	Chronic Obstructive Pulmonary Disease
CPHM	Cox Proportional Hazards Model
CRAN	The Comprehensive R Archive Network
CSS	Cascading Style Sheets
EBMT	European Group for Blood and Marrow Transplantation
EM	Expectation-Maximization
FGM	Farlie-Gumbel-Morgenstern
GAM	Generalized Additive Logistic
GPL-2	General Public License Version 2.0
GT	Global Test

HIV	Human Immunodeficiency Virus
HTML	Hypertext Markup Language
IDM	Illness-death model
IPCW	Inverse Probability of Censoring Weighting
JM	Joint Modeling / Joint Model
JMLM	Joint Modeling Landmark
KM	Kaplan-Meier
KMW	Kaplan-Meier Weights
LIDA	Lifetime Data Analysis
LM	Landmarking / Landmark Model
LMAJ	Landmark Aalen-Johansen
LR	Log-rank
LT	Laplace Transform
LT	Local Test
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
MSM	Multistate model
NP	Nonparametric Presmoothing
PAJ	Presmoothed Aalen-Johansen
PDF	Portable Document Format
PH	Proportional Hazard
PLM or PrLM	Presmoothed Aalen-Johansen
PLMAJ	Presmoothed Landmark Aalen-Johansen
PrKM	Presmoothed Kaplan-Meier
PSA	Prostate Specific Antigen
R	R Language
ROC	Receiver Operating Characteristic

SD	Standard Deviation
TCP	Transmission Control Protocol
TP	Transition Probability
UI	User Interface
WCH	Weighted Cumulative Hazard

Chapter 1

Introduction

1.1 General concepts in survival analysis

Survival analysis can be seen as a set of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. Such events are generally referred to as 'failures' that can be time until an electrical component fails, time to learning a professional skill, or promotion times for employees. In biomedical applications, some examples of events may be time to death or time to first recurrence of a tumor after an initial treatment. Among the wide existing literature concerning survival analysis, the contents of this section are mainly based on the books of Hougaard (2000) [1]; Klein and Moeschberger (1997) [2]; Tableman and Kim (2003) [3]; Kleinbaum and Klein (2012) [4]; and Hosmer, Lemeshow and May (2008) [5].

In a survival analysis, we usually refer to time variable as lifetime or survival time (T) which denote a non-negative random continuous variable that represents the lifetimes of individuals from a homogeneous population. The cumulative distribution function (c.d.f.) for the survival time is given by

$$F(t) = P(T \leq t) = \int_0^t f(x)\, dx \tag{1.1}$$

where $f(\cdot)$ represents the probability density function (p.d.f.).

The probability of an individual survives to time t is given by the survivor function

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)\, dx, \tag{1.2}$$

being $S(t)$ a monotone non-increasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

Conversely, we can express the p.d.f. as

$$f(t) = \lim_{\triangle t \to 0^+} \frac{P(t \leq T < t + \triangle t)}{\triangle t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{1.3}$$

The hazard function ($\lambda(t)$) gives the instantaneous rate of occurrence of the event or failure at $T = t$, given that the individual has survived up to time t. This can be defined as

$$\lambda(t) = \lim_{\triangle t \to 0^+} \frac{P(t \leq T < t + \triangle t \mid T \geq t)}{\triangle t} = \frac{f(t)}{S(t)} = -\frac{dS(t)/dt}{S(t)} = -\frac{d \log \Big(S(t) \Big)}{dt} \tag{1.4}$$

Accordingly, the cumulative hazard or cumulative risk is defined by

$$\Lambda(t) = \int_0^t \lambda(x)\, dx = -\log \Big(S(t) \Big) \tag{1.5}$$

Thus, survival and hazard functions provide an alternative but equivalent characterizations of the distribution of T. From (1.5), survival function can also be given by

$$S(t) = \exp \Big(-\Lambda(t) \Big) = \exp \Big(-\int_0^t \lambda(x)\, dx \Big) \tag{1.6}$$

Taking into account the dynamic nature of survival data, in practice, time to an event cannot be observed due to a deliberate design or random censoring. In particular, right censoring occurs if the event of interest has not been observed when the data was evaluated. Some reasons for that may be a loss of follow-up, drop out or termination of study.

Other types of censoring are left-censoring or interval-censoring. This happens, respectively, when the event of interest has already occured when observation of the individual begins (i.e., the time-to-event is lower than a given value) or that the lifetime is known only to lie within an interval instead of being observed exactly.

The concept of truncation is the major importance in survival analysis. This is a procedure in which another condition beyond the main event of interest is used to select patients. In case of left-truncation, only individuals with lifetime higher than truncation condition are included in the sample. As an example of right-truncation, we can consider the sampling scheme that only infected individuals who have developed AIDS prior to the end of the study are included in the study. In this case the time-to-event is the waiting in years from HIV infection to development of AIDS (Klein and Moeschberger (1997) [2]).

So far, we have dealt with independent and identically distributed survival data but without refering other possible factors that may have influence in the lifetime. In general, this can be done fitting regression models over the hazard function. Among the models, the Cox proportional hazard model (Cox (1972) [6]) is the most frequently used in the literature. This can be defined as follows

$$\lambda(t, x_i) = \lambda_0(t) \exp\Big(\sum_{j=1}^{q} \beta_j x_{ij}\Big) \tag{1.7}$$

where λ_0 is the baseline hazard function, $\beta = (\beta_1, \cdots, \beta_q)'$ is the vector of unknown coefficients and $X = (x_1, \cdots, x_q)'$ is the vector of the covariates.

As we can see, the hazard function is given by the product of two functions. The first one characterizes how the hazard function changes as a function of survival time, while the second changes as a function of subjects covariates. Traditionally, $\lambda_0(t)$ remains unspecified and the coefficients β are obtained through partial likelihood (Cox (1975) [7]) without specifying the baseline hazard function. It is also possible to consider parametric models of the baseline hazard from standard survival distributions such as exponential, weibull or gamma.

Part of the generalization of the Cox model, as a standard in biomedical applications, is due to the easy interpretation of the ratio of two hazard functions. For instance, let us consider a Cox model with only one covariate for two subjects with values x_1 and x_2. In this case, the hazard ratio function is obtained cancelling out $\lambda_0(t)$ as follows

$$HR(t, x_2, x_1) = \frac{\lambda(t, x_2))}{\lambda(t, x_1)} = \frac{\lambda_0(t)\exp(\beta x_2)}{\lambda_0(t)\exp(\beta x_1)} = \exp\Big(\beta(x_2 - x_1)\Big) \tag{1.8}$$

If X is a dichotomous covariate, such gender, with value of $x_2 = 1$ for males and $x_1 = 0$ for females, the hazard ratio in (1.8) becomes $HR(t, x_2, x_1) = \exp(\beta)$. This means that males die at twice the rate of females ($\exp(\beta) = 2$), taking $\beta = \ln(2)$ (Hosmer, Lemeshow and May (2008) [5]). In case of continuous covariates, two assumption are required to use a Cox model: the effect of the covariates should not vary over time (in accordance of the proportional hazard assumption that allows to simplify the expression (1.8)) and the effect of covariates must have a linear functional shape (or log-linear). On the presence of nonlinear effect, this may lead to the risk of a misspecified model with consequences in terms of bias or a decreasing power of tests (Struthers and Kalbfleish (1986) [8]; Anderson and Fleming (1995) [9]).

The lack of flexibility of the Cox model specification has conducted over the last decades to the development of a variety of nonparametric regression methods such as additive hazard regression model (Martinussen and Scheike (2006) [10] and the Cox models with additive predictors (Hastie and Tibshirami (1990) [11]). Several presmoothed approaches have been proposed to render the Cox models more flexible. Among them, penalized splines (Eilers and Marx (1996) [12]) are more commonly used where the nonparametric problem is replaced by a parametric equivalent, in which a vector of regression coefficients is estimated under a smoothness penalty.

1.2 Multi-state models

Multi-state models (Andersen *et al.* (1993) [13], Hougaard (1999) [14], Meira-Machado *et al.* (2009) [15], Meira-Machado and Sestelo (2019) [16]) are models for a time continuous stochastic process, which at any time occupies one of a set of discrete states. These models provide a relevant modeling framework to deal with complex longitudinal survival data in which individuals may experience more than one single event type. In such survival studies, besides overall survival, more than one endpoint can be observed making the use of multi-state models preferable over traditional survival methods (e.g., the Cox model and the Kaplan-Meier estimator of survival). The state structure of a multi-state model identifies the states and the transitions allowed among states. This structure can be represented schematically through diagrams with boxes representing the states and arrows the possible transitions that can occur. The complexity of a multi-state model greatly depends on the number of states defined and on the transitions allowed between these states. The simplest form of a multi-state model is the mortality model which consists of just two states (usually 'alive' and 'dead') and a single transition allowed between them. This corresponds to the usual survival analysis situation. Splitting the 'alive' state from the simple mortality model for survival data into two transient states, we therefore obtain the simplest progressive three-state model. The competing risks model (Andersen and Keiding (2002) [17]; Putter, Fiocco and Geskus (2007) [18]) can be seen as an extension of the simple mortality model for survival data in which each individual may 'die' due to any of several causes.

A well-known and more complex multi-state model is the illness-death model. This model, also known as the disability model, can be used to study the incidence of the disease and the rate of death. Figure 1.1 shows the schematic diagram of transitions involved in the model. In this irreversible version of the model, individuals may pass from the initial state ('health'), to the intermediate event or disease state and then to the absorbing state ('dead'). Individuals are at risk of death in each transient state (States 1 and 2). Many time-to-event data sets from medical studies with multiple end points can be reduced to this generic structure. There exists an extensive literature on multi-state models. Main contributions include books by Andersen *et al.* (1993) [13] and Hougaard (2000) [1]. Recent reviews on this topic may be found in the papers by Putter, Fiocco and Geskus (2007) [18], Meira-Machado *et al.* (2009) [15], and Meira-Machado and Sestelo (2019) [16]. Several other structures of multi-state models and corresponding biomedical examples of application can be found in Hougaard (1999) [14] and Hougaard (2000) [1].

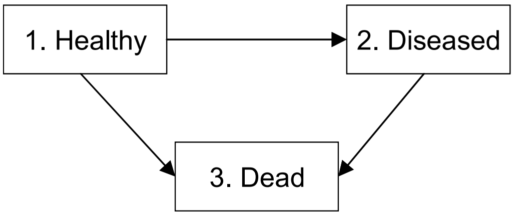


FIGURE 1.1: Illness-death model.

A wide range of biomedical situations have been modeled using multi-state methods, for example, HIV infection and AIDS (Gentleman *et al.* (1994) [19]), liver cirrhosis (Andersen and Esbjerj and Sorensen (2000) [20]), breast cancer (Pérez-Ocón *et al.* (2001) [21]; Putter, Fiocco and Geskus (2007) [18]) and problems following heart transplantation (Meira-Machado *et al.* (2009) [15]). The states are usually based on clinical symptoms (e.g., bleeding episodes), biological markers (CD4 T-lymphocyte cell counts, serum immunoglobulin levels), some scale of the disease (e.g., stages of cancer or HIV infection) or a non-fatal complication in the course of the illness (e.g., heart transplantation, cancer recurrence, etc.). In cancer studies, besides death other endpoints such as locoregional recurrence and distant metastasis are often observed. They have also been used in other areas of application including epidemiology, clinical trials, reliability studies in engineering, etc.

1.3 Main goals in Multi-state models and the Markov condition

The multi-state process can be fully characterized through transition intensities or transition probabilities. The transition intensities are the instantaneous hazards for movement from one state to another. These functions can be used to estimate the mean sojourn time in a given state and to determine the number of individuals in different states at a certain moment. Covariates may be incorporated in the models in order to explain differences among individuals in the course of the illness. In fact, an important goal in multi-state modeling is to study the relationships between the different predictors and the outcome. To this regard, several models have been used in literature. A common simplifying strategy is to decouple the whole process into various survival models by fitting separate intensities to all permitted transitions using semiparametric Cox proportional hazard regression models (Cox (1972) [6]), while making appropriate adjustments to the risk set.

To perform the inference of these quantities is essential to check the Markov assumption, which states that the relevant information for the future evolution of the process is provided by its current state, independently of the states previously visited and the transition times among them. Traditionally, the Markov condition is verified by modeling particular transition intensities on aspects of the history of the process using a proportional hazard model (Kay (1986) [22]). In the progressive illness-death model, the Markov condition is only relevant for the transition from the intermediate state 'disease' (State 2) to 'death' (State 3). Under this model, we can examine whether the time spent in the initial state (State 1) is important on the transition from the disease state to death or not. Therefore, unlike the mortality or the competing risks models, the illness-death model is not necessarily Markovian, since the prognosis for an individual in the intermediate state may be influenced by the subject specific arrival time.

Taking into consideration the Markov condition, the most common models for the inference of transition intensities are characterized through one of the two model assumptions that can be made about the dependence of the transition intensities and time: In first case, they may be modeled using separated Cox models assuming the process to be Markovian (also known as the clock forward modeling approach), which states that past and future are independent given the present state. In second one, they are modeled using a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. Semi-Markov models are also called 'clock reset' models because each time the patient enters a new state the time is reset to 0.

Multi-state models are also very useful to obtain prediction probabilities of future events such as the state occupation probabilities or the transition probabilities. These transition probabilities can be seen as a generalization of the occupation probabilities as far they permit predictions of the clinical prognosis of a patient at a certain point in his/her recovery or illness process. We are also interested to identify the effect of a covariate (or a vector of covariates) for the transition probabilities among states. One standard method, particularly well-suited to the setting with multiple covariates, is to consider estimators based on a Cox's regression model (Cox (1972) [6]) fitted marginally to each transition, with the corresponding baseline hazard function estimated by the Breslow's method (Breslow (1972) [23]). One alternative and flexible nonparametric approach is to consider local smoothing by means of kernel weights based on local constant (Nadaraya-Watson) regression. Right censoring is handled by applying inverse probability of censoring weighting. This is a fully nonparametric approach which provides flexible effects of the continuous covariates (Meira-Machado, de Uña-Álvarez and Datta (2015) [24], Rodríguez-Álvarez, Meira-Machado and Abu-Assi (2016) [25], Meira-Machado and Sestelo (2019) [16]).

1.4 Organization of the thesis

1.4.1 Main objectives

This thesis has five main objectives: (i) to describe a set of algorithms for simulating data from different classes of copulas and provide sampling algorithms to simulate multivariate survival data in a variety of scenarios; (ii) to review the most important nonparametric methods for the estimation of transition probabilities and develop new methods for estimating these quantities in illness-death models that are not necessarily Markovian; (iii) to propose new estimators for transition probabilities that combine landmarking and the joint modeling approach of longitudinal analysis with survival data sets to include repeated measures as covariates; (iv) to introduce new methods for testing the Markov condition in multi-state models involving 'global' and 'local' tests that are based on measuring the discrepancy between the Aalen-Johansen estimator (consistent in Markov processes) and recent approaches that do not rely on this assumption; (v) to develop software in form of R packages to implement the methods addressed in this thesis to be used in biomedical applications.

1.4.2 Thesis outline

In Chapter 2, we introduce the concept of copula, their properties and present different choices of generator for several families of copulas. We also describe dependence measures involving copulas given by Kendall's τ or Spearman's ρ. Due to the importance of simulation studies in statistical inference, we also present algorithms based on three of the most techniques for generating multivariate data from copulas: the conditional distribution method; methods based on the bivariate distribution of the copula or sampling algorithms based on numerical inversion of Laplace transforms. Finally, algorithms for generating survival data in time-to-event, recurrent, competing risk and illness-death models are also described.

In Chapter 3, we revisit different methods for nonparametric estimation of transitions probabilities in multi-state models and report recent contributions to deal with non-Markov settings in which the standard estimator Aalen-Johansen estimator does not provide consistent estimations. To tackle this issue, we propose estimators that are constructed using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. Simulation studies have confirmed the good accuracy of these estimators providing similar behavior when comparing to the landmark estimators.

Chapter 4 is devoted to introduce a new method for improving the accuracy of the transition probabilities estimates under the landmark approach. In fact, these types of estimators provide high variability since they are built by considering subsets of individuals that usually have small size, in particular, at the last moments of the survival studies. To this regard, we present a new estimator based on presmoothing method by replacing the indicator variable of the landmark estimators using logistic or generalized additive logistic models. Nonparametric presmoothing was also considered given by the Nadaraya-Watson kernel estimator. Simulation studies and the application to real data sets have provide good results with less variability of the estimates.

In Chapter 5, we propose new estimators for the estimation of the transition probabilities given a continuous covariate repeatedly measured over time. To this purpose, we combine the joint modeling analysis of longitudinal and survival data with the landmark approach in order to deal with several biomarkers for each individual instead of one single value as occurs using the classical Breslow's methods. Results have confirmed the ability of the proposed methods to reflect the evolution of the longitudinal measures on the transition probabilities among states.

In Chapter 6, we use contributions of the landmark approach for estimating transition probabilities to introduce new methods for testing the Markov condition. To be specific, we consider a 'local' test to check the existing of any moments that enable us to suspect that the process may be non-Markovian. We also developed a 'global' test which is obtained by combining results from 'local' tests over times. The proposed methods have been compared to the traditional global test given by Cox models that includes covariates depending on the history of the process, and the local test given by the log-rank statistics. Results from simulation studies and the application to real data sets have demonstrated the accuracy of the proposed test to detect the lack of Markovianity.

Chapter 7 contains a detailed description of the main functionalities associated to `survidm` R package. This software allows the inference for illness-death models of transition probabilities, occupation probabilities and sojourn distributions, as well as, the coefficients of intensities transitions and a graphical inspection of Markovianity. Other software developments covering the proposed methods in this thesis are also introduced as supplementary material [A].

Chapter 8 introduces the `MSM.app` web application, developed using the shiny packpage, which interactively enables us to show outputs and graphs of multi-state survival data analysis that can be used by everyone, even without any knowledge of the `R` language.

Chapter 2

Some of the most common copulas for simulating complex survival data

Simulation studies play an important role in the evaluation of the performance of a variety of statistical methods. Such assessment is performed under computer intensive procedures and cannot be achieved with studies of real data alone. These studies are increasingly employed in evaluating the properties of the proposed methods being the generation of data the most fundamental and important component. However, only a few of published studies provide sufficient details to allow readers to understand fully all the processes to generate the data. In this chapter, we present a collection of practical algorithms for simulating multivariate data from a wide class of multivariate copulas. This chapter also details important considerations necessary when generating the survival data in a variety of scenarios. A software application for `R` was developed to implement all the methods.

The contents of this chapter are mainly based on the paper published in *International Journal of Mathematics and Computers in Simulation* by Soutinho and Meira-Machado (2020) [26]

2.1 Introduction

Recent advances in computer and software technology have allowed simulation studies to be more accessible. However, performing simulations is not a simple issue. Important guidelines to achieve a good quality simulation study are given by Burton *et al.* (2006) [27]. Data generation is probably the most important step to achieve a good quality simulation study and require a rigorous planning. Unfortunately, only few published articles provide sufficient details to assess the integrity of the study design or to allow readers to understand fully all the processes required when designing their own simulation study. In addition, it is important to obtain simple and high-quality simulations that reflect the complex situations seen in practice, such as, for example, for survival data.

Longitudinal survival data often require the joint modeling of two or more random variables. For example, to model the relationship between survival time of a patient and the hemoglobin level; to model the relationship between two consecutive events of the same nature (recurrent events) or to model different stages in the evolution of an illness (multi-state models). Simulating data for such studies is a challenging issue that can be performed using copulas that provide a useful method for deriving joint distributions given the marginal distributions, especially when the variables are non-normal as in the case of time-to-event variables. In addition, in a bivariate context, copulas can be used to easily control the measures of dependence for the pairs of random variables.

A copula C is a multivariate distribution function that links a univariate marginal distribution to their full multivariate distribution. Copulas were first introduced by Sklar (1959) [28] and its terminology is derived from the Latin word copulare, to connect or to join. In this chapter we explore the topic of random generation in several families of copulas and, in particular, we present algorithms to generate 2-dimensional random vectors (X, Y) whose distribution is $H(x, y) = C(F(x), G(y))$ where F and G denote the marginal distribution functions and C is a copula. These algorithms are based on three of the most used techniques for generating multivariate data from copulas: (1) conditional distribution method; (2) based on the bivariate distribution of the copula and (3) sampling algorithms based on numerical inversion of Laplace transforms. A conceptual framework of these three methods and algorithms for generating survival data is presented in Figure 2.1.

The organization of this chapter is as follows. In Section 2.2, we discusses properties of copulas, their relationships to measures of dependence, and some of the most known families of copulas that have appeared in the literature. Section 2.3 provides practical algorithms for simulating data from a wide class of multivariate copulas. Sampling algorithms are also given to simulate multivariate survival data in a variety of scenarios.

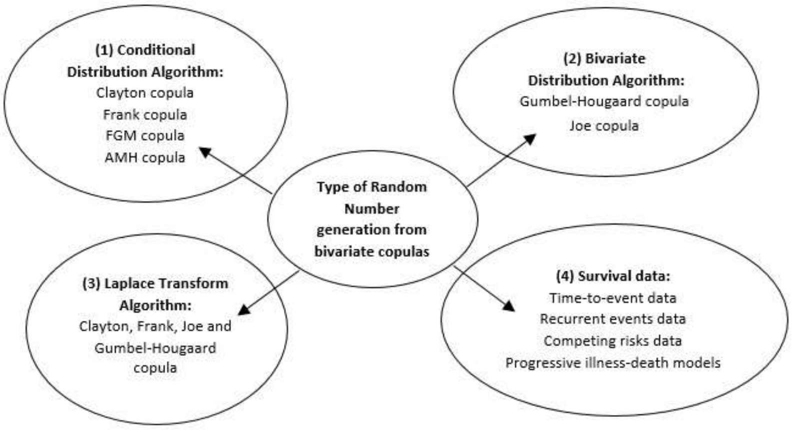

(1) Conditional Distribution Algorithm:
Clayton copula
Frank copula
FGM copula
AMH copula
(2) Bivariate Distribution Algorithm:
Gumbel-Hougaard copula
Joe copula
Type of Random Number generation from bivariate copulas
(3) Laplace Transform Algorithm:
Clayton, Frank, Joe and Gumbel-Hougaard copula
(4) Survival data:
Time-to-event data
Recurrent events data
Competing risks data
Progressive illness-death models

FIGURE 2.1: Copulas and Random Number Generation

Software developments are presented in Section 2.4. Finally, a discussion of the main conclusions are reported in Section 2.5.

2.2 Most common bivariate copulas: Definitions and properties of copulas

Copulas are functions that link multivariate distributions to their one-dimensional margins. These functions are restrictions to $[0,1]^2$ of bivariate distribution functions whose margins are uniform in $[0,1]$. Sklar (1959) [28] showed that if H is a bivariate distribution function with margins $F(x)$ and $G(y)$, then there exists a copula C such that $H(x,y) = C(F(x), G(y))$. Sklar also showed that if the marginal distributions are continuous, then there is a unique copula representation. In the multivariable case, if H is an p-dimensional cumulative distribution function with univariate margins $F_1, ..., F_p$, then there exists an p-dimensional copula C such that $F(x_1, ..., x_p) = C(F_1(x_1), ..., F_p(x_p))$. The case $p = 2$ has attracted special attention and will be considered from now on.

A function $\varphi : [0,1] \to [0,\infty]$ is called a *generator* if it is convex, decreasing and $\varphi(1) = 0$. The generalized inverse of φ (also known as pseudo-inverse) is denoted by $\varphi^{[-1]} = \inf\{u \in [0,1] \mid \varphi(u) \leq t\}, t \in [0,\infty]$.

A copula C is called Archimedean if there exists a generator φ such that $C(u,v) = \varphi^{]-1[}(\varphi(u) + \varphi(v)), (u,v) \in [0,1]^2$. The copula C determines the generator φ uniquely up to a multiplicative constant. In Table 2.1 we present the different choices of generator for several important families of Archimedean copulas.

Family	Space Parameter	Generator $\varphi(t)$	Generator inverse $\varphi^{-1}(s)$	Bivariate copula $C(u,v)$
Clayton (1978) [29]	$\theta \in (0,\infty]$	$\frac{1}{\theta}(t^{-\theta}-1)$	$(1+\theta s)^{-1/\theta}$	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$
Frank (1979) [30]	$\theta \in \mathbb{R} \setminus \{0\}$	$-\ln\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right]$	$-\frac{1}{\theta}\ln(1+e^{-s}(e^{-\theta}-1))$	$-\theta^{-1}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$
Gumbel (1960) [31]	$\theta \in [1,\infty)$	$(-\ln t)^{\theta}$	$e^{-s^{1/\theta}}$	$exp\{-[-(-\ln u)^{\theta}+(-\ln v)^{\theta}]^{1/\theta}\}$
Ali, M.M., Mikhail and Haq (1978) [32]	$\theta \in [-1,1)$	$\ln\left[\frac{1-\theta(1-t)}{t}\right]$	$\frac{1-\theta}{e^{s}-\theta}$	$\frac{uv}{1-\theta(1-u)(1-v)}$
Joe (1997) [33]	$\theta \in [1,\infty)$	$-\ln[1-(1-t)^{\theta}]$	$1-(1-e^{-s})^{1/\theta}$	$1-[(1-u)^{\theta}+(1-v)^{\theta}+(1-u)^{\theta}(1-v)^{\theta}]^{1/\theta}$

TABLE 2.1: Generators and their inverses for one-parameter Archimedean copulas.

Archimedean copulas are popular because they are easily derived and are capable of capturing wide ranges of dependence. Given a pair of variables (X,Y) whose distribution is H, and C the associated copula, this dependence can be measured by Kendall's tau τ or Spearman's ρ. Kendall's tau can be defined as the difference between the probabilities of concordance and discordance for any two independent pairs. In terms of copulas, Kendall's τ is defined by

$$\tau_C = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1. \tag{2.1}$$

The Spearman's ρ coefficient is defined as

$$\rho_C = 12\int_0^1\int_0^1 (C(u,v) - uv)dudv. \tag{2.2}$$

Table 2.2 illustrates the calculation of these correlation measures.

Modeling the multivariate dependence also involves quantifying tail-dependence. Tail-dependence describes the concordance between extreme values of the random variables X and Y. Lower tail-dependence λ_L and the upper tail-dependence λ_U can also be expressed in terms of bivariate copulas

$$\lambda_L = \lim_{u\to 0^+}\frac{C(u,u)}{u} \text{and} \lambda_U = \lim_{u\to 1^-}\frac{1-C(u,u)}{1-u}. \tag{2.3}$$

Family	Kendall's τ	$\tau \in \Omega$	Spearman's ρ
Clayton (1978) [29]	$\frac{\theta}{\theta+2}$	$[0,1)$	No simple form
Frank (1979) [30]	$1-\frac{4}{\theta}\{D_1(-\theta)-1\}$	$[-1,1]\setminus\{0\}$	$1-\frac{12}{\theta}\{D_2(-\theta)-D_1(-\theta)\}$
Gumbel (1960) [31]	$\frac{\theta-1}{\theta}$	$[0,1)$	No simple form
Ali, M.M., Mikhail and Haq (1978) [32]	$1-\frac{2}{3\theta}-\frac{2}{3\theta^2}(\theta-1)^2 ln(1-\theta)$	$[-0.181726, \frac{1}{3}]$	a^*
Joe (1997) [33]	$1+\frac{4}{\theta}E_J(\theta)$	$[0,1)$	No simple form
FGM	$\frac{2}{9}\theta$	$[-\frac{2}{9},\frac{2}{9}]$	$\frac{\theta}{3}$

TABLE 2.2: Copulas and their measures of dependence. $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$ denotes the "Debye" function; $a^* = \frac{12(1+\theta)dilog(1-\theta)-24(1-\theta)\ln(1-\theta)}{\theta^2} - \frac{3(\theta+12)}{\theta}; dilog(x) = \int_1^x \frac{\ln t}{1-t}dt;$ $E_J(\theta) = \int_0^1 \frac{(1-t^{\theta})\ln(1-t^{\theta})}{t^{\theta-1}}dt.$

One of the most popular families of copulas, that were studied by Farlie (1960) [34], Gumbel (1960) [31] and Morgenstern (1956) [35], is the Farlie-Gumbel-Morgenstern (FGM) family that is defined by

$$C(u,v) = uv(1+\theta(1-u)(1-v)), -1 \le \theta \le 1. \tag{2.4}$$

The FGM copula can be seen as a perturbation of the product copula which is obtained for $\theta = 0$. This copula is attractive because of its simplicity but is restrictive since is only useful when dependence between the two marginals is small. A maximum correlation of 33% is attained for the Spearman's coefficient while this correlation is limited to the interval $[-\frac{2}{9}, \frac{2}{9}]$ for Kendall's τ correlation.

To demonstrate the dependence properties of different copulas we simulate 500 pairs of exponential random variables (with rate 1) from the Clayton, Frank, Gumbel, AMH, Joe, and FGM copulas using the approaches outlined in next section. This is illustrated in Figure 2.2.

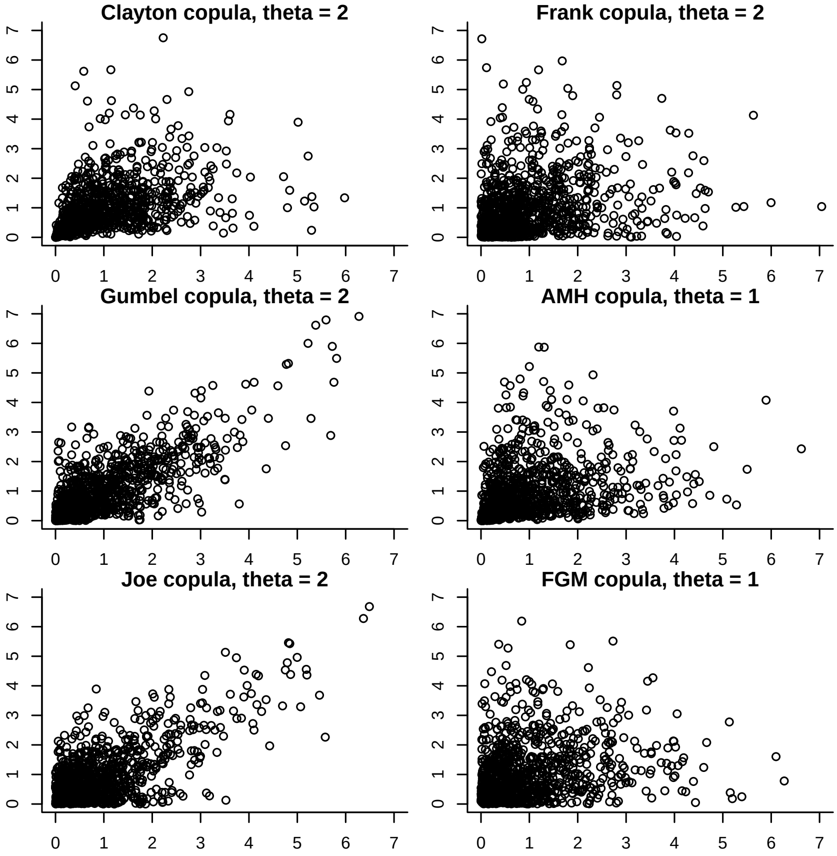


FIGURE 2.2: Simulated samples from copulas (cut at a level of 7).

The pairs of exponential variables are plotted in order to illustrate dependence properties of the copulas. For four of the six copulas, the dependence parameter θ is set to 2. For the remaining copulas the dependence parameter was set to 1. Note that the dependence parameter in FGM, is set such that the dependence between the two variables are maximized (the FGM is unable to accommodate larger dependencies).

2.3 Copulas and random number generation

Simulations have an important role in statistical inference. They are particularly useful to investigate properties of estimators and to study the quality of a model. Moreover, they are also necessary to understand the underlying multivariate distribution. The copula construction allows us to simulate outcomes from many multivariate distributions easily.

The goal of this section is to present practical algorithms to simulate bivariate random variables for all copulas mentioned in the previous section. Assume that (X, Y) is a 2-dimensional random vector whose distribution is

$$H(x, y) = C(F(x), G(y)) \tag{2.5}$$

where F denotes the marginal distribution of X, G the marginal distribution of Y and C is a copula.

2.3.1 Conditional distribution algorithm

One popular algorithm for simulating random variables is based on the conditional distribution approach. This approach separates the copula into several univariate components, each of which can be easily sampled. This method can be used in many copulas (Clayton, Frank, FGM, AMH). Assume that (X, Y) has a bivariate distribution function based on the two-dimensional Archimedean copula (Clayton, Frank, FGM or AMH). To generate data from a bivariate distribution function (X, Y) we first sample (u_1, u_2) from the copula-based distribution $C(u_1, u_2)$ with uniform margins and then we have to invert each u_i using the marginal distributions to obtain the data for the (X, Y). The procedure is to generate the observation of one margin, say U_1, and then to generate an observation for U_2 from its distribution given U_1. Consider two uniform random variables U_1 and U_2 with known copula C. Assuming sufficient regularity conditions, we obtain the conditional cumulative distribution function (c.d.f.)

$$C_{2|1}(u_2 \mid u_1) = P(U_2 \leq u_2 \mid U_1 \leq u_1) = \frac{\partial C(u_1, u_2)}{\partial u_1} \tag{2.6}$$

Thus, the procedure to sample (u_1, u_2) from a copula-based distribution $C(u_1, u_2)$ is based on the algorithm 1 shown below.

Algorithm 1

(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).

(2) Set $u_1 = v_1$.

(3) Find the conditional distribution $C_{2|1}(v_2 \mid v_1)$ and its quasi-inverse $C_{2|1}^{-1}(v_2 \mid v_1)$. Set $u_2 = C_{2|1}^{-1}(v_2 \mid v_1)$.

Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula $C(u_1, u_2)$.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.1: Generating bivariate outcomes from Clayton copula

(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).

(2) Set $u_1 = v_1$.

(3) Set $u_2 = [v_1^{-\theta}(v_2^{-\theta/(1+\theta)} - 1) + 1]^{-1/\theta}$.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.2: Generating bivariate outcomes from Frank's copula

(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).

(2) Set $u_1 = v_1$.

(3) Set $u_2 = -\frac{1}{\theta} \ln \left(1 + \frac{v_2(1-e^{-\theta})}{v_2(e^{-\theta v_1}-1)-e^{-\theta v_1}}\right)$.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.3: Generating bivariate outcomes from FGM copula

(1) Simulate two independent uniform $U(0, 1)$ random variables, say (v_1, v_2).

(2) Set $u_1 = v_1$.

(3) Set $a = 1 + \theta(1 - 2v_1)$; $b = \sqrt{a^2 - 4(a-1)v_2}$.

(4) Set $u_2 = 2v_2/(a + b)$.

(5) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 1.4: Generating bivariate outcomes from AMH copula

(1) Simulate two independent uniform $U(0,1)$ random variables, say (v_1, v_2).

(2) Set $a = 1 - v_1$; $b = 1 - \theta(1 + 2av_2) + 2\theta^2 a^2 v_2$; $c = 1 + \theta(2 - 4a + 4av_2) + \theta^2(1 - 4av_2 + 4a^2 v_2)$.

(3) Set $u_2 = (2t(a\theta - 1)^2)/(b + \sqrt{c})$.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

The conditional distribution algorithm can be extended to the general case of p variables. In higher dimensions, the full distribution of $(X_1, ..., X_p)$ is simulated by recursively simulating the conditional distribution of X_k given $X_1, ..., X_{k-1}$ for $k = 2, ..., p$ (Bouyè *et al.* (2000) [36]).

2.3.2 Bivariate distribution algorithm

For some copulas the conditional distribution is not directly invertible and so different algorithms are necessary. This is the case of the Gumbel-Hougaard copula and the Joe copula. One alternative and popular algorithm that can be used to simulate random variables from an Archimedean copula is based on the following Theorem.

Theorem Let U_1 and U_2 be uniform $U(0,1)$ random variables and let its bivariate distribution function be defined by the Archimedean copula generated by φ. Then, the function $K_C(t) = t - \varphi(t)/(\varphi'(t))$ is the distribution function of $C(U_1, U_2)$. Furthermore, the joint distribution of the random variables $X = \varphi(U_1)/[\varphi(U_1) + \varphi(U_2)]$ and $Y = C(U_1, U_2)$ is characterized by $H(x,y) = x \times K_C(y)$, for all $(x,y) \in I^2$ with X and Y independent, and X uniformly distributed on $(0,1)$. Following, we present a proof in case of copula C to be absolutely continuous.

A proof for the general case can be found in Genest and Rivest (1993) [37].

The joint density $h\left(x,y\right) = \frac{\partial^2}{\partial u_1 \partial u_2} C\left(u_1, u_2\right) \cdot \left| \frac{\partial\left(u_1,u_2\right)}{\partial\left(x,y\right)} \right|$ in terms of x and y, where $\partial^2 C\left(u_1, u_2\right)$ is given as follows and $\partial\left(u_1, u_2\right)/\partial\left(x,y\right)$ correspond to the Jacobian of the transformation $\varphi(u_1) = x\varphi(y)$, $\varphi(u_2) = (1-x)\,\varphi(y)$. Since

$$\frac{\partial\left(u_1, u_2\right)}{\partial\left(x,y\right)} = \frac{\varphi\left(y\right)\varphi'\left(y\right)}{\varphi'\left(u_1\right)\varphi'\left(u_2\right)} \tag{2.7}$$

and consequently

$$h\left(x,y\right)=\left(-\frac{\varphi''\left(y\right)\varphi'\left(u_1\right)\varphi'\left(u_2\right)}{\left[\varphi'\left(y\right)\right]^3}\right)\cdot\left(-\frac{\varphi\left(y\right)\varphi'\left(y\right)}{\varphi'\left(u_1\right)\varphi'\left(u_2\right)}\right)=\frac{\varphi''\left(y\right)\varphi'\left(y\right)}{\left[\varphi'\left(y\right)\right]^2} \tag{2.8}$$

Thus

$$H\left(x,y\right)=\int_0^x\int_0^y\frac{\varphi''\left(z\right)\varphi\left(z\right)}{\left[\varphi'\left(z\right)\right]^2}dzdw=x\cdot\left[z-\frac{\varphi\left(z\right)}{\varphi'\left(z\right)}\right]_0^y=x\cdot K_c\left(y\right) \tag{2.9}$$

and the conclusion follows.

The resulting simulation procedure follows algorithm 2.

Algorithm 2

(1) Simulate two independent uniform $U(0,1)$ random variables, say (v_1, v_2).

(2) Set $t = K_C^{-1}(v_2)$ where $K_C(t) = t - \varphi(t)/(\varphi^{'}(t))$

(3) Set $u_1 = \varphi^{-1}(v_1\varphi(t))$ and $u_2 = \varphi^{-1}((1-v_1)\varphi(t))$

Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula C.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Extensions of the results shown in the above theorem can be used to provide the corresponding simulation algorithm to the multi-dimensional case (Wu, Valdez and Sherris (2006) [38]). The main challenge for the practical implementation of this algorithm is to find the inverse function of K_C.

Algorithm 2.1: Generating bivariate outcomes from the Gumbel-Hougaard copula

(1) Simulate two independent uniform $U(0,1)$ random variables, say (v_1, v_2).

(2) Set $K_C(t) = t(1 - \ln(t)/\theta) = v_2$, and solve numerically for $0 < t < 1$.

(3) Set $u_1 = \exp[v_1^{1/\theta}\ln(t)]$ and $u_2 = \exp[(1-v_1)^{1/\theta}\ln(t)]$.

Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective copula C.

(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

Algorithm 2.2: Generating bivariate outcomes from Joe copula

(1) Simulate two independent uniform $U(0,1)$ random variables, say (v_1, v_2).

(2) Set $K_C(t) = t - \frac{\left[\ln(1-(1-t)^\theta)\right]\left[1-(1-t)^\theta\right]}{\left[\theta(1-t)^{\theta-1}\right]} = v_2$, and solve numerically for $0 < t < 1$.

(3) Set $u_1 = 1 - \{1 - [1-(1-t)^\theta]^{v_1}\}^{1/\theta}$ and $u_2 = 1 - \{1 - [1-(1-t)^\theta]^{1-v_1}\}^{1/\theta}$.

Then, the pairs (u_1, u_2) are uniformly distributed variables drawn from the respective

copula C.
(4) The desired simulated values are $x = F^{-1}(u_1)$ and $y = G^{-1}(u_2)$.

2.3.3 Laplace transform algorithm

Clayton, Frank, Joe and Gumbel-Hougaard copulas fall into the class of the so-called Laplace (Stieltjes) Transform Archimedean copulas (LT-Archimedean copulas). This LT representation leads to a useful way of simulating such copulas (Marshall and Olkin (1988) [39]; Joe (1997) [33]; Hofert (2008) [40]). For such copulas, the inverse of the generator function φ has a nice representation on a Laplace Transform of some function G. Algorithm 3, based on the LT representation, is given below:
Algorithm 3
(1) Generate a variable V with distribution function G with $\psi(t) = \int_0^{+\infty} e^{tx} dG(x), t \geq 0$, the Laplace-Stieltjes transform of G.
(2) Generate independent standard uniform random variables v_1, v_2.
(3) Set $u_i = \psi(-ln(v_i)/V)$.

Then, the vector (u_1, u_2) has the desired Archimedean copula dependence structure with generator $\varphi = \psi^{-1}$.

- For a Clayton copula, V is gamma distributed $Ga(1/\theta, 1)$ and $\psi(t) = (1+t)^{-1/\theta}$.
- For a Gumbel-Hougaard copula, V is stable distributed $St(1/\theta, 1, (cos(\Pi/(2\theta)))^{\theta}, 0; 1)$ (see Nolan (2007) [41]) and $\psi(t) = exp(-t^{1/\theta})$.
- For a Frank copula, V is discrete with $P(V = k) = (1-e^{-\theta})^k/(k\theta)$ and $\psi(t) = -\frac{1}{\theta} ln[1 + e^{-t}(e^{-\theta} - 1)], k \in \mathbb{N}$.
- For the AMH copula, V is discrete with $P(V = k) = (1-\theta)\theta^{k-1}$ and $\psi(t) = \frac{1-\theta}{e^t - \theta}$, $k \in \mathbb{N}$.
- For Joe copula, V is discrete with $P(V = k) = (-1)^{k+1}\binom{1/\theta}{k}$ and $\psi(t) = 1 - (1 - e^{-t})^{1/\theta}, k \in \mathbb{N}$.

Unfortunately, it is not known how to find G explicitly. If we know how to sample G, this algorithm provides a powerful tool for sampling these copulas with large dimensions.

2.3.4 Survival data and random number generation

Copulas have been widely studied in the last decades. Their first applications were mainly in actuarial sciences and finances but their use has spread to other areas such as survival analysis. The copula construction allows the selection of different marginal distributions for each outcome while accounting for the dependence between the random variables. They can be used to model and understand explanatory variables in survival analysis. The copula structure can also be used to study different survival models such as the bivariate survival. For example, suppose we are considering to examine the survival of twins. There is strong empirical evidence that supports the dependence of their lifetimes. Another problem that often appear in survival analysis and that can be modeled with copulas is the issue of competing risks. Though in many cases the outcomes (competing risks; see Figure 2.3) are assumed to be statistically independent there is strong evidence that this assumption is not realistic. To account for this dependence, one general approach is to apply copulas (Escarela and Carrière (2003) [42]; Kaishev, Dimitrova and Haberman (2007) [43]).

In many longitudinal studies subjects can experience several events across a follow-up period. The events of concern may be of the same nature (e.g., cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g., 'alive' and 'disease-free', 'alive with recurrence' and 'dead'). If the events are of the same nature these are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modeled through their intensity functions (Soutinho, Meira-Machado and Oliveira (2020) [44]; Harden and Kropko (2008) [45]; Kropko and Harden (2018) [46]). The dependence between the different outcomes can also be modeled using copulas (Cook and Lawless (2007) [47]; Hougaard (2000) [1]; Malehi *et al.* (2015) [48]; Rotolo, Legrand and Van Keilegom (2013) [49]).

The algorithms shown above can be used to generate survival data that can be used in many of these situations. One can use them to generate survival data subject to random right-censoring (Kalbfleisch and Prentice (1980) [50]; Hougaard (2000) [1]), arising from censored gap times (de Uña-Álvarez and Meira-Machado (2008) [51]; Moreira and Meira-Machado (2012) [52]), competing-risks (Putter, Fiocco and Geskus (2007) [18]) and multi-state models (Andersen *et al.* (1993) [13]; Meira-Machado *et al.* (2009) [15]; Meira-Machado and Sestelo (2016) [53]; Meira-Machado and Sestelo (2019) [16]). Below, we present the algorithms to generate data for four models (Figure 2.3).

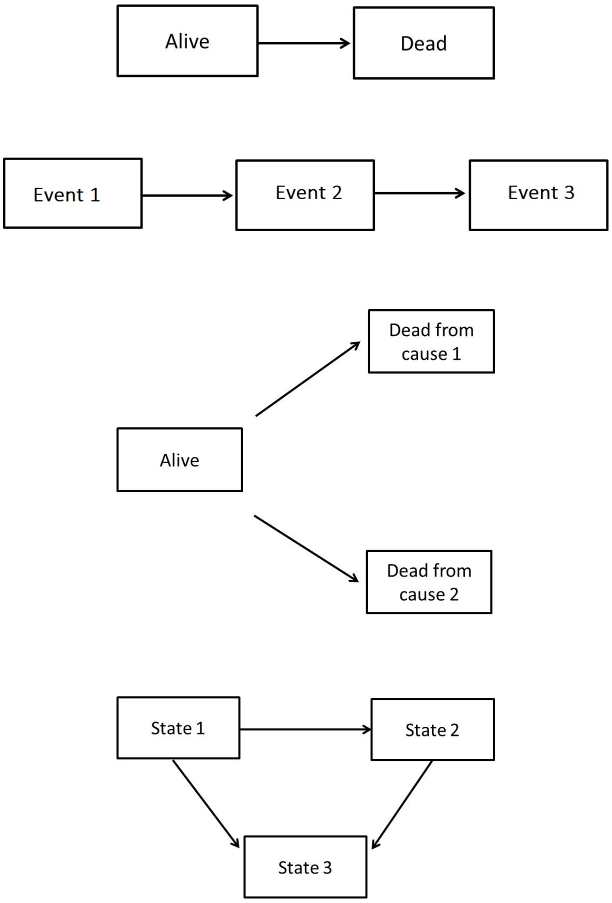


FIGURE 2.3: Schematic representation of some common multi-state models. Mortality model for survival analysis (top); recurrent events model (second row); competing risks model (third row) and progressive illness-death model (bottom).

Time-to-event data

Standard survival data measure the time from some particular time origin until the occurrence of one type of event. The main feature of survival data is censoring. Right-censoring is the most common type of censoring and can occur because of insufficient follow-up, loss to follow-up or failure unrelated to the study. In terms of notation, in this chapter, we denote the random variable survival time by Y. Next, we denote the random censoring variable by Z, which we assume to be independent of Y; and $\Delta = I(Y \leq Z)$ the indicator status indicating either a failure (i.e., $\Delta = 1$) or censorship occurred. Because of censoring rather than Y we observe (T, Δ) where $T = \min(Y, Z)$ is the observed time. If covariables, C, are present, the observed data consists of the triplets (T_i, Δ_i, C_i) $(i = 1, \ldots, n)$ of independent and identically distributed replicates of (T, Δ, C).

The procedure to generate such data is as follows:

(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.

(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(3) Set $T = \min(Y, Z)$ and $\Delta = I(Y \leq Z)$.

Recurrent events data

Recurrent events involve repeat occurrences of the same type of event over time ([47]). Recurrent events in longitudinal studies include recurrent leukaemia episodes, tumor recurrences in cancer patients (e.g. bladder cancer) or heart failure hospitalizations. Let (X, Y) be gap times corresponding to two consecutive events, which are observed subject to random right-censoring. The fact that the variables X and Y are recorded successively, rather than simultaneously, is important when the variables are subject to censoring. Again, we consider here random right censoring (denoted by Z). In the present context of successive events, we only observe the second gap time if the first failure time is uncensored. More precisely, the observable variables are given by $(T_1, T_2, \Delta_1, \Delta_2)$ where $T_1 = \min(X, Z)$, $\Delta_1 = I(X \leq Z)$, $T_2 = \min(Y, Z_2)$ and $\Delta_2 = I(Y \leq Z_2)$, where $Z_2 = (Z - X)I(X \leq Z)$ is the censoring variable for the second gap time.

The procedure to generate such data is as follows:
(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.
(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).
(3) Set $T_1 = \min(X, Z)$; $\Delta_1 = I(X \leq Z)$; $T_2 = \min(Y, Z - X) \times I(X \leq Z)$; $\Delta_2 = I(X + Y \leq Z)$.

Competing risks data

Competing risks data (Figure 2.3, third row) are encountered in many medical studies where the subjects under study are at risk for more than one mutually exclusive event. The observable data in these models is represented by the failure time T and the indicator status variable Δ, which in this case will take the value 0 if the competing risk process does not move from the initial state at the survival time T, or the value 1 and 2 for the possible causes of death 1 and 2. The observable data may also include a possibly covariable vector, which we shall ignore for the moment. The survival time and cause of death may be modeled as arising from the minimum of latent failure times corresponding to the different causes. The procedure to generate such data is as follows:

(1) Generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.

(2) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).

(3) If $X \leq Y$ then $D = 1$; otherwise $D = 2$.

(4) Set $T = \min(X, Y, Z)$; $\Delta = I(\min(X, Y) \leq Z) \times D$.

Alternative simulation designs for competing risks data are given by Beyersmann (2009) [54].

Progressive illness-death multi-state model

In some cases the events of concern may not be of the same nature, representing different stages in the disease process. Consider for example a cancer study, where X represents the time between tumor resection and recurrence (local or distant), and Y represents the time between development of a recurrence and death of the patient. Some individuals may die without observing a recurrence. The progressive illness-death model is probably the most popular one in the medical literature. The irreversible version of this model (Figure 2.3, bottom), describes the pathway from an initial state to an absorbing state either directly or through an intermediate state. Many event-history data sets from biomedical studies with multiple endpoints can be reduced to this generic structure.

To simulate the data in the progressive illness-death model, we separately consider the subjects passing through State 2 at some time, and those who directly go to the absorbing State 3. For the first subgroup of individuals, the successive gap times can be simulated using a two-dimensional copula, whereas those in the second group can be simulated from any continuous distribution.

The procedure to generate such data is as follows:

(1) Draw $\rho \sim Ber(p)$ where p is the proportion of subjects passing through State 2.

(2) If $\rho = 1$ then generate (X, Y) from a bivariate distribution function based on some known two-dimensional copula.

(3) If $\rho = 0$, one particular model (e.g., Uniform, Exponential or Weibull) is used to generate the transition time, W, from State 1 to State 3.

(4) An independent censoring time Z is generated, according to some particular model (e.g., Uniform or Exponential).

(5) If $\rho = 1$ then set $T_1 = \min(X, Z)$ and $\Delta_1 = I(X \leq Z)$. Set also $T = \min(X + Y, Z)$ and $\Delta = I(X + Y \leq Z)$.

(6) If $\rho = 0$ then set $T_1 = \min(W, Z)$ and $\Delta_1 = I(W \leq Z)$. Set also $T = T_1$ and $\Delta = \Delta_1$.

The stochastic behavior of the process in this model is characterized by the vector of random variables $(T_1, T, \Delta_1, \Delta)$, where T_1 is the sojourn in State 1, T the total time and Δ_1 and Δ the corresponding indicator statuses.

The general (and usual) censoring distributions assumed to model censoring are uniform and exponential. The parameters in these distributions can be determined by iterative algorithms to control the censoring percentage one wishes to obtain.

2.4 Software developement

In R, several packages provide functions for simulating survival data. A comprehensive list of these packages can be seen in the CRAN task view 'Survival Analysis' (Allignol and Latouche (2019) [57]). Some of them can be used to simulate data from complex processes, such as the genSurv package (Meira-Machado and Faria (2014) [58]) that permits to generate data with one binary time-dependent covariable and data stemming from a progressive illness-death model. Univariate and semi-competing risks data can be generated using the SimSCRPiecewise package. The survsim package (Crowther and Lambert (2013) [59]; Morina and Navarro (2004) [60]) can also be used to simulate simple and complex survival data such as recurrent event data and competing risks data. Complex multi-state models data with possibly nonlinear baseline hazards and nonlinear covariable effects can be simulated using functions available as part of the simMSM package.

To provide researchers with a tool for simulating complex survival data we develop an R package called survCopula. This package is composed by a set of functions which allow the user to simulate a cohort with the objective of studying its behavior in a variety of scenarios including survival, competing risks, recurrent events and some multi-state models. The main feature of the package is its ability for using different copulas for simulating correlated multivariate survival data in a variety of scenarios as discussed previously. They allow us to control the dependence between time variables with knowledge of the marginal distributions. This software and source code are all available at the GitHub repository at https: at the GitHub repository at https://github.com/gsoutinho/survCopula. Details on the usage of its functions can be obtained with the corresponding help pages after the package is installed. As a supplementary material [A.1], we described all the functions of this R package.

2.5 Discussion

Copulas have become a popular tool to create distributions that model correlated multivariate data. In this chapter a review of the most common copulas is presented with the goal to introduce the generators functions of some important families of Archimedean copulas as well as their dependence that can be measured by Kendall's τ or Spearman's ρ. Due to the important role of the simulation studies in statistical inference, this chapter also describes several algorithms to generate bivariate data from several copulas, and explored the use of these correlated data for generating multivariate survival data in a variety of scenarios. In fact, the use of copulas is suitable for this purpose since they can be used to introduce dependence between time and covariates, or between times of different transitions in more complex survival systems. In case of the Conditional distribution algorithm this can be applied in many copulas such as Clayton, Frank, FGM or AMH. Since some copulas are not directly invertible for the Gumbel-Hougaard and the Joe copulas was also discussed an alternative algorithm making use of the function the function K_C. A Laplace Transform algorithm is also described for some copulas and finally, four types of survival data and random number generation are presented covering different situations, including recurrent events, competing risks and models with multiple events of different types. We also demonstrated the application of these methods of copulas to the biomedical statistics namely in simulation studies involving different models in survival analysis or multistate models who have the advantage to take in consideration the dependence of variables.

In order to be used on biomedical practices a user-friendly software in the form of an R package is provided too. The package provides several functions that can be used to generate survival data in a variety of scenarios including competing risks, recurrent event and multi-state models. Users can choose the marginal distributions as well as the dependence between the correlated data which is induced in the joint distribution by means of copulas. As a future field of research we are interested to use copulas to simulate longitudinal and survival data. This type of data is particularly relevant in cancer studies in which longitudinal biomarkers may be associated to the survival time.

Chapter 3

Estimation of the Transition Probabilities in Multi-state Survival Data: New Developments and Practical Recommendations

Multi-state models can be successfully used for describing complicated event history data, for example, describing stages in the disease progression of a patient. In these models one important goal is the estimation of the transition probabilities since they allow for long term prediction of the process. Traditionally, these quantities have been estimated by the Aalen-Johansen estimator which is consistent if the process is Markovian. Recently, estimators have been proposed that outperform the Aalen-Johansen estimators in non-Markov situations. In this chapter, we review the most important nonparametric methods for the estimation of transition probabilities and consider a new proposal for these quantities in a multi-state system that is not necessarily Markovian. The proposed product-limit nonparametric estimator is defined in the form of a counting process, counting the number of transitions between states and the risk sets for leaving each state with an inverse probability of censoring weighted form. Several simulation studies were conducted under different data scenarios (Section 3.4). The proposed methods were also illustrated with a real data set on colon cancer (Section 3.5). Finally, the advantages and limitations of the different methods and some practical recommendations are discussed in Section 3.6.

The contents of the paper are partially based on the papers published in WSEAS Transactions on Mathematics by Soutinho and Meira-Machado (2020) [55] and *Computational Statistics* by Soutinho and Meira-Machado (2021) [108].

3.1 Notation

A multi-state model is a model for a time continuous stochastic process $(X(t), t \in [0, \infty))$ which at any time occupies one of a few possible states. In this chapter, we consider the progressive illness-death model and assume that all subjects are in State 1 at time $t = 0$, i.e., $P(X(0) = 1) = 1$. In terms of notation, we also may define the sojourn time in the initial state, $Z = \inf\{t : X(t) \neq 1\}$ and the total time of the process $T = \inf\{t : X(t) = 3\}$. Note that (Z, T) falls on a line with a strictly positive probability, since $T = Z$ for those individuals undergoing a direct transition from State 1 to the absorbing State 3. On the other hand, $Z < T$ indicates that the individual visits the intermediate State 2 at some time. In practice, several issues influence the observation of these two random variables. The most common issue is right-censoring, which happens when a subject leaves the study before an event occurs, or when the study ends before the event has occurred. Under right-censoring, only the censored versions of Z and T, along with their corresponding censoring indicators, are available. This censoring is modeled by considering a variable C, which we assume to be independent of the process (Z, T). Define $\widetilde{Z} = \min(Z, C)$ and $\widetilde{T} = \min(T, C)$ for the censored versions of Z and T and introduce $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$ for the respective censoring indicators of Z and T. The variables $\widetilde{Z}$ and $\widetilde{T}_{23} = \widetilde{T} - \widetilde{Z}$ are the observed sojourn times in states 1 and 2, respectively. Finally, the available data are $(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i)$, $1 \leq i \leq n$, i.i.d. copies of $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta)$.

3.2 Transition probabilities

As aforementioned multi-state models may be considered as a generalization of survival analysis where survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. This multi-state process can be fully characterized through transition probabilities between states h and j, that we express by $p_{hj}(s, t|\mathcal{H}_{s-}) = P(X(t) = j|X(s) = h, \mathcal{H}_{s-})$, for $h, j \in \mathcal{S}$ and $s < t$, where $\mathcal{H}_{s-}$ denotes the history of the multi-state process up to s. In particular, the history of the process has the information of the different transitions that occur to an individual over time, as well as the time at which these transitions take place.

The process can be also characterized through the transition intensities

$$\lambda_{hj}(t|\mathcal{H}_{t-}) = \lim_{\Delta t \to 0} \frac{P(X(t+\Delta t) = j|X(t) = h, \mathcal{H}_{t-})}{\Delta t} \tag{3.1}$$

The cumulative transition intensities are defined as $\Lambda_{hj}\,(t) = \int_0^t \lambda_{hj}(u)du$, with $\Lambda_{hh}(t) = -\sum_{j\neq h} \Lambda_{hj}(t)$ the (h,h)th diagonal element of the $K \times K$ matrix $\mathbf{\Lambda}(t)$. Similarly, we define the $K \times K$ matrix $\mathbf{P}(s,t)$ with the (h,j)th element $p_{hj}(s,t)$.

When the multi-state process is Markovian, the transition intensities can be simplified to

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{P(X(t+\Delta t) = j|X(t) = h)}{\Delta t} \tag{3.2}$$

and the transition probabilities to $p_{hj}(s,t) = P(X(t) = j|X(s) = h)$.

In particular, this means that under the Markov assumption, $P(X(t) = j|X(s) = h, X(u) = x) = P(X(t) = j|X(s) = h)$ for any $0 \leq u < s$ and $x \in \mathcal{S}$, and thus, that the future of the process after time s depends only on the state occupied at time s, not on the arrival time to that state or on the states previously visited.

For Markovian processes, the transition probability matrix $\mathbf{P}(s,t)$ can be recovered from the transition intensities through product integration (Aalen and Johansen (1978) [61]):

$$\mathbf{P}(s,t) = \prod_{s<u\leq t} \big(\mathbf{I} + d\mathbf{\Lambda}(u)\big) \tag{3.3}$$

where $\mathbf{I}$ is the $K \times K$ identity matrix, and where the cumulative transition intensities can be estimated by the Nelson-Aalen estimator (Andersen *et al.*(1993) [13])

$$\widehat{\Lambda}_{hj}(t) = \sum_{u\leq t} \frac{N_{hj}(u)}{Y_h(u)} \tag{3.4}$$

where $N_{hj}(t)$ is the number of observed direct transitions from state h to state j up to time t and $Y_h(t)$ is the number of individuals under observation in State h just before time t, and then, the Aalen-Johansen estimator takes the form

$$\widehat{\boldsymbol{P}}(s,t) = \prod_{s<u\leq t} \Big(\mathbf{I} + d\widehat{\boldsymbol{\Lambda}}(u)\Big) \tag{3.5}$$

For more simple models like the illness-death model, we can give explicit expressions for the elements of $\widehat{\boldsymbol{P}}(s,t)$. Expressions for general models are not possible.

Without loss of generality and for the purpose of simplicity, from this point on, we will consider the progressive illness-death model in which we have five different transition probabilities to estimate: $p_{11}(s,t)$, $p_{12}(s,t)$, $p_{13}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$. Using the introduced notation, the transition probabilities can be written as

$$\begin{aligned}
p_{11}(s,t) &= P\left(Z > t \mid Z > s\right), && (3.6)\\
p_{12}(s,t) &= P\left(Z \leq t, T > t \mid Z > s\right), && (3.7)\\
p_{13}(s,t) &= P\left(T \leq t \mid Z > s\right), && (3.8)\\
p_{22}(s,t) &= P\left(Z \leq t, T > t \mid Z \leq s, T > s\right), && (3.9)\\
p_{23}(s,t) &= P\left(T \leq t \mid Z \leq s, T > s\right). && (3.10)
\end{aligned}$$

from which it follows

$$\begin{aligned}
p_{11}(s,t) &= \frac{P\left(Z > t\right)}{P\left(Z > s\right)}, && (3.11)\\
p_{12}(s,t) &= \frac{P\left(s < Z \leq t, T > t\right)}{P(Z > s)}, && (3.12)\\
p_{13}(s,t) &= \frac{P\left(Z > s, T \leq t\right)}{P(Z > s)}, && (3.13)\\
p_{22}(s,t) &= \frac{P\left(Z \leq s, T > t\right)}{P\left(Z \leq s, T > s\right)}, && (3.14)\\
p_{23}(s,t) &= \frac{P\left(Z \leq s, s < T \leq t\right)}{P\left(Z \leq s, T > s\right)}. && (3.15)
\end{aligned}$$

Since we have two obvious relations $p_{12}(s,t) = 1 - p_{11}(s,t) - p_{13}(s,t)$ and $p_{23}(s,t) = 1 - p_{22}(s,t)$ this means that, in practice, we only need to estimate three transition probabilities.

The progressive illness-death model is characterized by three transition intensities: the disease intensity $\lambda_{12}(t)$, the mortality intensity without the disease $\lambda_{13}(t)$ and the mortality intensity among the diseased individuals, $\lambda_{23}(t, t_{12})$. The later transition intensity may depend on t_{12}, the time of the disease occurrence in the illness-death process: $\lambda_{23}(t, t_{12}) = \lim_{\Delta t \to 0} P(X(t + \Delta t) = 3 | X(t) = 2, T_{12} = t_{12})/\Delta t$ where T_{12} represent the potential transition times from State 1 to State 2. The process is called Markov if $\lambda_{23}(t, t_{12})$ is independent of t_{12}, otherwise it is called semi-Markov (i.e., future evolution not only depends on the current state, but also on the entry time into that same state).

In the particular case of the progressive illness-death model the transition probabilities can be obtained from the transition intensities as follows (Beyersmann, Schumacher and Allignol (2011) [62])

$$p_{11}(s, t) = \exp\left(-\int_s^t (\lambda_{12}(u) + \lambda_{13}(u))\, du\right) \tag{3.16}$$

$$p_{22}(s, t \mid t_{12}) = \exp\left(-\int_s^t \lambda_{23}(u, t_{12}) du\right) \tag{3.17}$$

$$p_{12}(s, t) = \int_s^t p_{11}(s, u-)\lambda_{12}(u) p_{22}(u, t \mid u) du \tag{3.18}$$

Here, $p_{22}(s, t \mid t_{12})$ denotes the transition probability p_{22} conditionally on a particular entry time t_{12}. If the process is Markov, $\lambda_{23}(t, t_{12}) = \lambda_{23}(t)$ and $p_{22}(s, t \mid t_{12}) = p_{22}(s, t)$. The two other transition probabilities $p_{13}(s, t)$ and $p_{23}(s, t)$ can be estimated from the two obvious relations aforementioned.

3.3 Nonparametric estimation of the transition probabilities

The standard nonparametric method to estimate a transition probability matrix is the time-honored Aalen-Johansen (AJ) estimator (Aalen and Johansen (1978) [61]). This estimator benefits from the assumption of Markovianity on the underlying stochastic process extending the time-honored Kaplan-Meier estimator (Kaplan and Meier (1958) [63]) to Markov chains.

Moreira, de Uña-Álvarez and Meira-Machado (2013) [64] propose a modification of the Aalen-Johansen estimator in the illness-death model based on a preliminary smoothing (also known as presmoothing, Dikta (1998) [65]; Cao *et al.* (2005) [66]) of the censoring probability for the total time (respectively, of the sojourn time in State 1), given the available information. The presmoothed Aalen-Johansen (PAJ) estimator proposed by Moreira, de Uña-Álvarez and Meira-Machado (2013) [64] is obtained by replacing the censoring indicators (in the transition probabilities $p_{11}(s,t)$ and $p_{22}(s,t)$) by an estimator of a binary (logistic) regression function. The authors verified through simulations that the use of presmoothing can lead to improved estimators with less variability.

Since the Markov assumption may be violated in practice, the consistency of the time-honored Aalen-Johansen estimator and of its presmoothed versions cannot be ensured in general. Exceptions to this are the estimators for $p_{11}(s,t)$ or for the so-called occupation probabilities, $p_{1j}(0,t)$ (Datta and Satten (2001) [67]).

Estimators for the transition probabilities in the progressive illness-death model which do not rely on the Markov assumption were introduced for the first time by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68]. The proposed estimators were defined in terms of multivariate Kaplan-Meier integrals with respect to the marginal distributions of Z and T. These authors showed the practical superiority of their estimators relative to the Aalen-Johansen in situations in which the Markov condition is strongly violated. However, their proposal has the drawback of requiring that the support of the censoring distribution contains the support of the lifetime distribution. Otherwise, they only report valid estimators for truncated transition probabilities. To avoid this issue, corrected estimators (labeled in this chapter as LIDA, the acronym of Lifetime Data Analysis, the journal in which this estimator was published for the first time) were proposed by de Uña-Álvarez and Meira-Machado (2015) [69] for $p_{12}(s,t)$ and $p_{22}(s,t)$.

The paper by de Uña-Álvarez and Meira-Machado (2015) [69] also introduces estimators based on subsampling. The idea behind subsampling, also referred to as landmarking (van Houwelingen (2007) [70]), is to consider the subset of individuals observed in State h by time s. The procedure is then based on (differences between) Kaplan-Meier estimators derived from these subsets of the data. Subsampling was later used by Putter and Spitoni (2018) [71] to derive a landmark Aalen-Johansen estimator (LMAJ) of the transition probabilities. The idea behind the proposed estimator is to use the Aalen-Johansen estimator of the state occupation probabilities derived from those subsets (consisting of subjects occupying a given state at a particular time) for which consistency has already been proved in multi-state models that are not necessarily Markov (Datta and Satten (2001) [67]). In this latter approach, the application of presmoothed estimators (PLMAJ) is possible too.

3.3.1 Aalen-Johansen estimator

The Aalen-Johansen estimator is the standard nonparametric estimator of the transition probabilities for Markov processes. Their estimation method extends the time-honored Kaplan-Meier estimator to Markov chains. The Kaplan-Meier estimator is the standard method to estimate the survival function from time-to-event data that are subject to right censoring. It is a step function with jumps at event times. The size of the steps depends on the number of events and the number of individuals at risk at the corresponding time. Explicit formulae of the Aalen-Johansen estimator (Aalen and Johansen (1978) [61]) for the illness-death model are given by the following expressions:

$$\widehat{p}_{11}^{\texttt{AJ}}(s,t) = \prod_{s<t_i\leq t}\left(1 - dN_1(t_i)/Y_1(t_i)\right) \tag{3.19}$$

$$\widehat{p}_{22}^{\texttt{AJ}}(s,t) = \prod_{s<t_i\leq t}\left(1 - dN_{23}(t_i)/Y_2(t_i)\right) \tag{3.20}$$

and

$$\widehat{p}_{12}^{\texttt{AJ}}(s,t) = \sum_{s<t_i\leq t}\widehat{p}_{11}^{\texttt{AJ}}(s,t_i^-)\frac{dN_{12}(t_i)}{Y_1(t_i)}\widehat{p}_{22}^{\texttt{AJ}}(t_i,t) \tag{3.21}$$

Where $dN_1(t_i) = dN_{12}(t_i) + dN_{13}(t_i)$ for the total number of transitions out of state 1 and let $Y_1(t_i)$ and $Y_2(t_i)$ be the number of healthy (i.e. in state 1) and diseased (i.e. in state 2) individuals, respectively, just prior to time t_i. Since $\widehat{p}_{11}^{\texttt{AJ}}(s,t)$ and $\widehat{p}_{22}^{\texttt{AJ}}(s,t)$ are Kaplan-Meier estimators, their variance may be estimated by Greenwood's formula. The expression for the variance of $\widehat{p}_{12}^{\texttt{AJ}}(s,t)$ can be found in Borgan (2005) [72].

3.3.2 Kaplan-Meier weighted estimators (LIDA)

For a general non-Markov illness-death process without recovery, Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] derived estimators for the transition probabilities defined in terms of multivariate "Kaplan-Meier integrals" with respect to the marginal distribution of the total time T. In particular, the estimators of $p_{12}(s,t)$ and $p_{22}(s,t)$ were proposed as an alternative to the Aalen-Johansen estimators in non-Markov situations. The transition probability $p_{11}(s,t)$ is defined as the ratio of observed survival distributions (and they can be estimated by the ordinary Kaplan-Meier estimator of survival (Kaplan and Meier (1958) [63]) of the sojourn time in State 1, which we denote by $\widehat{S}_0$).

The denominator of $p_{12}(s,t)$ can be estimated in the same way. The remaining quantities involve expectations of particular transformations of the pair (Z,T), $E\left[\varphi\left(Z,T\right)\right]$, which can not be estimated so simply

$$\widehat{p}_{12}^{\texttt{LIDA}}(s,t) = \frac{\widehat{E}(\varphi_{s,t}(Z,T))}{\widehat{S}_0(s)} \tag{3.22}$$

and

$$\widehat{p}_{22}^{\texttt{LIDA}}(s,t) = \frac{\widehat{E}(\widetilde{\varphi}_{s,t}(Z,T))}{\widehat{E}(\widetilde{\varphi}_{s,s}(Z,T))} \tag{3.23}$$

where $\varphi_{s,t}(u,v) = I(s < u \leq t, v > t)$ and $\widetilde{\varphi}_{s,t}(u,v) = I(u \leq s, v > t)$ and $\widehat{E}(\varphi_{s,t}(Z,T))$ is the "Kaplan-Meier integral"

$$\widehat{E}(\varphi_{s,t}(Z,T)) = \sum_i W_i \varphi_{s,t}(\widetilde{Z}_i, \widetilde{T}_i) \tag{3.24}$$

where W_i is the Kaplan-Meier weight attached to $\widetilde{T}_i$ when estimating the marginal distribution of T from the $\left(\widetilde{T}_i, \Delta_i\right)$'s (equal to minus the jump at $\widetilde{T}_i$ of the Kaplan-Meier estimator of survival of the total time $\widehat{S}$). See Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] for more details.

The methods proposed by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] have the drawback of requiring that the support of the censoring distribution contains the support of the lifetime distribution. An assumption that is often not fulfilled in medical applications due to limitations in the patient's following-up. To avoid this potential problem, corrected estimators were proposed by de Uña-Álvarez and Meira-Machado (2015) [69] for $p_{12}(s,t)$ and $p_{22}(s,t)$:

$$\widehat{p}_{12}^{\texttt{cLIDA}}(s,t) = \frac{\widehat{S}_0(s) - \widehat{S}_0(t) - \widehat{E}(\gamma_{s,t}(Z,T))}{\widehat{S}_0(s)} \tag{3.25}$$

and

$$\widehat{p}_{22}^{\texttt{cLIDA}}(s,t) = 1 - \frac{\widehat{E}(\widetilde{\gamma}_{s,t}(Z,T))}{\widehat{S}(t) - \widehat{S}_0(s)} \tag{3.26}$$

where $\gamma_{s,t}(u,v) = I(u > s, v \leq t)$ and $\widetilde{\gamma}_{s,t}(u,v) = I(u \leq s, s < v \leq t)$.

All these quantities can be estimated nonparametrically using Kaplan-Meier weights.

3.3.3 Landmark Estimators

Recently, de Uña-Álvarez and Meira-Machado (2015) [69] have used the idea of subsampling to introduce (landmark) estimators of the transition probabilities which do not rely on the Markov property. The idea of the new methods is to use a procedure based on (differences between) Kaplan-Meier estimators derived from a subset of the data consisting of all subjects observed to be in a given state at a given time. Following the notation introduced in Section 3.1, given the time point s, to estimate $p_{1j}(s,t)$ for $j = 1,2,3$ the landmark analysis is restricted to the individuals observed in State 1 at time s. For the subpopulation $Z > s$, the censoring time C is still independent of the pair (Z,T) and, therefore, Kaplan-Meier-based estimation will be consistent. Similarly, to estimate $p_{2j}(s,t)$, $j = 2,3$, the landmark analysis proceeds from the sample restricted to the individuals observed in State 2 at time s. Then, we may formally introduce the landmark estimators as follow

$$\widehat{p}_{11}^{\,\mathtt{LM}}(s,t) = \widehat{S}_0^{\,\mathtt{KM}(s)}(t) \tag{3.27}$$

$$\widehat{p}_{12}^{\,\mathtt{LM}}(s,t) = \widehat{S}^{\,\mathtt{KM}(s)}(t) - \widehat{S}_0^{\,\mathtt{KM}(s)}(t) \tag{3.28}$$

$$\widehat{p}_{13}^{\,\mathtt{LM}}(s,t) = 1 - \widehat{S}^{\,\mathtt{KM}(s)}(t) \tag{3.29}$$

$$\widehat{p}_{22}^{\,\mathtt{LM}}(s,t) = \widehat{S}^{\mathtt{KM}[s]}(t) \tag{3.30}$$

$$\widehat{p}_{23}^{\,\mathtt{LM}}(s,t) = 1 - \widehat{S}^{\,\mathtt{KM}[s]}(t) \tag{3.31}$$

where $\widehat{S}_0^{\mathtt{KM}(s)}$ and $\widehat{S}^{\mathtt{KM}(s)}$ are the Kaplan-Meier estimators for the distributions of Z and T, respectively, but computed from the subsample $\mathcal{S}_1 = \left\{i : \widetilde{Z}_i > s\right\}$; whereas $\widehat{S}^{\,\mathtt{KM}[s]}$ is the Kaplan-Meier estimator of the distribution of T but computed from the subsample $\mathcal{S}_2 = \left\{i : \widetilde{Z}_i \leq s < \widetilde{T}_i\right\}$.

As we can see, the Kaplan-Meier estimator plays an important role for the landmark estimators proposed by de Uña-Álvarez and Meira-Machado (2015) [69]. Under our notation, the Kaplan-Meier product-limit estimator of the survival function $S(t) = P(T > t)$ can be expressed using Kaplan-Meier weights as follows

$$\widehat{S}^{\mathtt{KM}}(t) = 1 - \sum_{i=1}^{n} W_i I(\widetilde{T}_{(i)} \leq t) \tag{3.32}$$

where

$$W_i = \frac{\Delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[j]}}{n-j+1} \right] \tag{3.33}$$

is the Kaplan-Meier weight attached to $\widetilde{T}_{(i)}$.

Similarly, one could introduce the Kaplan-Meier formula based on the $(\widetilde{Z}_i, \Delta_{1i})$'s for the distribution of Z.

The subsampling approach combined with the Aalen-Johansen estimate of the state occupation probabilities was later used by Putter and Spitoni (2018) [71] to introduce the termed Landmark Aalen-Johansen estimator. The landmark Aalen-Johansen estimators of the transition probabilities may then be introduced as

$$p_{hj}^{\texttt{LMAJ}}(s,t) = \widehat{\pi}^{\texttt{LM}}(s) \prod_{s<u\leq t} \left(\mathbf{I} + d\widehat{\mathbf{\Lambda}}^{\texttt{LM}}(u) \right) \tag{3.34}$$

with $\widehat{\pi}^{\texttt{LM}}(s)$ a $1 \times K$ vector with $\widehat{\pi}^{\texttt{LM}}(s) = 1$ for the jth element, and other values equal to 0. Here, the estimator of the cumulative transition intensities, $\widehat{\Lambda}^{\texttt{LM}}$, is Nelson-Aalen estimator computed on a landmark data set which selects subjects observed to be in State h at time s (Putter and Spitoni 2018) [71].

Simulation studies published in the paper by Putter and Spitoni (2018) [71] show that the landmark Aalen-Johansen estimator (`LMAJ`) and the landmark estimator (`LM`) perform similarly. In fact, the two landmark estimators (`LM` and `LMAJ`) of the transition probabilities $p_{11}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$ are equivalent.

As a weakness, the landmark estimators proposed by de Uña-Álvarez and Meira-Machado (2015) [69] and Putter and Spitoni (2018) [71] may provide large standard errors in estimation in some circumstances. This may occur for small sample sizes and/or large proportion of censored data. In such cases the estimators based on a landmark approach may result in a wiggly estimator with fewer jump points. A valid approach that can be used to reduce the variability of these estimators is to consider a modification of the landmark estimator based on presmoothing (Meira-Machado (2016) [73]; Soutinho, Meira-Machado and Oliveira (2020) [44]).

3.3.4 Weighted Cumulative Hazard Estimators

In this section we propose new estimators for the transition probabilities $p_{11}(s,t)$, $p_{13}(s,t)$ and $p_{22}(s,t)$. The estimators are constructed using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. The proposed estimator (`WCH` - weighted cumulative hazard) for the transition probability $p_{11}(s,t)$ is given by

$$\begin{aligned} \widehat{p}_{11}^{\,\texttt{WCH}}(s,t) &= \widehat{P}\left(Z > t \mid Z > s\right), \\ &= \prod_{v \in R_1} \{1 - \widehat{\Lambda}_{11}(dv)\}, \end{aligned} \tag{3.35}$$

where $\Lambda_{11}(dv)$ is the cumulative conditional hazard of Z given $Z > s$. Assuming that $Z \perp C$, $\Lambda_{11}(dv)$ can be estimated by

$$\widehat{\Lambda}_{11}(dv) = \frac{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{Z}_i = v, \Delta_{1i} = 1)}{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{Z}_i \geq v)}$$

and where $R_1 = \{\widetilde{Z}_i : \widetilde{Z}_i \leq t\}$.

Estimator (3.35) is equivalent to the estimator proposed by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68], de Uña-Álvarez and Meira-Machado (2015) [69] and the so-called Aalen-Johansen estimator (Aalen and Johansen (1978) [61].

Similar ideas can be used to obtain estimators for $p_{13}(s,t)$ and $p_{22}(s,t)$. Note that $p_{13}(s,t) = P(T \leq t \mid Z > s) = 1 - P(T > t \mid Z > s)$. Then,

$$\begin{aligned} \widehat{p}_{13}^{\,\texttt{WCH}}(s,t) &= 1 - \widehat{P}\left(T > t \mid Z > s\right), \\ &= 1 - \prod_{v \in R_{13}} \{1 - \widehat{\Lambda}_{13}(dv)\} \prod_{v \in R_{123}} \{1 - \widehat{\Lambda}_{123}(dv)\}, \end{aligned} \tag{3.36}$$

where $\Lambda_{13}(dv)$ is the cumulative conditional hazard of T given $Z > s$ for those individuals going directly into State 3 without visiting State 2; and $\Lambda_{123}(dv)$ is the cumulative conditional hazard of T given $Z > s$ for those visiting State 2. Assuming that $(Z,T) \perp C$, $\Lambda_{13}(dv)$ can be estimated by

$$\widehat{\Lambda}_{13}(dv) = \frac{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{Z}_i = \widetilde{T}_i, \widetilde{T}_i = v, \Delta_{2i} = 1)}{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{T}_i \geq v)} \tag{3.37}$$

and where $R_{13} = \{\widetilde{T}_i : \widetilde{T}_i \leq t\}$; whereas $\Lambda_{123}(dv)$ can be estimated by

$$\widehat{\Lambda}_{123}(dv) = \frac{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{Z}_i < \widetilde{T}_i, \widetilde{T}_i = v, \Delta_{2i} = 1)}{\sum_{i=1}^{n} I(\widetilde{Z}_i > s, \widetilde{T}_i \geq v)} \tag{3.38}$$

and where $R_{123} = \{\widetilde{T}_i : \widetilde{T}_i \leq t\}$.

Then, $p_{12}(s,t)$ can be estimated by $\widehat{p}_{12}^{\,\texttt{WCH}}(s,t) = 1 - \widehat{p}_{11}^{\,\texttt{WCH}}(s,t) - \widehat{p}_{13}^{\,\texttt{WCH}}(s,t)$. Note that estimators of $\widehat{p}_{1j}^{\,\texttt{WCH}}(s,t)$, $j = 1,2$ are equivalent to the landmark estimators proposed by de Uña-Álvarez and Meira-Machado (2015) [69].

Since $p_{22}(s,t) = \frac{P(Z<s,T>t)}{P(Z<s,T>s)} = \frac{P(T>t|Z<s)}{P(T>s|Z<s)}$. The key to estimating $p_{22}(s,t)$ is to estimate $P(T > u \mid Z < s)$ for $u \in \{s,t\}$. These quantities can be estimated by

$$\widetilde{P}(T > u \mid Z < s) \quad = \quad \prod_{v \in R_{23}} \{1 - \widetilde{\Lambda}_{23}(dv)\}, \tag{3.39}$$

where $R_{23} = \{\widetilde{T}_{23i} : \widetilde{T}_{23i} \leq u - \widetilde{Z}_i, \widetilde{Z}_i < \widetilde{T}_i\}$; and $\Lambda_{23}(dv)$ can be estimated by

$$\widetilde{\Lambda}_{23}(\Delta v) = \frac{\sum_{i=1}^{n} I(\widetilde{Z}_i \leq s, \widetilde{Z}_i < \widetilde{T}_i, \widetilde{T}_{23i} = v, \Delta_{2i} = 1)/\widehat{G}(\widetilde{Z}_i + v)}{\sum_{i=1}^{n} I(\widetilde{Z}_i \leq s, \widetilde{Z}_i < \widetilde{T}_i, \widetilde{T}_{23i} \geq v, \Delta_{1i} = 1)/\widehat{G}(\widetilde{Z}_i + v)}. \tag{3.40}$$

The resultant estimator is labeled as $\widehat{p}_{22}^{\texttt{WCH}}(s,t)$.

Since $p_{22}(s,t) = P(T > t|Z < s, T > s)$, an alternative estimator is given by

$$\widetilde{p}_{22}^{\texttt{WCH}}(s,t) = \widetilde{P}(T > u \mid Z < s, T > s) \quad = \quad \prod_{v \in R_{23}^{\star}} \{1 - \widetilde{\Lambda}_{23}^{\star}(dv)\}, \tag{3.41}$$

where $R_{23}^{\star} = \{\widetilde{T}_i : \widetilde{T}_i \leq t\}$ and where $\Lambda_{23}^{\star}(dv)$ can be estimated by

$$\widetilde{\Lambda}_{23}^{\star}(\Delta v) = \frac{\sum_{i=1}^{n} I(\widetilde{Z}_i \leq s, \widetilde{T}_i > s, \widetilde{T}_i = v, \Delta_{2i} = 1)/\widehat{G}(v)}{\sum_{i=1}^{n} I(\widetilde{Z}_i \leq s, \widetilde{T}_i > s, \widetilde{T}_i \geq v, \Delta_{1i} = 1)/\widehat{G}(\max(\widetilde{Z}_i, v))}. \tag{3.42}$$

The estimator $\widetilde{p}_{22}^{\texttt{WCH}}(s,t)$ is equivalent to the landmark estimator $\widehat{p}_{22}^{\texttt{LM}}(s,t)$ proposed by de Uña-Álvarez and Meira-Machado (2015) [69].

The estimation of the variance is important for inference purposes. Resampling techniques such as bootstrap provide here a practical solution to the problem of variance estimation and inference. These methods can be used to construct confidence limits based on the percentile bootstrap.

3.4 Simulation study

In this section we investigate the performance of the proposed estimators through simulations. More specifically, the estimators introduced in Section 3.3 are considered. In particular we aim to compare the performance of the Aalen-Johansen estimator which benefits from the assumption of Markovianity on the underlying stochastic process, with alternative estimators which are free of the Markov condition. The simulation addresses also the question about the more efficient estimator in different scenarios.

To simulate the data in the irreversible illness-death model, we separately consider the subjects passing through State 2 at some time, and those who directly go to the absorbing State 3. For the second subgroup of individuals, times to death without illness are generated from the hazard function $h_{13}(t) = 0.024t$. For the first subgroup of individuals, the successive gap times $(Z, T - Z)$ are simulated using two cause-specific hazard functions, λ_{12} and λ_{23} for each of the events (illness and death). The cause-specific hazard for the intermediate event was defined as $\lambda_{12}(t) = \frac{0.29}{t+1}$. For the individuals that experienced the disease, times to death after the disease were generated using three different hazards:

$$\begin{aligned}\lambda^1_{23}(t, t_{12}) &= 0.05\\ \lambda^2_{23}(t, t_{12}) &= \frac{1}{0.25(t_{12}+1)^{0.8}}\\ \lambda^3_{23}(t, t_{12}) &= 0.04 \times log(t+1)\end{aligned}$$

where $t > 0$, denotes the time since the start point, and t_{12} is the transition time from State 1 to State 2.

The use of these three different hazard functions provides three different scenarios. The first scenario can be considered Markovian since the hazard of death after the disease was set constant being independent of t and t_{12}. In the two remaining scenarios, the hazard for death after disease depended on the these times. The second scenario is semi-Markovian since the process depends not only of the current state, but also how long it has been in the current state (time refers to time since entering the intermediate state). The third scenario is non-Markov since the hazard for death after disease depends on the time since entry in study.

An independent uniform censoring time C is generated, according to models $U[0,20]$ and $U[0,30]$. For the Markovian scenario, the first model, presents 19% of censoring on the first gap time Z, 40% censoring on the total time T and 42% on the second gap time $T-Z$, for those individuals who entered in state 2. The second model changes these censoring levels to 33%, 58% and about 36%, respectively. The first model in the semi-Markovian scenario, reveals 23% of censoring for Z, 40% for T and 41% for $T-Z$. Second model increases these censoring levels to 36%, 49% and 41%, respectively. Finally, in the non-Markovian scenario, the model $U[0,20]$ provides 25% of censoring on the first gap time, 43% on the total time and 41% on the second gap time. Under the model $U[0,30]$ censoring increases to 26%, 44% and about 42%, respectively.

For each simulated scenario we consider several different points (s,t) pairs, corresponding to combinations of times 2, 4, 8 and 12 representing the differences between closer and distant times. Sample sizes $n = 100$ and $n = 250$ are considered. In each simulation, 1000 samples are generated. From these samples we obtained the mean for all generated data sets. As a measure of efficiency, we took the Mean Squared Error (MSE) but we also computed the standard deviations (SD) and the Bias.

Tables 3.1 (Markov scenario), 3.2 (semi-Markov scenario) and 3.3 (non-Markov scenario) report the results for transition probability $p_{13}(s,t)$. When one is confident of the Markov assumption, the Aalen-Johansen is preferred over non-Markovian estimators since it reports a smaller variance in estimation. This is in agreeing with results reported in Table 3.1. Results reported in the Tables 3.2 and 3.3 also reveal that the Aalen-Johansen estimator (labeled as `AJ`) might still perform reasonably well in situations where the process shows only mild deviations from Markovianity. However, when there is strong evidence that the process is not Markov the use of a non-Markov estimator is preferable due to their greater accuracy. This can be observed from results reported in Tables 3.2 and 3.3.

All three simulation scenarios reveal that the performance of the methods is poorer at the right tail. This was expected because for larger values of s and t, the censoring effects are stronger. The SD decreases with an increase of the sample size and with the decrease of the censoring percentage, which was also expected.

Results in Tables 3.1, 3.2 and 3.3 reveal a poor performance of the original non-Markov estimators by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68], referred to as `LIDA` estimators.

The other two non-Markovian methods (the corrected LIDA, labeled as cLIDA, and the Weighted Cumulative Hazard method, WCH) obtain in all settings a negligible bias (decreasing as the sample size increases), while the LIDA estimator shows a systematic bias.

Tables 3.2 and 3.3 show that the Markov-free estimators cLIDA and WCH may behave much more efficiently than the Aalen-Johansen. This is because of the failure of the Markov assumption from which the Aalen-Johansen estimator is built. This is more evident in the semi-Markov scenario, with higher lag times $t - s$. In these cases, the Aalen-Johansen show a systematic bias which does not decrease with an increasing sample size. In these cases the application of the Aalen-Johansen method is not recommended here, due to possible biases. The poor behavior of the Aalen-Johansen estimator can also be seen in Figure 3.1, in which we show the boxplots of the estimates of the transition probabilities based on the 1000 Monte Carlo replicates for the four estimators, with different sample sizes. From these plots it can be seen that the cLIDA and WCH methods are unbiased estimators and confirm the less variability of the Aalen-Johansen estimator. The WCH method (which in this case is equivalent to the LM method) is the preferred since is the unbiased method reporting less variability.

TABLE 3.1: Bias and standard deviation (SD) for the three estimators of $p_{13}(s,t)$. Markov scenario with two sample sizes and two censoring levels.

		$\widehat{p}_{13}^{\texttt{AJ}}(s,t)$		$\widehat{p}_{13}^{\texttt{LIDA}}(s,t)$		$\widehat{p}_{13}^{\texttt{cLIDA}}(s,t)$		$\widehat{p}_{13}^{\texttt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)								
n=100	$C \sim U[0,30]$	<0.0001	00299	0.0345	0.0533	<0.0001	0.0316	<0.0001	0.0316
	$C \sim U[0,20]$	<0.0001	0.0300	0.0600	0.0552	<0.0001	0.0312	<0.0001	0.0311
n=250	$C \sim U[0,30]$	<0.0001	0.0187	0.0335	0.0368	<0.0001	0.0197	<0.0001	0.0197
	$C \sim U[0,20]$	<0.0001	0.0193	0.0584	0.0383	<0.0001	0.0204	<0.0001	0.0204
(s,t)=	(2,8)								
n=100	$C \sim U[0,30]$	0.0024	0.0527	0.0659	0.0926	0.0021	0.0567	0.0025	0.0563
	$C \sim U[0,20]$	0.0014	0.0594	0.118	0.0907	0.0010	0.0630	<0.0001	0.0627
n=250	$C \sim U[0,30]$	<0.0001	0.0340	0.0677	0.0626	<0.0001	0.0363	<0.0001	0.036
	$C \sim U[0,20]$	<0.0001	0.0375	0.1147	0.0607	<0.0001	0.0406	<0.0001	0.0401
(s,t)=	(4,12)								
n=100	$C \sim U[0,30]$	-0.0039	0.0736	0.0669	0.1034	-0.0029	0.0791	-0.0038	0.0767
	$C \sim U[0,20]$	-0.0024	0.0824	0.1084	0.1181	-0.0013	0.0911	-0.0014	0.0893
n=250	$C \sim U[0,30]$	0.0011	0.0462	0.0671	0.0733	0.0018	0.0508	0.0017	0.0496
	$C \sim U[0,20]$	<0.0001	0.0503	0.1132	0.0708	<0.0001	0.0541	<0.0001	0.0533
(s,t)=	(8,12)								
n=100	$C \sim U[0,30]$	-0.0015	0.0816	0.0424	0.103	-0.0031	0.0844	-0.0029	0.0839
	$C \sim U[0,20]$	-0.0019	0.099	0.0645	0.117	-0.0016	0.1034	-0.0027	0.1012
n=250	$C \sim U[0,30]$	<0.0001	0.0529	0.0412	0.0658	<0.0001	0.0551	<0.0001	0.0546
	$C \sim U[0,20]$	<0.0001	0.0621	0.0615	0.0785	<0.0001	0.0652	<0.0001.	0.0639

TABLE 3.2: Bias and standard deviation (SD) for the three estimators of $p_{13}(s,t)$. Semi-Markov scenario with two sample sizes and two censoring levels.

		$\widehat{p}_{13}^{\mathtt{AJ}}(s,t)$		$\widehat{p}_{13}^{\mathtt{LIDA}}(s,t)$		$\widehat{p}_{13}^{\mathtt{cLIDA}}(s,t)$		$\widehat{p}_{13}^{\mathtt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)								
n=100	$C \sim U[0,30]$	0.0085	0.0312	0.0181	0.0488	0.0015	0.0323	0.0016	0.0323
	$C \sim U[0,20]$	0.0057	0.0311	0.0333	0.0534	<0.0001	0.0331	<0.0001	0.0331
n=250	$C \sim U[0,30]$	0.0071	0.0197	0.0158	0.0337	<0.0001	0.0202	<0.0001	0.0202
	$C \sim U[0,20]$	0.0065	0.0201	0.0333	0.0378	<0.0001	0.0208	<0.0001	0.0208
(s,t)=	(2,8)								
n=100	$C \sim U[0,30]$	0.0273	0.0579	0.0498	0.087	0.0014	0.0617	0.0014	0.0610
	$C \sim U[0,20]$	0.0254	0.0596	0.0863	0.0932	<0.0001	0.0639	<0.0001	0.0634
n=250	$C \sim U[0,30]$	0.0276	0.0353	0.0474	0.0590	0.0019	0.0369	0.0022	0.0370
	$C \sim U[0,20]$	0.0257	0.0354	0.0840	0.0629	<0.0001	0.0383	<0.0001	0.0378
(s,t)=	(4,12)								
n=100	$C \sim U[0,30]$	0.0328	0.0757	0.0761	0.1059	0.0035	0.0798	0.0035	0.0783
	$C \sim U[0,20]$	0.0308	0.0841	0.1193	0.1079	0,0028	0.0902	0.0022	0.0883
n=250	$C \sim U[0,30]$	0.0270	0.0459	0.0672	0.0727	-0.0024	0.0498	-0.0024	0.0486
	$C \sim U[0,20]$	0.0271	0.0530	0.1106	0.0744	-0.0027	0.0573	-0.0020	0.0562
(s,t)=	(8,12)								
n=100	$C \sim U[0,30]$	0.0085	0.0842	0.0505	0.1056	<0.0001	0.0872	-0.0012	0.0861
	$C \sim U[0,20]$	0.0077	0.0987	0.0738	0.1216	0.0013	0.1040	-0.0013	0.1004
n=250	$C \sim U[0,30]$	0.0095	0.0530	0.0508	0.0673	<0.0001	0.0550	<0.0001	0.0542
	$C \sim U[0,20]$	0.0085	0.0596	0.0720	0.0750	-0.0010	0.0620	-0.0011	0.0608

TABLE 3.3: Bias and standard deviation (SD) for the three estimators of $p_{13}(s,t)$. Non-Markov scenario with three sample sizes and two censoring levels.

		$\widehat{p}_{13}^{\mathtt{AJ}}(s,t)$		$\widehat{p}_{13}^{\mathtt{LIDA}}(s,t)$		$\widehat{p}_{13}^{\mathtt{cLIDA}}(s,t)$		$\widehat{p}_{13}^{\mathtt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)								
n=100	$C \sim U[0,30]$	0.0018	0.0282	0.0147	0.0488	-0.0019	0.0286	-0.0019	0.0285
	$C \sim U[0,20]$	0.0035	0.0318	0.0376	0.0563	<0.0001	0.0325	<0.0001	0.0325
n=250	$C \sim U[0,30]$	0.0034	0.0185	0.0123	0.0345	<0.0001	0.0188	<0.0001	0.0188
	$C \sim U[0,20]$	0.0030	0.0191	0.0309	0.0413	<0.0001	0.0194	<0.0001	0.0194
(s,t)=	(2,8)								
n=100	$C \sim U[0,30]$	0.0131	0.0563	0.0252	0.0936	-0.0017	0.0588	-0.0014	0.0586
	$C \sim U[0,20]$	0.0119	0.0567	0.0792	0.0925	<0.0001	0.0603	<0.0001	0.0596
n=250	$C \sim U[0,30]$	0.0164	0.0349	0.0257	0.0584	0.0027	0.0367	0.0028	0.0364
	$C \sim U[0,20]$	0.0140	0.0365	0.0686	0.0711	<0.0001	0.0392	<0.0001	0.0390
(s,t)=	(4,12)								
n=100	$C \sim U[0,30]$	0.0176	0.0749	0.0316	0.1103	<0.0001	0.0801	-0.0015	0.0786
	$C \sim U[0,20]$	0.0212	0.0832	0.0900	0.1197	0.0045	0.0906	0.0034	0.0869
n=250	$C \sim U[0,30]$	0.0204	0.0468	0.0240	0.0783	<0.0001	0.0506	<0.0001	0.0498
	$C \sim U[0,20]$	0.0179	0.0527	0.0815	0.0841	-0.0017	0.0583	-0.0015	0.0567
(s,t)=	(8,12)								
n=100	$C \sim U[0,30]$	0.0145	0.0814	0.0259	0.1077	0.0039	0.0842	0.0038	0.0835
	$C \sim U[0,20]$	0.0060	0.1005	0.0610	0.1246	-0.0038	0.1032	-0.0034	0.1031
n=250	$C \sim U[0,30]$	0.0087	0.0527	0.0189	0.0709	<0.0001	0.0541	<0.0001	0.0539
	$C \sim U[0,20]$	0.0114	0.0619	0.0596	0.0815	0.0016	0.0645	0.0017	0.0636

Tables 3.4, 3.5 and 3.6 report the results for five different estimators for the transition probability $p_{22}(s,t)$. Results reported in Table 3.4 reveal that the Aalen-Johansen estimator is the preferred since it reports unbiased estimates with smaller variance in estimation. This was expected since the process is Markovian in this scenario. Again, it is important

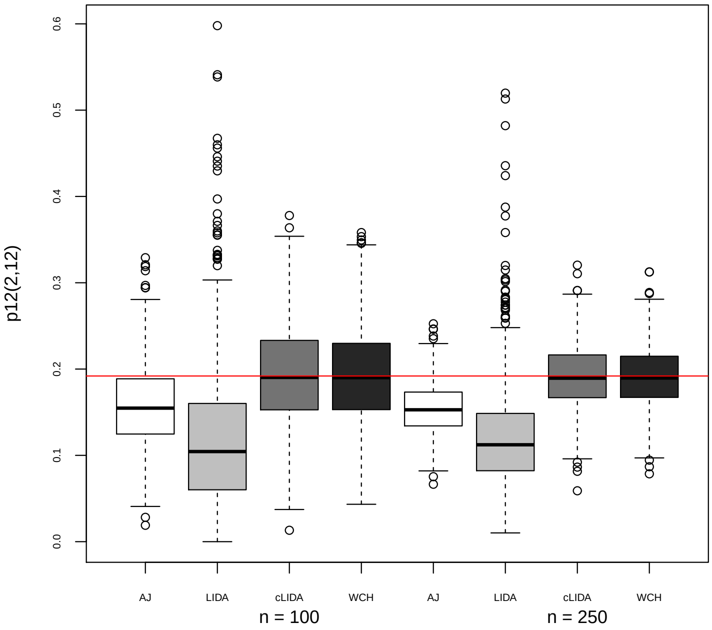


FIGURE 3.1: Boxplots of the M = 1000 estimates of the transition probabilities of the $\widehat{p}_{12}^{\mathtt{AJ}}$, $\widehat{p}_{12}^{\mathtt{LIDA}}$, $\widehat{p}_{12}^{\mathtt{cLIDA}}$ and $\widehat{p}_{12}^{\mathtt{WCH}}$ with two different samples sizes for semi-Markovian scenario. Censoring times were generated from an uniform distribution on $[0, 30]$.

mentioning that this estimator which assumes the process to be Markovian still performs reasonably well in situations where the process shows only mild deviations from Markovianity. This occurs for example in the semi-Markov scenario with small lag times $t - s$. In these cases, the Aalen-Johansen reports estimates with small bias but less variability and therefore low mean squared errors. As the lag times $t - s$ increase so the bias resulting in a clear biased estimator. This behavior is also present in the non-Markov scenario (Table 3.6). Results shown in Tables 3.5 and 3.6 reveal that when there is strong evidence that the process is not Markov that the use of a non-Markov estimator is preferable. With the exception of the `LIDA` method all the remaining non-Markov methods (`cLIDA`, `LM` and `WCH`) are valid alternative estimators due to their greater accuracy. Again, the performance of the `LIDA` method is poorer even worst than the Aalen-Johansen estimator. Simulation results reveal that the `LIDA` estimator is systematically (downward) biased whereas the three non-Markov methods `cLIDA`, `LM` and `WCH` are asymptotically unbiased. The best performance is attained by the non-Markov methods (`cLIDA`, `LM` and `WCH`) which lead to more efficient estimation of the transition probabilities. This can be seen in all measures (bias, standard deviation and mean square error). However, when considering all scenarios and all pairs (s,t) neither of the two methods seems to be uniformly best for estimating $p_{22}(s,t)$. However, the landmark method LM reveals in most cases less variability and therefore better results (with less mean square errors) than the remaining non-Markov estimators.

For completeness purposes we show in Figures 3.2, 3.3 and 3.4 the boxplots of the estimates of the transition probability $p_{22}(s,t)$ based on 1000 Monte Carlo replicates for the five estimators, with different sample sizes. The boxplots shown in these figures are in agreement with our findings reported in Table 3.4, 3.5 and 3.6. From these plots it can be seen that the LIDA estimator of $p_{23}(s,t)$ is systematically (downward) biased and that the AJ estimator may also lead to biased estimates (but with less variability) under deviations from Markovianity. Under a Markov scenario (Figure 3.2), all estimators but the LIDA estimator revealed to be unbiased and with a variance that decrease with the sample size. The AJ estimator is preferable in this case because it provides less variability. When the multi-state model is not Markov, this is no longer the case. Despite of offering a small variability, the bias associated to Aalen-Johansen estimator in non-Markov scenarios (Figures 3.3 and 3.4) makes this approach unappropriated. The methods labeled as LM and WCH are recommended in these cases.

TABLE 3.4: Bias and standard deviation (SD) for the five estimators of $p_{22}(s,t)$. Markov scenario with two sample sizes and two censoring levels.

		$\widehat{p}_{22}^{\mathtt{AJ}}(s,t)$		$\widehat{p}_{22}^{\mathtt{LIDA}}(s,t)$		$\widehat{p}_{22}^{\mathtt{cLIDA}}(s,t)$		$\widehat{p}_{22}^{\mathtt{LM}}(s,t)$		$\widehat{p}_{22}^{\mathtt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)										
n=100	$C \sim U[0,30]$	0.0017	0.0547	-0.0409	0.0960	<0.0001	0.0597	<0.0001	0.0599	<0.0001	0.0598
	$C \sim U[0,20]$	<0.0001	0.0547	-0.0926	0.1330	-0.0012	0.0598	-0.0014	0.0596	-0.0013	0.0596
n=250	$C \sim U[0,30]$	0.0010	0.0341	-0.0361	0.0552	<0.0001	0.0371	<0.0001	0.0372	<0.0001	0.0372
	$C \sim U[0,20]$	-0.0012	0.0365	-0.0778	0.0818	<0.0001	0.0410	<0.0001	0.0408	<0.0001	0.0408
(s,t)=	(2,8)										
n=100	$C \sim U[0,30]$	<0.0001	0.0788	-0.1129	0.1626	<0.0001	0.0969	0.0013	0.0962	0.0013	0.0963
	$C \sim U[0,20]$	<0.0001	0.0812	-0.2364	0.2174	-0.0011	0.1017	-0.0013	0.0993	<0.0001	0.0992
n=250	$C \sim U[0,30]$	0.0030	0.0488	-0.0949	0.0976	0.0033	0.0597	0.0028	0.0592	0.0029	0.0592
	$C \sim U[0,20]$	-0.0029	0.0526	-0.2143	0.1363	-0.0025	0.0651	-0.0029	0.0634	-0.0028	0.0635
(s,t)=	(4,12)										
n=100	$C \sim U[0,30]$	-0.0015	0.0845	-0.1550	0.1687	<0.0001	0.0981	-0.0016	0.0967	-0.0016	0.0972
	$C \sim U[0,20]$	0.0018	0.0958	-0.3451	0.2312	0.0026	0.1119	0.0028	0.1078	0.0017	0.1081
n=250	$C \sim U[0,30]$	<0.0001	0.0555	-0.1376	0.1055	0.0021	0.0635	0.0019	0.0622	0.0017	0.0627
	$C \sim U[0,20]$	-0.0021	0.0589	-0.2968	0.1566	-0.0023	0.0689	-0.0017	0.0668	-0.0019	0.0675
(s,t)=	(8,12)										
n=100	$C \sim U[0,30]$	<0.0001	0.0789	-0.1262	0.1568	-0.0026	0.0821	-0.0028	0.0819	-0.0030	0.0830
	$C \sim U[0,20]$	-0.0046	0.0952	-0.3220	0.2814	-0.0056	0.0992	-0.0065	0.0979	-0.0084	0.1024
n=250	$C \sim U[0,30]$	-0.0018	0.0497	-0.1030	0.0930	-0.0019	0.0513	-0.0018	0.0507	-0.0018	0.0514
	$C \sim U[0,20]$	<0.0001	0.0578	-0.2594	0.1730	<0.0001	0.0615	<0.0001	0.0605	<0.0001	0.0618

TABLE 3.5: Bias and standard deviation (SD) for the five estimators of $p_{22}(s,t)$. Semi-Markov scenario with two sample sizes and two censoring levels.

		$\widehat{p}_{22}^{\texttt{AJ}}(s,t)$		$\widehat{p}_{22}^{\texttt{LIDA}}(s,t)$		$\widehat{p}_{22}^{\texttt{cLIDA}}(s,t)$		$\widehat{p}_{22}^{\texttt{LM}}(s,t)$		$\widehat{p}_{22}^{\texttt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)										
n=100	$C \sim U[0,30]$	0.0207	0.0842	-0.0105	0.1081	-0.0013	0.0989	<0.0001	0.0976	<0.0001	0.098
	$C \sim U[0,20]$	0.0191	0.0908	-0.0333	0.1266	<0.0001	0.1047	<0.0001	0.1040	<0.0001	0,1048
n=250	$C \sim U[0,30]$	0.0230	0.0538	-0.0039	0.0668	0.0027	0.0625	0.0033	0.0614	0.0034	0.0620
	$C \sim U[0,20]$	0.0186	0.0576	-0.0279	0.0813	-0.0038	0.0677	-0.0037	0.0664	-0.0036	0.0668
(s,t)=	(2,8)										
n=100	$C \sim U[0,30]$	0.0805	0.0901	-0.0182	0.1388	<0.0001	0.1280	0.0022	0.1207	0.0027	0.1208
	$C \sim U[0,20]$	0.0776	0.0948	-0.0704	0.1606	-0.0011	0.1318	0.0000	0.1225	<0.0001	0.1250
n=250	$C \sim U[0,30]$	0.0799	0.0544	-0.0122	0.0858	0.0025	0.0749	0.0042	0.0704	0.0041	0.0712
	$C \sim U[0,20]$	0.0794	0.0590	-0.0524	0.1075	<0.0001	0.0839	-0.0010	0.0761	-0.0010	0.0773
(s,t)=	(4,12)										
n=100	$C \sim U[0,30]$	0.0665	0.1004	-0.0626	0.1498	-0.0017	0.1269	-0.0024	0.1179	-0.0034	0.1216
	$C \sim U[0,20]$	0.0732	0.1115	-0.1436	0.1847	0.004	0.1481	0.0058	0.1337	<0.0001	0.1414
n=250	$C \sim U[0,30]$	0.0703	0.0612	-0.0442	0.1000	0.0026	0.0822	0.0026	0.0736	0.0027	0.0759
	$C \sim U[0,20]$	0.0714	0.0691	-0.1196	0.1285	0.0011	0.0941	0.0043	0.0822	0.0049	0.0860
(s,t)=	(8,12)										
n=100	$C \sim U[0,30]$	0.0207	0.1201	-0.0949	0.1845	0.0017	0.1376	0.0027	0.1308	<0.0001	0.1382
	$C \sim U[0,20]$	0.0138	0.1381	-0.2551	0.2722	-0.0037	0.1572	-0.0056	0.1484	-0.0136	0.1712
n=250	$C \sim U[0,30]$	0.0189	0.0742	-0.0693	0.1117	0.0018	0.0825	0.0027	0.0791	0.0028	0.0823
	$C \sim U[0,20]$	0.0197	0.0893	-0.1973	0.1748	0.0016	0.1031	0.0025	0.0962	-0.0016	0.1052

TABLE 3.6: Bias and standard deviation (SD) for the five estimators of $p_{22}(s,t)$. Non-Markov scenario with three sample sizes and two censoring levels.

		$\widehat{p}_{22}^{\texttt{AJ}}(s,t)$		$\widehat{p}_{22}^{\texttt{LIDA}}(s,t)$		$\widehat{p}_{22}^{\texttt{cLIDA}}(s,t)$		$\widehat{p}_{22}^{\texttt{LM}}(s,t)$		$\widehat{p}_{22}^{\texttt{WCH}}(s,t)$	
		bias	SD	bias	SD	bias	SD	bias	SD	bias	SD
(s,t)=	(2,4)										
n=100	$C \sim U[0,30]$	0.0093	0.0492	-0.0074	0.0661	<0.0001	0.0573	<0.0001	0.0571	<0.0001	0.0576
	$C \sim U[0,20]$	0.0058	0.0512	-0.0322	0.0857	-0.0016	0.0586	-0.0018	0.0584	-0.0022	0.0588
n=250	$C \sim U[0,30]$	0.0071	0.0308	-0.0078	0.0408	-0.0011	0.0362	-0.0011	0.0360	-0.0014	0.0362
	$C \sim U[0,20]$	0.0082	0.0318	-0.0262	0.0513	<0.0001	0.0373	<0.0001	0.0370	<0.0001	0.0372
(s,t)=	(2,8)										
n=100	$C \sim U[0,30]$	0.0347	0.0797	-0.0300	0.1260	-0.0026	0.1000	-0.0017	0.0992	-0.0023	0.0996
	$C \sim U[0,20]$	0.0334	0.0843	-0.1050	0.1680	0.0033	0.1058	0.0017	0.1050	0.0014	0.1051
n=250	$C \sim U[0,30]$	0.0309	0.0490	-0.0268	0.0781	-0.0030	0.0629	-0.0028	0.0621	-0.0035	0.0624
	$C \sim U[0,20]$	0.0307	0.0523	-0.0992	0.1037	-0.0046	0.0663	-0.0045	0.0642	-0.0052	0.0646
(s,t)=	(4,12)										
n=100	$C \sim U[0,30]$	0.0323	0.0890	-0.0511	0.1340	0.0015	0.1065	<0.0001	0.1026	-0.0025	0.1042
	$C \sim U[0,20]$	0.0325	0.1027	-0.1729	0.1922	0.0019	0.1259	0.0021	0.1213	-0.0015	0.1240
n=250	$C \sim U[0,30]$	0.0322	0.0552	-0.0410	0.0871	<0.0001	0.0668	-0.0010	0.0643	-0.0036	0.0648
	$C \sim U[0,20]$	0.0387	0.0626	-0.1504	0.1286	0.0064	0.0793	0.0068	0.0746	0.0041	0.0760
(s,t)=	(8,12)										
n=100	$C \sim U[0,30]$	0.0085	0.0934	-0.0463	0.1270	-0.0035	0.1015	-0.0040	0.0999	-0.0097	0.1030
	$C \sim U[0,20]$	0.0145	0.1061	-0.1854	0.2219	0.0032	0.1157	0.0041	0.1125	-0.0046	0.1220
n=250	$C \sim U[0,30]$	0.0115	0.0575	-0.0340	0.0783	<0.0001	0.0619	<0.0001	0.0606	-0.0054	0.0639
	$C \sim U[0,20]$	0.0111	0.0689	-0.1534	0.1366	<0.0001	0.0740	<0.0001	0.0727	-0.0067	0.0766

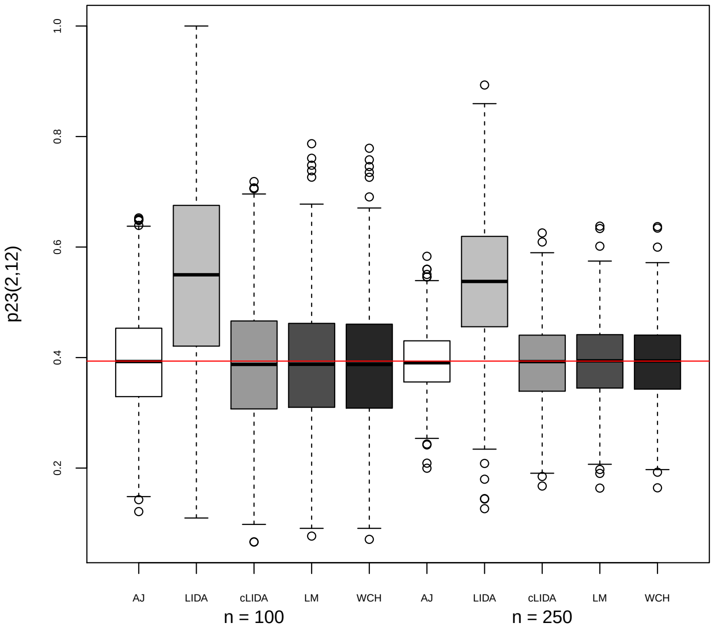


FIGURE 3.2: Boxplots of the M = 1000 estimates of the transition probabilities of the $\widehat{p}_{22}^{\mathtt{AJ}}$, $\widehat{p}_{22}^{\mathtt{LIDA}}$, $\widehat{p}_{22}^{\mathtt{cLIDA}}$, $\widehat{p}_{22}^{\mathtt{LM}}$ and $\widehat{p}_{22}^{\mathtt{WCH}}$ with two different samples sizes for Markovian scenario. Censoring times were generated from an uniform distribution on $[0, 30]$.

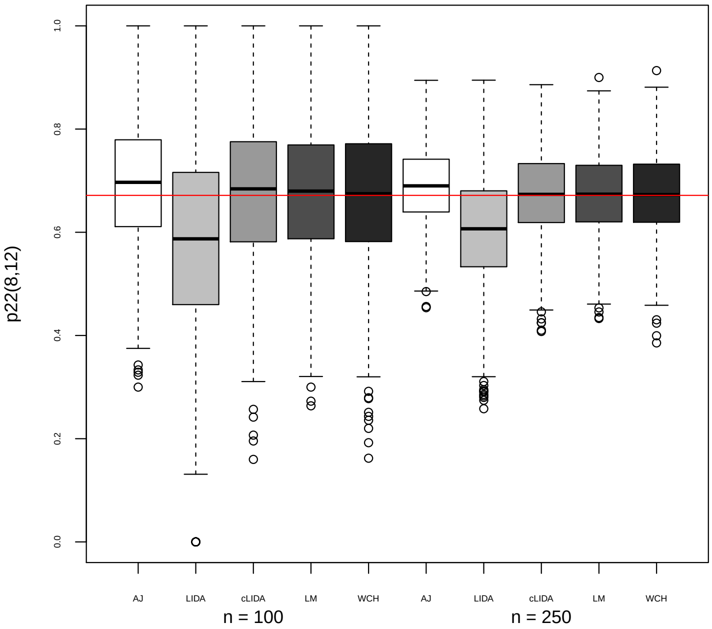


FIGURE 3.3: Boxplots of the M = 1000 estimates of the transition probabilities of the $\widehat{p}_{22}^{\mathtt{AJ}}$, $\widehat{p}_{22}^{\mathtt{LIDA}}$, $\widehat{p}_{22}^{\mathtt{cLIDA}}$, $\widehat{p}_{22}^{\mathtt{LM}}$ and $\widehat{p}_{22}^{\mathtt{WCH}}$ with two different samples sizes for semi-Markovian scenario. Censoring times were generated from an uniform distribution on $[0, 30]$.

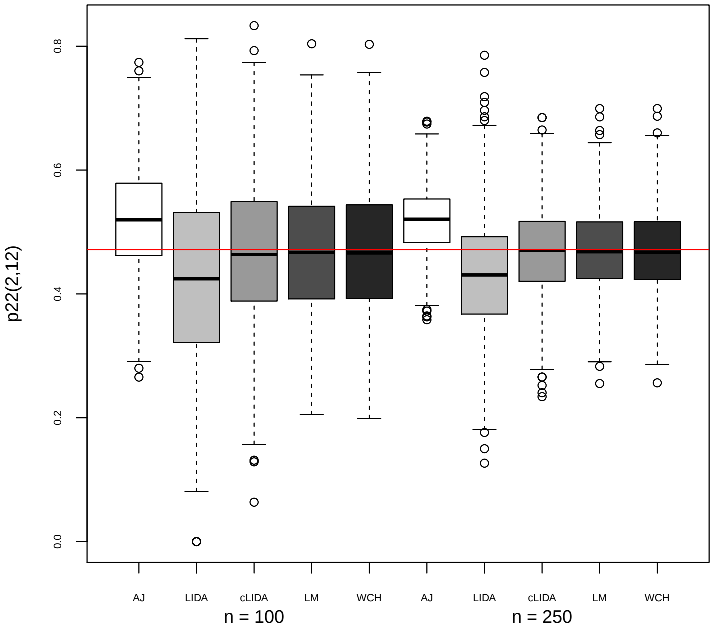


FIGURE 3.4: Boxplots of the M = 1000 estimates of the transition probabilities of the $\widehat{p}_{22}^{\texttt{AJ}}$, $\widehat{p}_{22}^{\texttt{LIDA}}$, $\widehat{p}_{22}^{\texttt{cLIDA}}$, $\widehat{p}_{22}^{\texttt{LM}}$ and $\widehat{p}_{22}^{\texttt{WCH}}$ with two different samples sizes for non-Markovian scenario. Censoring times were generated from an uniform distribution on $[0, 30]$.

3.5 Example of application

Colorectal cancer is one of the most commonly diagnosed cancers worldwide. It is also the one of most frequent causes of cancer-related death in both men and women. Several lifestyle-related factors have been linked to colorectal cancer including diet, weight and exercise. Survival rates for colorectal cancer vary worldwide but they have been associated to several clinical and pathological factors including age, tumor size, lymph nodes with detectable cancer, etc.

Surgical resection is the best treatment option for patients with colorectal cancer and the most powerful tool for assessing prognosis following potentially curative surgery. In a large percentage of the patients with colorectal cancer, the diagnosis is made at a sufficiently early stage when all apparent disease tissue can be surgically removed. Unfortunately, some of these patients have residual cancer, which leads to recurrence of the disease and death (in some cases). Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of mortality. This mortality is higher in cases of early recurrences. Traditionally, the effect of these covariates is studied using the Cox proportional hazards model (Cox (1972) [6]) with time-dependent covariates. The analysis of such studies can also be successfully performed using a multi-state model (Putter, Fiocco and Geskus (2007) [18]; Meira-Machado *et al.* (2009) [15]).

In this section we re-analyzed data from one of the first successful trials of adjuvant chemotherapy for colon cancer (Moertel *et al.* (1990) [74]). In this data set we have a total of 929 patients from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer. In this study, patients were followed from the date of cancer diagnosis until censoring or death. A total of 468 patients developed a recurrence and among these 414 died; 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up. Cancer recurrence affects the patient's outcome and can be included as a transient state in a progressive illness-death model with states 'alive and disease-free' (State 1), 'alive with recurrence' (State 2) and 'death' (State 3).

Besides recurrence, the sojourn time in State 1, the total time of the process and the corresponding indicator statuses are known for each individual. Additional covariates such as age, sex, tumor size and lymph nodes with detectable cancer, are also available. In this chapter we consider early recurrence as recurrence within 1 year after primary surgery of colorectal cancer. Considering overall survival after recurrence, the median survival time in the early recurrence group was 498 days with overall survival rates of 26.6%, 10.8% and 6.3% for 2, 3 and 5 years after surgery. As expected, better results (p-value < 0.001) were obtained for patients in the late recurrence group (i.e., with a time to recurrence greater than 1 year after surgery). The median survival time in this group was 1292 days with overall survival rates of 86.7%, 65.4% and 31.3% for 2, 3 and 5 years after surgery. These results confirm that recurrence has a negative impact in the prognosis and to a premature recurrence.

Statistical methods for analyzing data in an illness-death multi-state model depend on the Markov assumption. By ignoring the disease history behavior (e.g., states previously visited and the transition times among them), these models may carry severe limitations which can make the model inappropriate. It is a fact that the future health of individuals with an early recurrence may be different from those who have been healthy for a long time. In addition, the risk of death is known to increase shortly after the recurrence, which reveals that the length of stay in the recurrence state is relevant for prognosis, thus invalidating the memoryless property of Markov processes. Accordingly, the Markov assumption can be checked by including covariates depending on the history. This 'global' test for Markovianity based on the Cox model (using time to recurrence as a covariate) reported a coefficient of negative sign for the recurrence time, not revealing a possible increased risk of death shortly after relapse (p-value = 0.154).

Since several estimators introduced in this chapter are consistent regardless the Markov condition they can be used to introduce a 'local' test for the Markov condition by measuring the discrepancy in the estimates obtained from these estimators to those obtained using the Aalen-Johansen estimators (only consistent if the process is Markov). Graphical comparisons of the transition probabilities between the two approaches are reported in Figure 3.5. This figure depicts the discrepancy between the landmark non-Markovian estimator (LM) and the Aalen-Johansen estimator (Markovian), for $p_{12}(s,t)$ and $p_{22}(s,t)$, for $s = 365$, $s = 730$ and $s = 1095$ ($D_{ij} = p_{ij}^{\mathtt{LM}}(s,t) - p_{ij}^{\mathtt{AJ}}(s,t)$). The 95% pointwise confidence bands are based on simple bootstrap are also shown, revealing clear differences between the two methods in large intervals for $s = 365$. In this case, since there exists a deviation of the plot with respect to the straight line $y = 0$, one gets some evidence on the lack of Markovianity of the underlying process beyond one year after surgery. On the other hand, the plots depicted on the second and third row no not reveal evidence against the Markov assumption. In summary, these plots reveal some evidence, at least for $s = 365$, that the use of Markov-free estimators such as those proposed in this chapter are more suited to estimate the transition probabilities $p_{12}(s,t)$, $p_{13}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$. This topic of testing the Markov assumption in multi-state models will be more deeply analysed in Chapter 6 by the introduction of 'local' and 'global' tests based on the differencies between AJ and LM estimators.

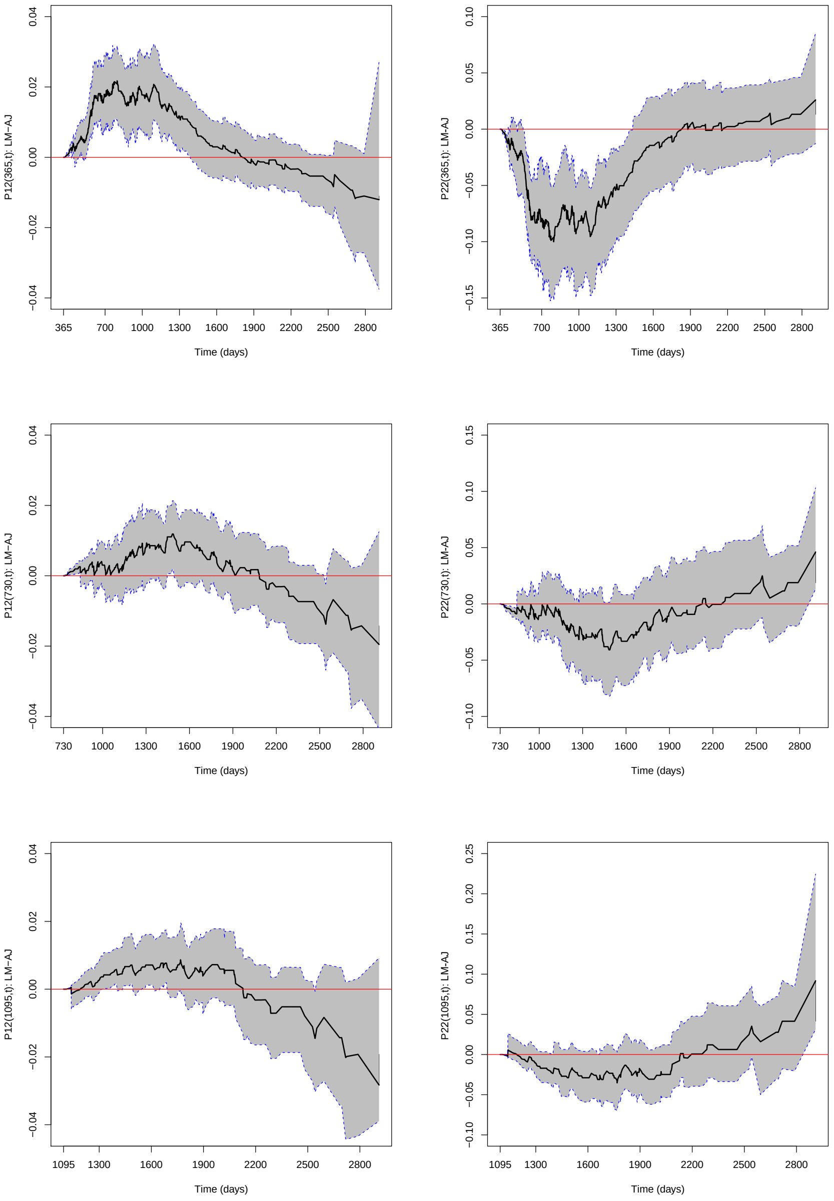


FIGURE 3.5: Local graphical test for the Markov condition, for $s = 365$ (top), $s = 730$ (middle) and $s = 1095$ (bottom). Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (LM). Colon cancer data.

With this example of application we are also interested in illustrating differences between the estimated transition probabilities from the estimators introduced in Section 3.3. These quantities can be used to obtain prediction probabilities of future events (e.g., recurrence and death in cancer studies). In Figure 3.6 we present, as an example, estimated transition probabilities for $p_{12}(s,t)$ and $p_{22}(s,t)$, with $s = 365$, $s = 730$ and $s = 1095$ (corresponding to 1, 2 and 3 years) for the colon cancer data, showing that a choice between the different methods makes a big difference. As shown in our simulations the estimators by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] provide curves that are almost always below those obtained by the new estimators. This is more clear in the transition probability $p_{22}(s,t)$ for higher values of s. This is in agreement with our simulation results that suggested a systematic negative bias for the estimator by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] (i.e. a downward biased estimator).

Since few events ('death') are observed at higher time values, consistency problems are expected at the right tail of the distribution when using the estimator by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68]. These features can be seen in all plots but especially in the figures of the transition probability $p_{22}(s,t)$. While both LM and WCH estimators decrease smoothly with time the estimator by Meira-Machado, de Uña-Álvarez and Cadarso-Suárez (2006) [68] shows a sharp decrease to zero.

All plots depicted in right hand side of Figure 3.6 reveal a similar behavior of the LM and WCH estimators of the transition probability $p_{22}(s,t)$. These plots report the survival fraction along time, among the individuals in the recurrence state 1 year (Figure 3.6, top), and 2 years (Figure 3.6, middle) and 3 years (Figure 3.6, bottom) after surgery. They reveal that patients with an early recurrence have lower survival probabilities. When comparing the two Markov-free methods with the Aalen-Johansen estimator (AJ) one can observe some differences for $s = 365$ which are less evident as s increases. These discrepancies can be explained by the failure of the Markov assumption as shown in Figure 3.5. Similarly, differences can also be observed between AJ and WCH estimators for the transition probability $p_{12}(s,t)$. Summarizing, it becomes clear from this application that, at least for $s = 365$, the use of Markov-free estimators such as LM and WCH are preferred over the Aalen-Johansen estimator.

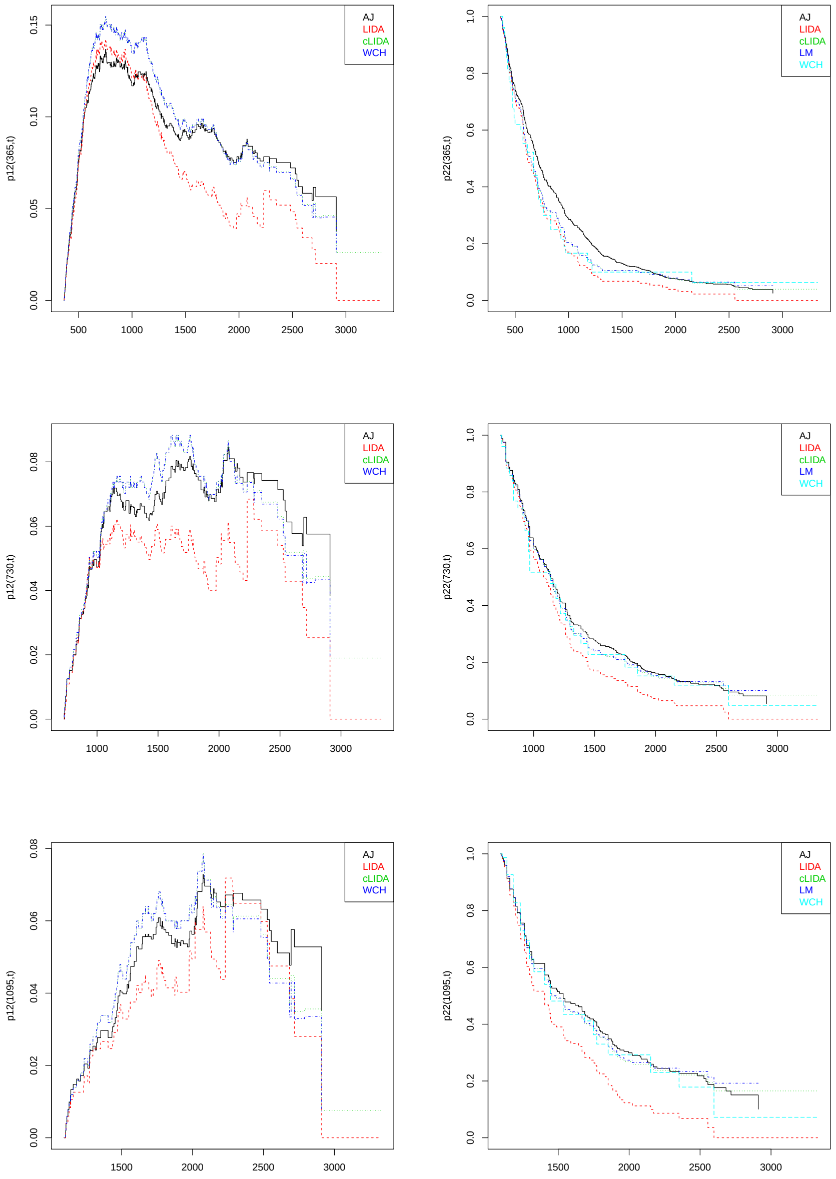


FIGURE 3.6: Estimated transition probabilities for $p_{12}(s,t)$ and $p_{22}(s,t)$, $s = 365$ (top), $s = 730$ (middle) and $s = 1095$ (bottom). Colon cancer data.

3.6 Discussion

There has been a remarkable surge of activity lately on the topic of nonparametric estimation of transition probabilities in multi-state models. Most recent contributions on this topic are in the context of non-Markov multi-state models since the Aalen-Johansen estimator is still the preferred and standard estimator when one is confident of the Markov assumption. In a recent paper, de Uña-Álvarez and Meira-Machado (2015) [69] has used the idea of subsampling to introduce the landmark estimators that are consistent regardless the Markov condition. In this chapter, we revisit the topic of the nonparametric estimation of the transition probabilities, by introducing competing estimators in a multi-state system that is not necessarily Markovian and that overcomes the referred assumption on the censoring support. The new set of estimators are constructed using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. One of the proposed estimator is equivalent to the landmark estimators. To evaluate the performance of all estimators, several simulation studies were conducted under different data scenarios. Based on these results recommendations are given as to which estimators to use. Results obtained from several simulation studies conducted under different data scenarios show that the new method and the proposals introduced by de Uña-Álvarez and Meira-Machado (2015) [69] are quite similar providing accurate estimates.

The comparison between estimated transition probabilities was also the basis to introduce a graphical local test for the Markov assumption. Following this approach, new methods, based on measuring the discrepancy of the Aalen-Johansen estimator which gives consistent estimators in Markov processes, and recent approaches that do not rely on this assumption, will be presented in Chapter 6. Our simulation results also show that the Aalen-Johansen estimator provides biased estimates if the Markov assumption does not hold. In most of these cases the use of a non-Markov estimator is preferable due to their greater accuracy. Therefore, one important issue is how to test the Markov assumption.

Chapter 4

A comparison of Presmoothing methods in the estimation of transition probabilities

The estimation of transition probabilities is of major importance in the analysis of survival data with multiple events. These quantities play an important role in the inference in multi-state modeling providing in a simple and summarized manner long-term predictions of the process. Recently, de Uña-Álvarez and Meira-Machado (2015) [69] proposed nonparametric estimators based on subsampling, also known as landmarking, which have already proved to be more efficient than other nonparametric estimators in case of strong violation of the Markov condition. However, as the idea behind the landmarking is to use specific portions of data, when the subsample sizes are reduced or in the presence of heavily censored data, this may lead to higher variability of the estimates.

To avoid the high variability of the landmark estimators, in this chapter we introduce estimators based on presmoothing which are obtained by replacing the censoring indicator variables in the classical definitions by values of a regression estimator. Results of simulation studies confirm the good behavior of the proposed methods. We also illustrate and compare these new methods to the nonparametric landmark estimator through a real data set on colon cancer.

The contents of this chapter are mainly based on the paper published in Communications in *Statistics - Simulation and Computation* by Soutinho, Meira-Machado and Oliveira (2020) [44].

4.1 Introduction

The landmark estimators are based on the computation of the so-called occupation probabilities in a subset of individuals that happen to be in a given state at a particular time. Because of this, since the Kaplan-Meier estimator is a step function with jumps located only at the uncensored observations, in some cases the computation of the transition probabilities are based on reduced data and usually in the presence of heavily censored data the accuracy of their estimation might not be acceptable, in particular, at the right tail of distribution.

To avoid this problem, Meira-Machado (2016) [73] proposed an approach that can be used to reduce the variability of the landmark estimator. This approach is based on spline smoothing while the other is based on a preliminary parametric estimation (presmoothing) of the probability of censoring. The main idea of presmoothing is that each censoring indicator is replaced by a smooth fit of a binary regression of the indicator on observables. The use of presmoothed estimators is a good alternative in these situations, since they give mass to all the event times, including the censored observations. Successful applications of presmoothed estimators include estimation of the survival function (Dikta (1998) [65]; Meira-Machado, Sestelo and Gonlçalves (2016) [75]), nonparametric curve estimation (Cao and Jácome (2004) [76]), regression analysis (de Uña-Álvarez and Rodríguez-Campos (2004) [77]; Jácome and Iglesias (2010) [78]), estimation of the bivariate distribution of censored gap times (de Uña-Álvarez and Amorim (2011) [79]), and the estimation of the transition probabilities (Amorim, de Uña-Álvarez and Meira-Machado (2011) [80]). All these references concluded that the presmoothed estimators have improved variance when compared to purely nonparametric estimators.

Following, we review some recent developments on the presmoothed estimation of the transition probabilities. Specifically, we focus on the choice of the preliminary smoothing function which may be based on a certain parametric family such as logistic, probit, cauchit; or on a nonparametric regression smoother such as the Nadaraya-Watson kernel estimator (Wand and Jones (1997) [81]). Thus, in Section 4.2 we introduce the proposed presmoothing estimators. The performance of four sets of estimators is investigated through simulations in Section 4.3, while in Section 4.4 the methods are compared through the analysis of medical data from a controlled clinical trial in liver cirrhosis. Software development is presented in 4.5. Main conclusions are reported in Section 4.6.

4.2 Presmoothing methods

In terms of notation let's consider the same of Section 3.1, in which the movement of the individuals through a progressive illness-death model is given by $(X(t), t \in [0, \infty))$, but now with state space $\{0, 1, 2\}$, with the available data represented by $(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i)$, $1 \leq i \leq n$, i.i.d. copies of $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta)$. Under the landmark approach described in 3.3.3, the presmoothed Kaplan-Meier estimators are obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights with some smooth fit before the Kaplan-Meier formula is applied. Thus, the presmoothed Kaplan-Meier estimator of the survival function of T is given by

$$\widehat{S}^{\texttt{PrKM}}(t) = 1 - \sum_{i=1}^{n} \widetilde{W}_i I(\widetilde{T}_{(i)} \leq t), \tag{4.1}$$

where

$$\widetilde{W}_i = \frac{p_n(\widetilde{T}_{(i)})}{n - i + 1} \prod_{j=1}^{i-1} \left[1 - \frac{p_n(\widetilde{T}_{(j)})}{n - j + 1} \right], 1 \leq i \leq n, \tag{4.2}$$

and $p_n(t)$ stands for an estimator of the binary regression function $p(t) = P(\Delta = 1 \mid \widetilde{T} = t)$, i.e., the conditional probability that the observation at time t is not censored.

We may now formally introduce the presmoothed landmark estimators as follows

$$\widehat{p}_{00}^{\texttt{PrLM}}(s, t) = \widehat{S}_0^{\texttt{PrKM}(s)}(t) \tag{4.3}$$

$$\widehat{p}_{01}^{\texttt{PrLM}}(s, t) = \widehat{S}^{\texttt{PrKM}(s)}(t) - \widehat{S}_0^{\texttt{PrKM}(s)}(t) \tag{4.4}$$

$$\widehat{p}_{02}^{\texttt{PrLM}}(s, t) = 1 - \widehat{S}^{\texttt{PrKM}(s)}(t) \tag{4.5}$$

$$\widehat{p}_{11}^{\texttt{PrLM}}(s, t) = \widehat{S}^{\texttt{PrKM}[s]}(t) \tag{4.6}$$

$$\widehat{p}_{12}^{\texttt{PrLM}}(s, t) = 1 - \widehat{S}^{\texttt{PrKM}[s]}(t) \tag{4.7}$$

where $\widehat{S}_0^{\texttt{PrKM}(s)}$ and $\widehat{S}^{\texttt{PrKM}(s)}$ are the presmoothed Kaplan-Meier estimators for the distributions of Z and T, respectively, but computed from the respective subsamples.

One useful parametric candidate for the binary regression function $p(t)$ belongs to a parametric family of binary regression curves, such as logit or probit. When the parametric model specified for $p(t)$ is correct the corresponding semiparametric presmoothed estimator is at least as efficient as the original nonparametric Kaplan-Meier estimator. Importantly, the validity of a given model for the presmoothing function can be checked graphically or formally, by applying a goodness-of-fit test such as the test proposed by Hosmer and Lemeshow (2003) [82] for the logistic model or the Kolmogorov-Smirnov type version of the model-based bootstrap approach described in Dikta, Kvesic and Schmidt (2006) [83]. This implies that the risk of a misspecified model can be controlled in practice. An alternative and more flexible approach is to model the binary regression function through an additive regression model.

Nonparametric presmoothing (Cao *et al.* (2005) [66]) is useful when there is a clear risk of a miss-specification of the parametric model. In this case, one may use the Nadaraya-Watson kernel estimator for $p(\cdot)$ based on the binary responses Δ_i with covariates T_i. The idea is to estimate $p(\widetilde{T}_i)$ by $p_n(\widetilde{T}_i)$ where

$$p_n(t) = \frac{\sum_{i=1}^{n} K\left((t - T_i)/a_n\right) \Delta_i}{\sum_{i=1}^{n} K\left((t - T_i)/a_n\right)} \tag{4.8}$$

and K is a known probability density function (the kernel function) and a_n is a sequence of bandwidths. As a drawback, nonparametric regression requires the specification of a bandwidth for the computation of the smooth fit $p_n(t)$. As a_n decreases, the roughness of the resulting estimator will increase. When $a_n \to 0$, then, the presmoothed estimator coincides in the limit with the classical estimators. On the other hand, as the bandwidth a_n increases, oversmoothed estimates will be obtained removing important features of the underlying structure of the survival function to be estimated.

Different presmoothed estimators are obtained depending on the choice of the preliminary smoothing function. Simulation studies reported in the next section show that in both cases the presmoothed estimators may be much more efficient than the completely nonparametric estimator, since they often have less variance while providing smoother curves with the expected behavior.

4.3 Simulation studies

In this section, we report results from two simulation studies where the aim is to compare the finite sample performance of our presmoothed estimators of the transition probabilities. To be specific, we compare the original landmark estimator of de Uña-Álvarez and Meira-Machado (2015) [69] (labeled as `LM`) with the semiparametric presmoothed estimator which is obtained through the use of a logistic regression model for the presmoothing function (labeled as `PLM`) and the presmoothed estimator based on a nonparametric kernel regression model (labeled as `NP`). For completeness purposes, we also included the presmoothed estimator which is obtained through the use of a generalized additive logistic regression model (labeled as `GAM`).

We first simulated data from a scenario used by Amorim, de Uña-Álvarez and Meira-Machado (2011) [80], which these authors found to be challenging both in terms of bias and variance. To be specific, we separately consider the subjects passing through the intermediate state (State 1, in this chapter) at some time and those who directly go to the ultimate state (State 2). For the first subgroup of individuals we generated replicates of $(Z, T-Z)$ according to the bivariate distribution

$$F_{12}(x,y) = F_1(x)F_2(y)\left[1 + \{1 - F_1(x)\}\{1 - F_2(y)\}\right] \tag{4.9}$$

with exponential marginal distribution functions with rate parameter 1. For the second subgroup of individuals ($Z = T$), the value of Z is simulated according to an exponential with rate parameter 1. Random censoring was simulated from uniform distributions $U[0, \tau_G]$ for τ_G equal to 3 and 4. The model with $\tau_G = 4$ results in 24% of censoring on the first gap time Z, and in 47% of censoring on the second gap time $T-Z$. The model with $\tau_G = 3$ increases these censoring levels to 32% and about 57%, respectively. Since we are assuming correlated times for Z and $T-Z$, the simulated model does not satisfies the Markov property. Details of the simulation procedure can be found in Amorim, de Uña-Álvarez and Meira-Machado (2011) [80].

To summarize the results we fixed the values of (s,t) for four different points, corresponding to combinations of the percentiles 20%, 40%, 60% and 80% of the exponential marginal distribution functions with rate parameter 1. In each simulation, 1000 samples were generated with sample sizes 50, 100 and 250. As a measure of efficiency, we took the

Mean Squared Error (MSE) but we also computed the standard deviations (SD) and the Bias for each point (s,t).

The performance of the empirical transition probabilities $\widehat{p}_{ij}^{\,\texttt{LM}}(s,t)$, $\widehat{p}_{ij}^{\,\texttt{PLM}}(s,t)$, $\widehat{p}_{ij}^{\,\texttt{GAM}}(s,t)$ and $\widehat{p}_{ij}^{\,\texttt{NP}}(s,t)$ in the simulations are summarized in Table 4.1. The semiparametric presmoothing (labeled as PLM) requires the estimation of the binary regression functions, such as $p_{0n}(t) = P(\Delta_1 = 1|\widetilde{Z} = t)$ and $p_{1n}(t) = P(\Delta = 1|\widetilde{T} = t)$, to presmooth the Kaplan-Meier estimators $\widehat{S}_0(t)$ and $\widehat{S}(t)$, respectively. After some algebra, it is seen that the function $p_{0n}(t)$ is written as $(\tau_G - t)\ /\ (\tau_G - t + 1)$. The binary function $p_{1n}(t)$, for those individuals that observe a transition to State 1, is given by $\tau_G /\ \big(1 + \eta(t)\big)$, where $\eta(t) = \lambda_G(t)/\lambda_T(t)$, and where $\tau_G(t) = 1/(\tau_G - t)$ is the hazard rate of the censoring variable and

$$\lambda_T(t) = \Big(-e^{-t}\,(4-2t) + e^{-2t}\,(8+4t)\Big)\ /\ \Big(e^{-t}\,(2t-2) + e^{-2t}\,(3+2t)\Big) \tag{4.10}$$

is the hazard rate of T under restriction $\widetilde{Z} < \widetilde{T}$. Plots for these functions shown in Figure 4.1 reveal that they are monotonous and so they can be adequately estimated by logistic regression.

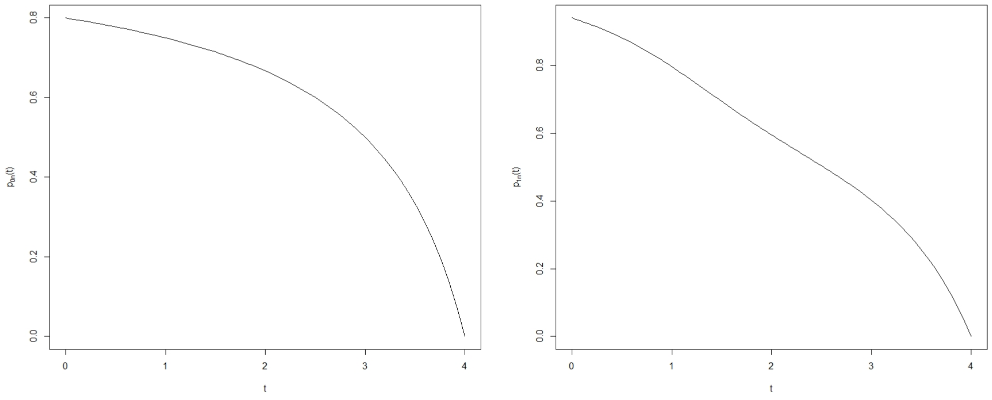


FIGURE 4.1: Theoretical characteristics of the binary regression functions $p_{0n}(t)$ (letf) and $p_{1n}(t)$ (right) for censoring times uniformly distributed between 0 and 4.

In our simulations we have used the logistic regression models $p_n(t;\beta) = 1/(1 + \exp(\beta_0 + \beta_1 t))$. The β parameters in model $p_{0n}(\cdot;\beta)$ were estimated by maximizing of the conditional likelihood of the Δ_1's given $\widetilde{Z}$; and via maximization of the conditional likelihood of the Δ's given $\widetilde{T}$, in model $p_{1n}(\cdot;\beta)$. The presmoothing labeled as `GAM` was implemented by fitting a generalized additive logistic model through the `mgcv` `R` package

(Wood (2019) [84]). To compute the transition probabilities with a nonparametric presmoothing we have used the plug-in bandwidth selector of Cao *et al.* [66] and biweight kernels. As would be expected, results reported in this table reveal that the estimation of the transition probabilities is performed with less accuracy as s and t grow. This behavior was expected since the performance of all methods in lifetime data is usually poorer at the right tail where the censoring effects are stronger. At these points the SD is in most cases larger. The SD decreases with an increase in the sample size and with a decrease of the censoring percentage as usual. All methods obtain in all settings a low bias. It can also be seen that the SD clearly dominates the performance of the proposed estimators in most the cases revealing a clear advantage of the presmoothed estimators when compared with the unsmoothed landmark estimators (labeled as `LM`). This can be observed by the relative efficiency between the unsmoothed landmark estimator and the three presmoothed estimators that was measured by the ratio between their corresponding MSEs. In almost all cases, the use of presmoothing leads to estimators with less SD and less MSE.

Because of space limitation, Table 4.1 does not report the results for transition probability $\widehat{p}_{00}(s,t)$. However, the behavior of the estimators for this transition can be seen in Figures 4.2 and 4.3 through the boxplots of the mean squared errors based on the 1000 Monte Carlo replicates for the four estimators, with different sample sizes and two censoring levels. For completeness purposes we decided to show the plots for three transitions and different fixed values of (s,t). The boxplots shown in these figures reveal some results which are in agreement with our findings reported in Table 4.1. These plots confirm the less variability of the presmoothed estimators. Results shown in Table 4.1 and Figures 4.2 and 4.3 suggest that the use of presmoothing leads to better results for all transition probabilities while neither one of the three presmoothed estimators (`PLM`, `GAM` and `NP`) seems to be uniformly the best.

While reducing the variance, presmoothing may introduce some bias in estimation. Simulation results reported in Table 4.1 serve to illustrate this issue too. When the parametric model specified for the presmoothing function is incorrect, the corresponding semiparametric estimator may lose efficiency, providing estimators with a large bias. However, the validity of a given parametric model for presmoothing can be checked by applying a goodness-of-fit test such as the test proposed by Hosmer and Lemeshow (2003) [82]. We have conducted simulation studies to evaluate the capability of the Hosmer and Lemeshow goodness-of-fit test to detect deviations from the logistic regression model.

Results (not reported here) suggest small deviations in this scenario. In most cases the test was not able to reject the logistic model. Though the simulated scenario seems to be favorable to the estimator based on a parametric preliminary smoothing, the estimator based on nonparametric presmoothing is competitive, attaining better results (with less variance and less mean squared errors) in a large number of points.

To assess the effect of a misspecification of the logistic regression model in the estimation of the transition probabilities through the use of parametric presmoothing methods, a second simulation study was performed. In this scenario, data were generated according to the progressive three-state model. This model can be viewed as a particular case of the illness-death model where no transitions are observed on disease-free mortality transition. The vector of gap times $(Z, T - Z)$ is simulated as follows. The first gap time Z is simulated according to a mixture of two lognormal distributions, $LogN(0, 3)$ and $LogN(3, 0.2)$, with equal probability. Given $Z = x$, the second gap time, $T_{12} = T - Z$, is drawn from a Lognormal distribution with mean $log(x)/2$ and standard deviation equal to 0.2. This scenario does not follow the Markov assumption since the hazard for the second gap time depends on the time to progression to the intermediate state. The censoring time C was independently generated following a uniform distribution $U[10, 30]$. Note that under this scenario, censored observations are expected to be concentrated near the center of the distribution and therefore revealing a misspecification of the fitted logistic regression model.

Results shown in Table 4.2 reveal the impact of a misspecification of the fitted logistic regression model in the estimation of the transition probabilities, leading to estimators with increased bias, in some points, and higher values for the MSE. In such cases the use of an estimator based on nonparametric presmoothing is preferable.

TABLE 4.1: Bias and standard deviation (SD) for estimators of $p_{ij}(s,t)$. The relative MSEs are also given. Scenario 1: illness-death model with correlated exponential gap times.

		$\widehat{p}^{\texttt{LM}}_{01}(s,t)$	$\widehat{p}^{\texttt{PLM}}_{01}(s,t)$	$\widehat{p}^{\texttt{NP}}_{01}(s,t)$	$\widehat{p}^{\texttt{GAM}}_{01}(s,t)$	Relative MSE		
		bias (SD)	bias (SD)	bias (SD)	bias (SD)	LM/PLM	LM/NP	LM/GAM
(s,t)=(0.2231,0.5108)								
n	C							
50	$U[0,4]$	0.0001 (0.0585)	-0.0074 (0.0521)	-0.0079 (0.0499)	-0.0039 (0.0543)	1.2333	1.3405	1.1516
	$U[0,3]$	0.0024 (0.0593)	-0.0063 (0.0515)	-0.0075 (0.0480)	0.0007 (0.0531)	1.3039	1.4922	1.2464
100	$U[0,4]$	-0.0015 (0.0411)	-0.0089 (0.0367)	-0.0092 (0.0355)	-0.0045 (0.0377)	1.1896	1.2575	1.1770
	$U[0,3]$	0.0010 (0.0424)	-0.0074 (0.0360)	-0.0084 (0.0350)	-0.0030 (0.0387)	1.3285	1.3870	1.1916
250	$U[0,4]$	0.0011 (0.0247)	-0.0063 (0.0219)	-0.0065 (0.0215)	-0.0019(0.0225)	1.1850	1.2175	1.2032
	$U[0,3]$	0.0007 (0.0265)	-0.0068 (0.0226)	-0.0081 (0.0226)	-0.0020 (0.0238)	1.2597	1.2252	1.2329
(s,t)=(0.2231,0.9163)								
n	C							
50	$U[0,4]$	0.0012 (0.0746)	-0.0027 (0.0660)	0.0003 (0.0644)	$<$ 0.0001 (0.0682)	1.2777	1.3446	1.1974
	$U[0,3]$	0.0031(0.0751)	-0.0009(0.0655)	0.0009(0.0641)	0.0012(0.0693)	1.3189	1.3760	1.1778
100	$U[0,4]$	-0.0009 (0.0526)	-0.0058 (0.0465)	-0.0022 (0.0463)	-0.0019 (0.0477)	1.2582	1.2848	1.2130
	$U[0,3]$	0.0018 (0.0539)	-0.0017 (0.0460)	0.0027 (0.0466)	0.0008 (0.0489)	1.3722	1.3377	1.2196
250	$U[0,4]$	0.0016 (0.0324)	-0.0029 (0.0288)	0.0012 (0.0293)	0.0011 (0.0296)	1.2532	1.2281	1.1984
	$U[0,3]$	-0.0002 (0.0336)	-0.0033 (0.0287)	0.0011 (0.0300)	-0.0002 (0.0301)	1.3540	1.2482	1.2479
(s,t)=(0.5108,1.6094)								
n	C							
50	$U[0,4]$	0.0007(0.0988)	0.0018 (0.0870)	0.0064 (0.0880)	0.0008 (0.0919)	1.2866	1.2539	1.1541
	$U[0,3]$	0.0018 (0.1112)	0.0079 (0.9390)	0.0130 (0.0959)	-0.0013 (0.0987)	1.3920	1.3214	1.2682
100	$U[0,4]$	-0.003 (0.0709)	-0.0029 (0.0620)	0.0042 (0.0634)	-0.0030 (0.0649)	1.3101	1.2493	1.1931
	$U[0,3]$	0.0040 (0.0756)	0.0070 (0.0640)	0.0147 (0.0665)	0.0046 (0.0680)	1.3827	1.2369	1.2335
250	$U[0,4]$	$<$0.0001 (0.0458)	$<$0.0001 (0.0398)	0.0070 (0.0410)	0.0005 (0.0416)	1.3271	1.2141	1.2133
	$U[0,3]$	-0.0002 (0.0485)	0.0027 (0.0394)	0.0109 (0.0424)	0.0005 (0.0425)	1.5079	1.2224	1.2975
(s,t)=(0.9163,1.6094)								
n	C							
50	$U[0,4]$	0.0041 (0.1332)	-0.0013 (0.1201)	-0.0016 (0.1178)	-0.0006 (0.1199)	1.2304	1.2801	1.2347
	$U[0,3]$	-0.0016 (0.1422)	-0.0042 (0.1279)	-0.0095 (0.1273)	-0.0182 (0.1254)	1.2345	1.2428	1.2643
100	$U[0,4]$	-0.0043 (0.0848)	-0.011 (0.0760)	-0.0082 (0.0766)	-0.0075 (0.0783)	1.2226	1.2152	1.1629
	$U[0,3]$	-0.0020 (0.1014)	-0.0027 (0.0872)	-0.0002 (0.0893	-0.0069 (0.0882)	1.3507	1.2906	1.3134
250	$U[0,4]$	-0.0014 (0.0550)	-0.0070 (0.0493)	-0.0046 (0.0487)	-0.0032 (0.0503)	1.2230	1.2666	1.1933
	$U[0,3]$	-0.0025 (0.0586)	-0.0044 (0.0507)	-0.0011 (0.0530)	-0.0030 (0.0532)	1.3314	1.2272	1.2132
		$\widehat{p}^{\texttt{LM}}_{12}(s,t)$	$\widehat{p}^{\texttt{PLM}}_{12}(s,t)$	$\widehat{p}^{\texttt{NP}}_{12}(s,t)$	$\widehat{p}^{\texttt{GAM}}_{12}(s,t)$	LM/PLM	LM/NP	LM/GAM
(s,t)=(0.2231,0.5108)								
n	C							
50	$U[0,4]$	-0.0192 (0.2470)	-0.0218 (0.2452)	-0.0026 (0.2189)	0.0219 (0.1982)	1.0124	1.2812	1.5505
	$U[0,3]$	-0.0087 (0.2427)	-0.0163 (0.2404)	0.0217 (0.2124)	0.0559 (0.1830)	1.0159	1.2941	1.6226
100	$U[0,4]$	-0.0063 (0.1666)	-0.0112 (0.1655)	-0.0066 (0.1657)	-0.0006 (0.1592)	1.0096	1.0113	1.0968
	$U[0,3]$	-0.0018 (0.1636)	-0.0063 (0.1619)	-0.0016 (0.1594)	0.0086 (0.1503)	1.0186	1.0537	1.1809
250	$U[0,4]$	0.0046 (0.0975)	0.0025 (0.0971)	-0.0016 (0.0968)	0.0030 (0.0974)	1.0082	1.0164	1.0016
	$U[0,3]$	-0.0004 (0.1036)	-0.0004 (0.0998)	-0.0048 (0.0986)	-0.0007 (0.1012)	1.0783	1.1025	1.0477
(s,t)=(0.2231,0.9163)								
n	C							
50	$U[0,4]$	-0.0026 (0.2291)	-0.0082 (0.2250)	0.0124 (0.1918)	0.0401 (0.1590)	1.0357	1.4215	1.9630
	$U[0,3]$	0.0035 (0.2389)	-0.0137 (0.2458)	0.0475 (0.1903)	0.0579 (0.1558)	0.9416	1.4848	2.0796
100	$U[0,4]$	-0.0003 (0.1561)	-0.0013 (0.1513)	0.0032 (0.1526)	0.0109 (0.1412)	1.0645	1.0457	1.2155
	$U[0,3]$	-0.0005 (0.1624)	-0.0039 (0.1541)	-0.0016 (0.1567)	0.0084 (0.1437)	1.1097	1.0738	1.2728
250	$U[0,4]$	0.0026 (0.0988)	0.0029 (0.0965)	-0.0016 (0.0991)	0.0019 (0.0976)	1.0476	0.9941	1.0249
	$U[0,3]$	-0.0021 (0.0991)	-0.0031 (0.0958)	-0.0048 (0.0975)	-0.0038 (0.0970)	1.0693	1.0304	1.0420
(s,t)=(0.5108,1.6094)								
n	C							
50	$U[0,4]$	-0.0064 (0.1801)	-0.0120 (0.1686)	0.0020 (0.1517)	0.0212 (0.1378)	1.1366	1.4113	1.6733
	$U[0,3]$	-0.0182 (0.1910)	-0.0412 (0.1997)	0.0028 (0.1574)	0.0442 (0.1290)	0.8852	1.4862	1.9847
100	$U[0,4]$	0.0013 (0.1253)	-0.0032 (0.1179)	0.0012 (0.1188)	0.0004 (0.1174)	1.1280	1.1115	1.1394
	$U[0,3]$	-0.0041 (0.1332)	-0.0158 (0.1218)	-0.0050 (0.1217)	-0.0042 (0.1190)	1.1763	1.1979	1.2525
250	$U[0,4]$	0.0002 (0.0768)	-0.0019 (0.0731)	-0.0019 (0.0754)	-0.0019 (0.0739)	1.1047	1.0376	1.0800
	$U[0,3]$	-0.0018 (0.0845)	-0.0076 (0.0786)	-0.0012 (0.0786)	-0.0061 (0.0801)	1.1470	1.1558	1.1058
(s,t)=(0.9163,1.6094)								
n	C							
50	$U[0,4]$	0.0011 (0.1953)	-0.0020 (0.1908)	0.0076 (0.1784)	0.0326 (0.1529)	1.0485	1.1968	1.5640
	$U[0,3]$	-0.0002 (0.2146)	-0.0135 (0.2114)	0.0144 (0.1921)	0.1023 (0.1438)	1.0263	1.2414	1.4827
100	$U[0,4]$	0.0007 (0.1412)	-0.0001 (0.1360)	-0.0057 (0.1409)	0.0028 (0.1366)	1.0782	1.0026	1.0675
	$U[0,3]$	-0.0040 (0.1533)	-0.0110 (0.1462)	-0.0154 (0.1437)	0.0161 (0.1325)	1.0947	1.1277	1.3219
250	$U[0,4]$	-0.0017 (0.0856)	-0.0019 (0.0815)	-0.0095 (0.0819)	-0.0025 (0.0822)	1.1031	1.0788	1.0854
	$U[0,3]$	-0.0036 (0.0928)	-0.0045 (0.0858)	-0.0110 (0.0886)	-0.0044 (0.0891)	1.1696	1.0821	1.0846

TABLE 4.2: Bias and standard deviation (SD) for estimators of $p_{ij}(s,t)$. The relative MSEs are also given. Scenario 2: progressive three-state model.

	$\widehat{p}_{01}^{\mathtt{LM}}(s,t)$	$\widehat{p}_{01}^{\mathtt{PLM}}(s,t)$	$\widehat{p}_{01}^{\mathtt{NP}}(s,t)$	$\widehat{p}_{01}^{\mathtt{GAM}}(s,t)$	Relative MSE		
	bias (SD)	bias (SD)	bias (SD)	bias (SD)	`LM/PLM`	`LM/NP`	`LM/GAM`
(s,t)=(8,10)							
n							
100	-0.0004 (0.0190)	-0.0140 (0.0081)	-0.0127 (0.0088)	-0.0079 (0.0150)	1.3722	1.5101	1.2463
250	0.0004 (0.0123)	-0.0144 (0.0049)	-0.0137 (0.0054)	-0.0109 (0.0092)	0.6512	0.6962	0.7344
500	-0.0002 (0.0086)	-0.0147 (0.0034)	-0.0144 (0.0036)	-0.0115 (0.0065)	0.3334	0.3464	0.4382
(s,t)=(10,14)							
n							
100	0.0002 (0.0323)	0.0161 (0.0253)	0.0178 (0.0266)	0.0131 (0.0285)	1.1609	1.0168	1.0599
250	0.0001 (0.0204)	0.0175 (0.0158)	0.0108 (0.0181)	0.0038 (0.0185)	0.7506	0.9400	1.1755
500	0.0001 (0.0151)	0.0178 (0.0113)	0.0064 (0.0133)	0.0032 (0.0137)	0.5159	1.0502	1.1486
(s,t)=(12,18)							
n							
100	0.0010 (0.0761)	-0.0163 (0.0615)	-0.0082 (0.0601)	0.0273 (0.0617)	1.4298	1.5724	1.2729
250	0.0013 (0.0472)	-0.0154 (0.0380)	-0.0069 (0.0414)	-0.0025 (0.0437)	1.3246	1.2631	1.1653
500	0.0010 (0.0329)	-0.0154 (0.0262)	-0.0042 (0.0301)	-0.0017 (0.0305)	1.1702	1.1725	1.1626
(s,t)=(16,18)							
n							
100	0.0013 (0.0712)	0.0086 (0.0635)	0.0210 (0.0598)	0.0703 (0.0638)	1.2339	1.2644	05793
250	0.0007 (0.0423)	0.0109 (0.0376)	0.0083 (0.0365)	0.0153 (0.0378)	1.1682	1.2732	1.0733
500	0.0015 (0.0309)	0.0122 (0.0271)	0.0103 (0.0261)	0.0056 (0.0286)	1.0841	1.2152	1.1281
	$\widehat{p}_{12}^{\mathtt{LM}}(s,t)$	$\widehat{p}_{12}^{\mathtt{PLM}}(s,t)$	$\widehat{p}_{12}^{\mathtt{NP}}(s,t)$	$\widehat{p}_{12}^{\mathtt{GAM}}(s,t)$	`LM/PLM`	`LM/NP`	`LM/GAM`
(s,t)=(8,10)							
n							
100	0.0109 (0.3476)	0.0099 (0.3483)	-0.0091 (0.2273)	0.0245 (0.2458)	0.9961	2.3404	2.1752
250	0.0068 (0.2161)	0.0045 (0.2172)	-0.0008 (0.1872)	0.0144 (0.1687)	0.9905	1.3348	1.6334
500	0.0033 (0.1362)	-0.0007 (0.1383)	0.0027 (0.1368)	0.0023 (0.1346)	0.9700	0.9913	1.0240
(s,t)=(10,14)							
n							
100	-0.0060 (0.2852)	-0.0089 (0.2815)	-0.0096 (0.1909)	-0.0201 (0.0480)	1.0256	2.2333	4.1961
250	0.0030 (0.1852)	-0.0016 (0.1764)	0.0074 (0.1499)	0.0039 (0.1344)	1.1027	1.5240	1.9011
500	0.0029 (0.1272)	0.0017 (0.1211)	-0.0008 (0.1244)	0.0022 (0.1196)	1.1033	1.0462	1.1301
(s,t)=(12,18)							
n							
100	0.0141 (0.1786)	0.0077 (0.1792)	0.0149 (0.1339)	-0.2905 (0.1001)	0.9980	1.8120	0.3781
250	-0.0120 (0.1626)	-0.0247 (0.1488)	-0.0106 (0.1282)	0.0107 (0.0993)	1.1684	1.6103	2.7210
500	-0.0053 (0.1141)	-0.0135 (0.1090)	-0.0047 (0.0903)	-0.0006 (0.0873)	1.0823	1.6001	1.7157
(s,t)=(16,18)							
n							
100	0.0027 (0.1518)	-0.0034 (0.1423)	0.0220 (0.1137)	0.0019 (0.1042)	1.1382	1.7275	2.3236
250	0.0038 (0.0863)	-0.0020 (0.0808)	0.0182 (0.0831)	0.0053 (0.0813)	1.1413	1.0312	1.1256
500	0.0023 (0.0559)	-0.0016 (0.0475)	0.0147 (0.0476)	-0.0008 (0.0490)	1.3836	1.2611	1.3046

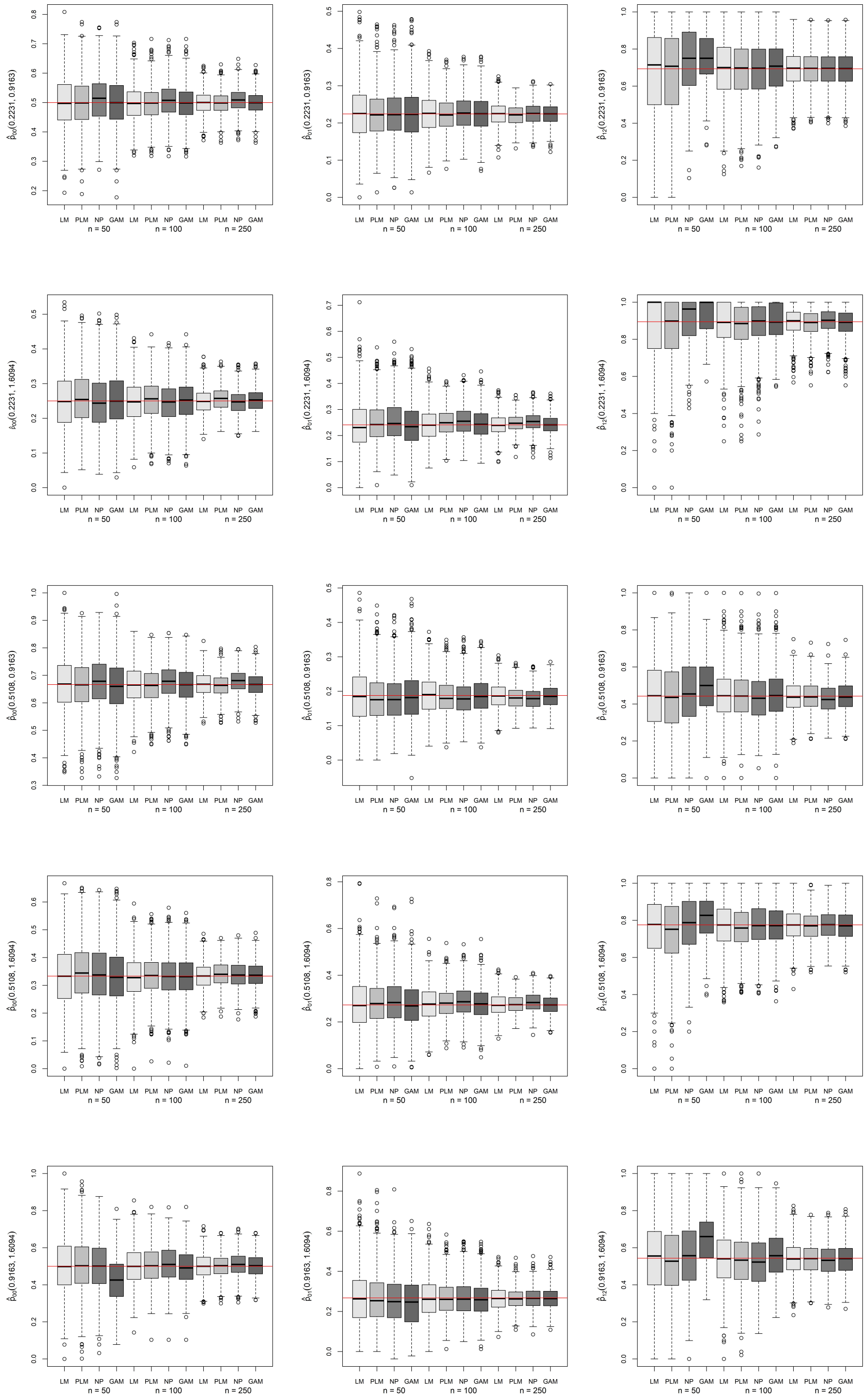


FIGURE 4.2: Boxplots of the $M = 1000$ estimates of the transition probabilities. Horizontal solid red line correspond to the true value of the transition probability. Censoring times uniformly distributed between 0 and 3.

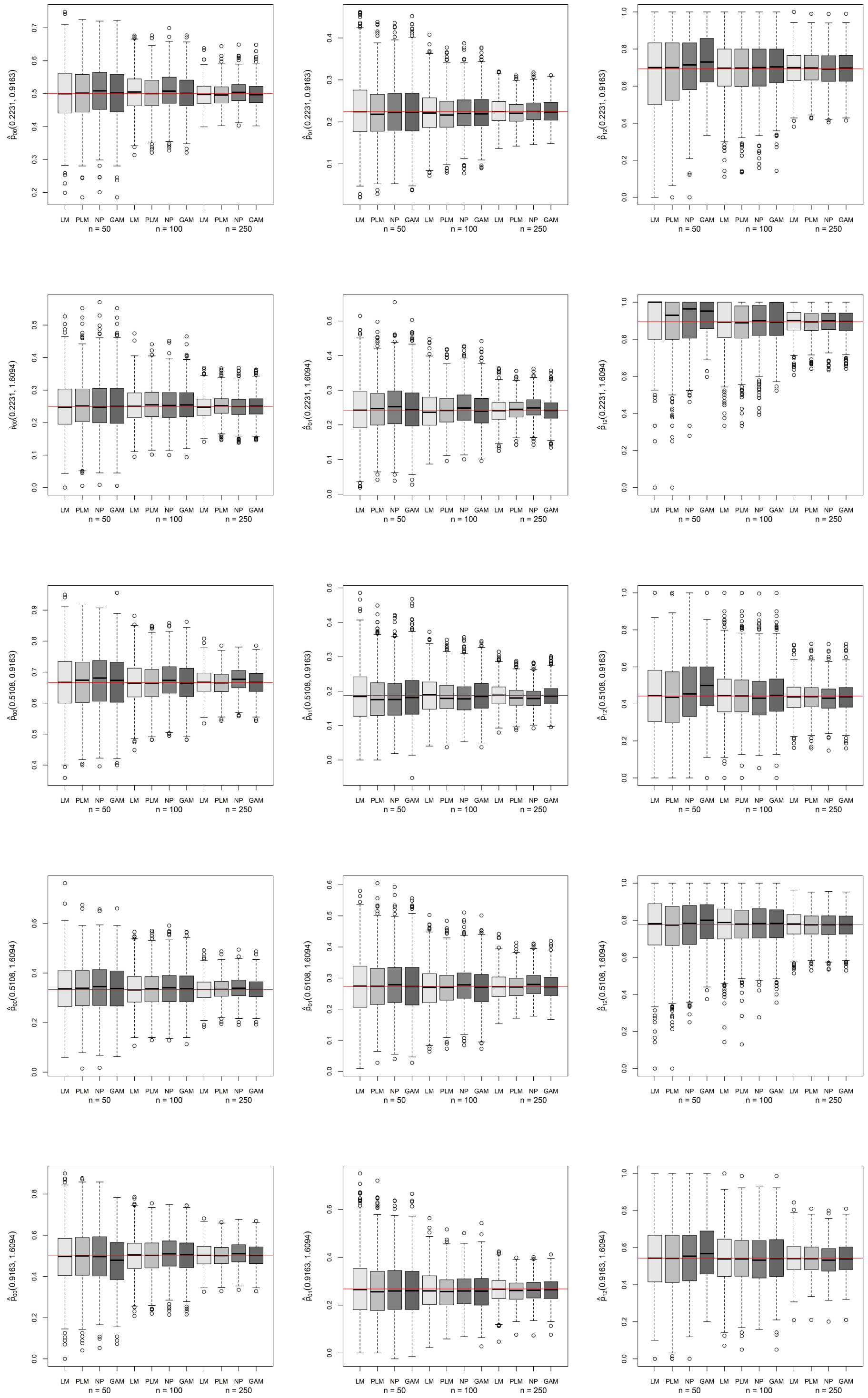


FIGURE 4.3: Boxplots of the $M = 1000$ estimates of the transition probabilities. Horizontal solid red line correspond to the true value of the transition probability. Censoring times uniformly distributed between 0 and 4.

4.4 Real data application

For illustration purposes, we apply the proposed methods of Section 4.2 to data based on a case study presented by the PROVA study group which aimed to evaluate the effect of propranolol and sclerotherapy on risk of first variceal bleeding and survival in patients with liver cirrhosis. The trial included 286 patients in whom cirrhosis has been diagnosed in the past and where endoscopy had shown oesophageal varices but who had not yet experienced a transfusion-requiring bleeding from the varices (PROVA (1991) [85]). From the 286 patients enrolled in the trial, 50 had a bleeding episode and among these 29 died. Forty six patients died without experienced variceal bleeding. The rest of the patients (190) remained alive and without variceal bleeding up to the end of the follow-up. The patients were observed for up to 42 months, with an average of 15 months. For each patient, the two event times (time to variceal bleeding and time to death) and the corresponding indicator statuses are recorded. The process can be modeled using a progressive illness-death model with transient states 'alive and without variceal bleeding', 'alive with variceal bleeding', and an absorbing state 'dead'.

The data set for the PROVA study group is of small size with a high censoring percentage, near to 74%. The time alive and without variceal bleeding, Z, is more often censored in the time interval between 1 and 4 years. We have only 11 individuals that experienced variceal bleeding one year after enrollment in the study. A few (15) complete (i.e. uncensored) observations for patients with a survival time greater than one year that died without experiencing variceal bleeding. At 90 days (3 months) we have 260 individuals alive and without variceal bleeding and only 11 individuals alive with variceal bleeding. From those 11 that experiencing variceal bleeding only 3 died leading to 73% of censored observations. Similarly, at 1 year after enrollment, we have 21 individuals alive with variceal bleeding and a few 5 complete observations (76% censoring). This means that it is expected that any estimator for the transition probabilities will behave poorly in these cases. Presmoothing will help to improve the estimation in these situations, in particular in the right tail of the distributions of Z and T (i.e., for larger times for being alive and without variceal bleeding, and also for larger values of the time since variceal bleeding, $T - Z$).

It is of practical interest to determine whether the Markov condition holds within a particular data set in order to determine which method is more appropriate to estimate the transition probabilities. The Markov condition may by checked by studying the effect of the time of variceal bleeding on the mortality after bleeding. Results from fitting this covariate in a Cox model reported a significant coefficient (p-value $= 0.0003$; regression coefficient: 0.0061) for the bleeding time revealing the failure of the Markov condition for the PROVA data set. These findings are in agreement with those obtained by Andersen and Esbjerj and Sorensen (2000) [20] and Andersen and Keiding (2002) [17].

In cases like this one, where there is strong evidence that the process does not verifies the Markov property, the landmark estimators can be preferable due to their greater accuracy. Since the construction of these estimators is based on subsamples of the complete data, this will generally lead to small sample sizes and usually to heavily censored data as shown in Figure 4.4. This fact is more evident for the individuals with variceal bleeding (observed in State 1) but it is also present for those in the alive and without bleeding state (State 0) for higher landmarking times. To avoid this problem, a valid approach is to consider a modification of the landmark estimator based on presmoothing. Here we present some figures to illustrate the estimators obtained by replacing the censoring indicator variables in the classical definitions by values of several regression estimators.

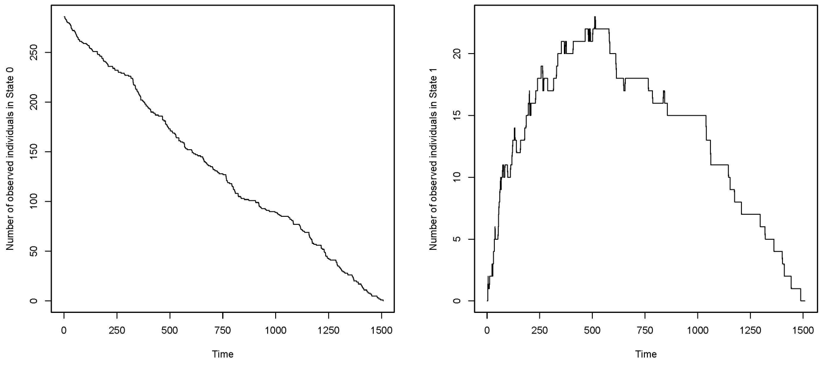


FIGURE 4.4: Number of individuals observed in State 0 and State 1 along time. PROVA data.

Figure 4.5 shows estimated curves of the transition probabilities $\widehat{p}_{hj}(s,t)$, for fixed values of $s = 90$ and $s = 365$ days along time t. The estimators labeled as 'Unsmoothed' correspond to the original unsmoothed landmark estimator proposed by de Uña-Álvarez and Meira-Machado (2015) [69] whereas the estimator labeled as 'logit' correspond to the semiparametric presmoothed estimator introduced in Section 4.2 with a preliminary presmoothing based on a parametric binomial logistic family. For completeness purposes we also included semiparametric presmoothed estimators based on a different binomial family such as 'probit' or 'cauchit'. We also considered estimators with a preliminary presmoothing function based on an additive logistic model (labeled in Figure 4.5 as 'logit.gam'). Finally, the estimator labeled as 'nonparametric' correspond to the presmoothed estimators with a preliminary presmoothing based on a nonparametric regression model using the Nadaraya-Watson kernel estimator. None of these methods require the process to be Markov.

As expected, all estimators report roughly the same estimates for lower values of t whereas some deviations are observed for higher values of t. It can also be seen that the original unsmoothed landmark estimator (de Uña-Álvarez and Meira-Machado (2015) [69]) reveals higher variability on the right hand side. Plots for the transition probabilities $\widehat{p}_{00}(s,t)$ (first row) and $\widehat{p}_{11}(s,t)$ (last row), $s = 90$ and 365, report the survival fraction along time, among the individuals 'alive and without variceal bleeding' and among those with a 'bleeding episode'. Results for the estimated curves for $\widehat{p}_{00}(s,t)$, $s = 90$ and 365, reveal also that the choice of the family function for the regression estimator can lead to major differences in the estimates. Some differences can also be observed in the right tail of the estimated transition probabilities $\widehat{p}_{01}(s,t)$ (second row) and $\widehat{p}_{02}(s,t)$ (third row). Plots for $\widehat{p}_{02}(s,t)$, report one minus the the survival fraction along time, among the individuals alive and without variceal bleeding at time s. Plots for $\widehat{p}_{01}(s,t)$, allow for an inspection along time of the probability of being 'alive with bleeding episode' for the individuals who are without variceal bleeding at time s. Since the 'bleeding' state is transient, these curves are first increasing and then decreasing. Major differences are observed in the right tail when comparing the methods based on a parametric presmoothing with their counterparts ('unsmoothed' and 'nonparametric') for the estimation of the transition probabilities. This can be possibly explained by a misspecification of the parametric model. The logistic model is likely the most common statistical model that is usually taken to apply to a binary dependent variable. The validity of a logistic model for the presmoothing function can be checked by applying a goodness-of-fit test such as the test proposed by Hosmer and Lemeshow (2003) [82]. Results obtained for all landmarking times (i.e., s) are depicted in Figure 4.6, revealed that the test was able to reject the logistic model when used to estimate the transition probabilities $\widehat{p}_{0j}(90,t)$. Note that the choice of this parametric model is a common choice for a parametric presmoothing. However, curves depicted in Figure 4.5 reveal that the presmoothed landmark estimators with a preliminary presmoothing based on a nonparametric regression provides in all cases curves with the expected behavior, similar to those obtained from the nonparametric original landmark estimators (labeled as 'unsmoothed') but with a better description of the tail behavior. Since there is a clear risk of a misspecification of the parametric model, the use of estimators based on nonparametric presmoothing as those labeled with 'nonparametric' are preferable.

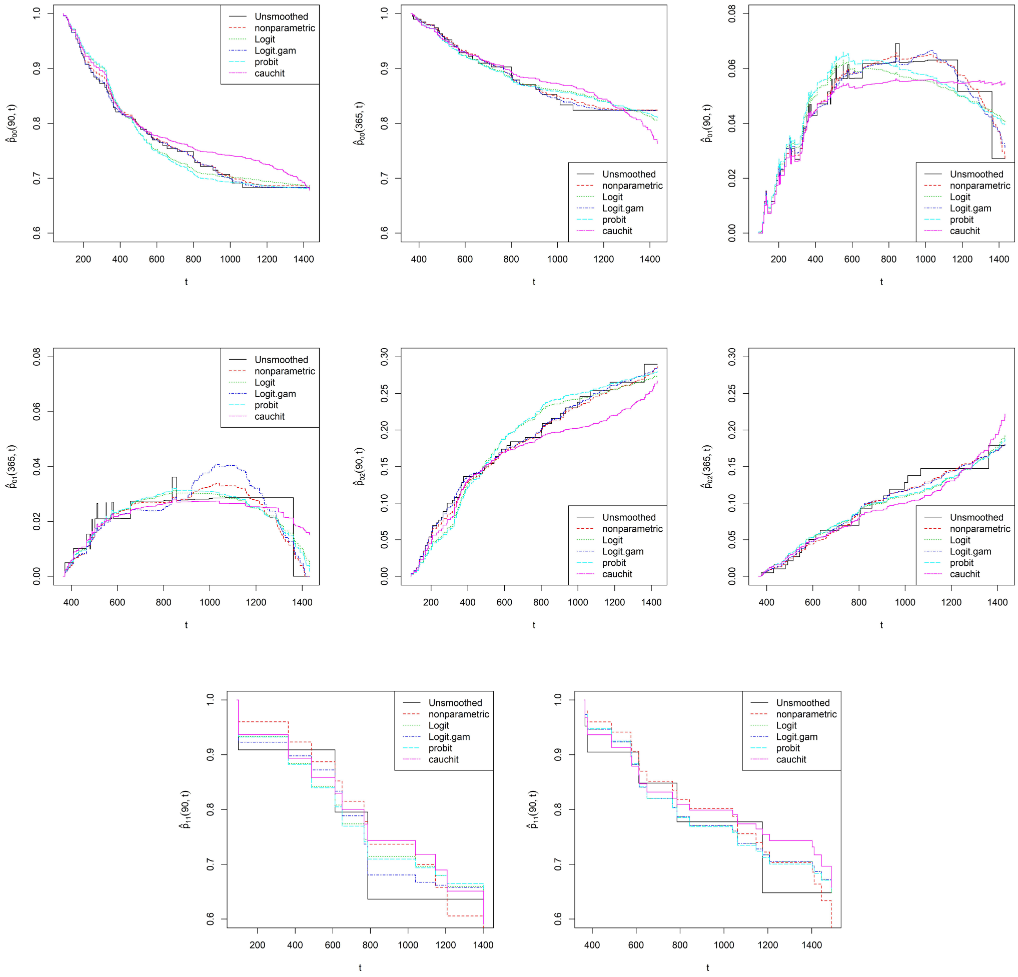


FIGURE 4.5: Estimated transition probabilities for $s = 90$ and $s = 365$. PROVA data.

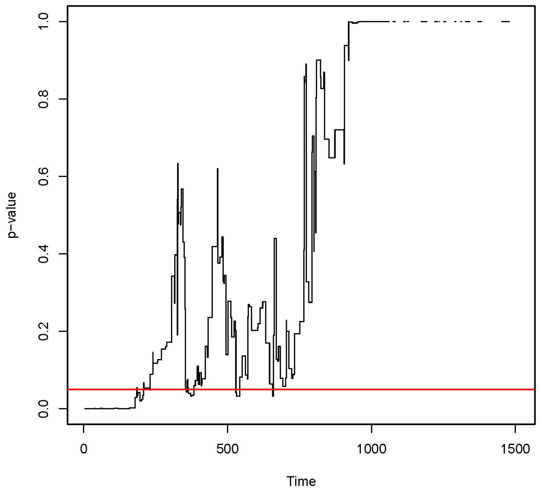


FIGURE 4.6: p-values for the Hosmer and Lemeshow test for logistic regression model for all possible landmarking times. PROVA data.

4.5 Software development

To provide the biomedical researchers with a comprehensive and easy-to-use tool for obtaining estimates of the transition probabilities, we developed an R package (R Core Team (2018) [86]) called presmTP (Soutinho, Meira-Machado and Oliveira (2019) [87]). The main function presmTP can be used to calculate the estimated transition probabilities from the unsmoothed landmark estimators but presmoothed estimates can also be obtained through the use of a parametric family of binary regression curves, such as logit, probit or cauchit. The additive logistic regression model and nonparametric regression are also alternatives implemented in function presmTP. The package dependencies include the survPresmooth package (López-de Ullibarri and Jácome (2013) [88]) and the mgcv package (Wood (2019) [84]). presmTP package (version 1.1.0) is available at the CRAN repository at https://cran.r-project.org/web/packages/presmTP. As a supplementary material, in Section A.2, we present the functions that comprise the R package

4.6 Discussion

Several recent contributions for the estimation of the transition probabilities in the context of multi-state processes that do not satisfy the Markov property have been reported. Many of these contributions were made for the illness-death model. Recently, the problem of estimating the transition probabilities in illness-death model has been reviewed, and new (landmark) estimators have been proposed which are built by considering specific subsets of individuals. As a weakness, the proposed estimators may provide large standard errors in estimation.

In this chapter, the relative performance of several presmoothed landmark estimators for the transition probabilities was investigated through simulations. Results suggest that the use of presmoothing techniques can lead to competitive estimators that may outperform the original landmark estimators providing estimators with less variability. It is worth mentioning that presmoothing can introduce some bias in estimation while reducing the variance. This bias component is larger when there is some misspecification in the chosen parametric model. The risk of introducing a large bias through a misspecified model can be controlled in practice by testing the validity of the model applying a goodness-of-fit test. Nonparametric presmoothing is useful when there is no parametric candidate for the presmoothing function.

Chapter 5

Joint models of longitudinal and multistate survival data: estimation of the conditional transition probabilities

The topic of joint modeling of longitudinal and survival data has received remarkable attention in recent years. In cancer studies for example, these models can be used to assess the impact that a longitudinal marker has on the time to death or relapse. The goal of this chapter is to introduce feasible estimation methods for the transition probabilities conditionally on covariates observed with repeated measures combining existing methods for joint modeling of longitudinal and survival data with the landmark landmark approach. This way, it is possible to extend to time-dependent covariates for each individual so that the trajectory of the longitudinal outcomes are included in the regression models. From the application to a data set we have confirmed that the proposed joint modeling landmark estimators have good small sample properties and are more efficient than competing estimators that do not take in consideration all information provided by the longitudinal measures of the covariate.

This chapter is partially based on Soutinho, Meira-Machado and Oliveira (2020) [89] published in *35th International Workshop on Statistical Modelling (IWSM2020)*.

5.1 Introduction

In medical studies, besides studying time-to-event, one main goal is to evaluate the impact of a set of repeated measures as time-dependent covariates on the transition intensities. These longitudinal measures involve collection of data at different time points for each study subject which are characterized by the dependence within subject repeated observations over time (Xu and Bai (2017) [90]). Some examples are the use of repeated measures of lung function in clinical studies on chronic obstructive pulmonary disease (COPD); blood tests performed for patients enrolled in drug trials or to evaluate the prostate specific antigen (PSA) (Edwards (2000) [91]; Molenberghs and Verbeke (2001) [92]; and Hickey *et al.* (2018) [93]). In order to produce valid inferences in these cases a joint modeling analysis of longitudinal and survival outcomes is required (Rizopoulos (2010) [94]).

A common strategy, that greatly simplifies the inference, is to model both processes separately, using linear mixed effects models for the longitudinal model and Cox regression models for the time-to-event analysis (Cox (1972) [6]). However, this approach is not recommended, in practice, since this will generally lead to biased effect size estimates due to the correlation between the two outcomes (Ibrahim, Chu and Chen (2010) [95]; Hickey *et al.* (2018) [93]). Joint modeling approach still remains a matter of interesting works and there are many contributions in literature covering the topic such as the works by Self and Pawitan (1992) [96] and Tsiatis *et al.* (1995) [97], addressing the joint modeling on the CD4 biomarker to the development of AIDS, or the many contributions given by Rizopoulos (2010 and 2012) [94, 98].

Less has been done, however, in the framework of joint modeling of longitudinal and multi-state models. We can stand out some of the following: Williamson *et al.* (2008) [99] extended some methods for joint modeling with only one time-to-event to allow for competing risks. Liu and Huang (2009) [100] applied a joint random effects model to data of a clinical research on AIDS to study the interaction of the CD4 cell repeated measures and recurrent opportunistic diseases and their effect on survival. Andrinopoulou *et al.* (2014) [101] considered a joint model for two echocardiographic markers with competing risk failure time data to assess the valve functions during the follow-up period. Later, the same data set was used by Andrinopoulou *et al.* (2017) [102] to investigate the effect of the features of the longitudinal process on the prediction for the events regarding the time-dependent trajectory and time-dependent cumulative effects. Ferrer *et al.* (2016)

[56] proposed a joint model for a longitudinal process and a multi-state process which is divided into two sub-models linked by a function of shared random effects. Contributions have also been done in the form of software (Król *et al.* 2017) [103]. Background concepts related to the extension of the joint modeling to multi-state models can be found in Ferrer *et al.* (2016) [56] and Hickey *et al.* (2018) [104].

There has been little research on the estimation of the transition probabilities conditional on current or past covariate measures. An averaged Beran's conditional estimator was introduced by Dabrowska and Lee (1996) [105] to yield a consistent estimate of the transition probabilities. The authors considered a vector of sojourn times in past states as the covariate. Two competing nonparametric estimators for the conditional transition probabilities were introduced by de Uña-Álvarez and Cadarso-Suárez (2006) [68]. The proposed methods are based on local smoothing by means of kernel weights based on local constant (Nadaraya-Watson) regression. Two different schemes of inverse censoring probability reweighting were used to deal with right censoring. Their proposals are fully nonparametric approaches which are not recommended to multiple covariates due to the problem in multivariate nonparametric regression estimation, the so-called curse of dimensionality. One standard method that is particularly well-suited to the setting with multiple covariates is to consider estimators based on a Cox's regression model (Cox (1972) [6]) fitted marginally to each allowed transition, with the corresponding baseline hazard function estimated by the Breslow's method (Breslow, 1972) [23]. A direct binomial regression technique was considered by Azarang *et al.* (2017) [106].

In this chapter, we revisit the problem of estimation of the transition probabilities of an irreversible, possibly non-Markov model. However, unlike the previous contributions, we are interested in estimating these probabilities given a continuous covariate measured repeatedly over time. To this end, we used the subsampling approach proposed by de Uña-Álvarez and Meira-Machado (2015) [69] combined with methods proposed by Rizopoulos (2012) [98]. The joint modeling approach allows to extend to time-dependent covariates for each individual so that the trajectory of the repeated longitudinal outcomes are taken into account. This approach also deals quite well with imbalance data where the number of measures by individual or the moments at which they were collected are not the same (Rizopoulos (2012) [98]). Standard methods in this setup usually rely on a parametric specification of the covariate effects or do not have in consideration for all the longitudinal outcomes. As an example, recently Hoff *et al.* (2019) [107] present semiparametric regression models and inverse probability weights in combination with the LMAJ estimator to perform covariate adjusted analyses. Nevertheless, in the estimation of the transition probabilities repeated measures are not included.

The rest of the chapter is organized as follows: the joint modeling approach is introduced in Section 5.2. Section 5.3 describes the application of the proposed method to a data set. Finally the main conclusions are given in Section 5.4.

5.2 Methods

5.2.1 Notation and preliminaries

Let's consider the irreversible version of the illness-death model represented by individuals who start in the 'alive and disease-free' state (State 0, in this chapter) and may move to the 'diseased' state (State 1) or to the absorbing 'dead' state (State 2). Individuals in the 'diseased' state will eventually move to the 'dead' state without any possibility of recovery.

The underlying stochastic process of the progressive illness-death model may be represented by $\{X(t), t \geq 0, X(0) = 0\}$, where $X(t)$ represents the state occupied by the process at time t, for which we assume all individuals to be in State 0 at time 0. The process is characterized by the joint distribution of (Z, T), where $Z = \inf\{t : X(t) \neq 0\}$ denotes the sojourn time in State 0 and $T = \inf\{t : X(t) = 2\}$ is the total survival time of the process. The right-censoring is modelled by considering a censoring variable C, which we assume to be independent of the process. Due to censoring, rather than (Z, T) we observe $\widetilde{Z} = \min(Z, C)$ and $\widetilde{T} = \min(T, C)$, the censored versions of Z and T, and we have $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$ for the respective censoring indicators of Z and T. In this chapter, we also consider a time-dependent covariate which we denote by $Y(t)$ and a vector of baseline covariates denoted by W. Finally, the available data are $(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i, Y_i(t), W_i)$, $1 \leq i \leq n$, i.i.d. copies of $(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta, Y(t), W)$.

5.2.2 Joint model specification

In this section, we present the details of the joint modeling framework under the multi-state point of view. For simplicity and without loss of generality, consider the conditional transition probability $p^i_{11}(s, t \mid Y_i(t), W_i)$ where $Y_i(t)$ denotes the longitudinal covariate and W_i a vector of baseline covariates for the i-th individual who belong at time s to the intermediate state (State 1, in this chapter).

Assuming the landmark approach, we will consider a subsample of the data consisting on those individuals observed at State 1 at time s. The joint modeling approach for multi-state models is built assuming two submodels: a linear mixed effects submodel and a illness-death submodel with three transition intensities. To estimate the conditional transition probability $p^i_{11}(s, t \mid Y_i(t), W_i)$ we only need to consider the transition intensity from the intermediate State 1 ('diseaded') to State 2 ('dead'). Let $Y_i(t)$ denote the value of the longitudinal outcome at time t for the i-th individual. The observed longitudinal data consists of the measurements $Y_{ij} = \{Y_i(t_{ij}), j = 1, \ldots, n_i\}$, in which n_i correspond to the number of repeated measures for the i-th individual. The true and unobserved value of the longitudinal outcome at time t is denoted by $m_i(t)$. We aim to estimate $P(T > t \mid Z \leq s, T > s, \mathcal{M}_i(s), W_i)$, where $\mathcal{M}_i(s) = \{m_i(u), 0 \leq u < s\}$ denotes the history of the true unobserved longitudinal process up to time point s.

Under Gaussian assumptions, we have that the unobserved true longitudinal outcome $m_i(t)$ is explained according to time and covariates with fixed and random effects that take into consideration the correlation between the repeated measures of the same individual. In particular, we have the following mixed effects model:

$$\begin{aligned} Y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= U_i^T(t)\beta + V_i^T(t) b_i + \varepsilon_i(t), \end{aligned} \tag{5.1}$$

where $U_i(t)$ and $V_i(t)$ denote the vectors of the design matrices for the fixed and random effects, β and b the corresponding parameters, and $\varepsilon_i(t) \sim N(0, \sigma^2)$ is the measurement error term which is assumed to be independent of b_i, $b_i \sim N(0, D)$.

To model the transition intensities depending on the effect of $m_i(t)$, we consider the following proportional hazards model:

$$\begin{aligned} \lambda^i_{hj}\left(t|\mathcal{M}_i(t), W_i\right) &= \lim_{dt \to 0} P\left(X_i(t + dt) = j \mid X_i(t) = h, \mathcal{M}_i(t), W_i\right) \\ &= \lambda_{hj,0}\left(t\right) \exp\left\{\alpha_{hj} m_i(t) + \gamma^T_{hj} W_i\right\}, \end{aligned} \tag{5.2}$$

where $\lambda_{hj,0}(t)$ is a baseline hazard function, γ_{hj} is a vector of regression coefficients and the parameter α_{hj} quantifies the effect of the longitudinal outcome on the time-to-event for the transition from h to j.

Without loss of generality, and for the purpose of simplicity, we focus now our attention in the progressive illness-death model, in particular, on the individuals belong to intermediate state 1 given by the subset $\mathcal{S}_1$. In this case, the maximum likelihood estimation is based on the maximization of the log-likelihood of the joint distribution of the time-to-event (given by the total survival time) and longitudinal outcomes $(T_i, \Delta_i, Y_i(t))$. Thus, regarding the vector of time-independent random effects and the correlation between the repeated measurements in the longitudinal process (conditional independence), we formally have (Rizopoulos (2010) [94])

$$\begin{aligned} p\left(T_i, \Delta_i, Y_i | b_i; \theta\right) &= p\left(T_i, \Delta_i | b_i; \theta\right) p\left(Y_i | b_i; \theta\right) && (5.3)\\ p\left(Y_i | b_i; \theta\right) &= \prod_j p\left\{Y_i\left(t_{ij}\right) | b_i; \theta\right\} && (5.4)\end{aligned}$$

where $\theta = \left(\theta_t^T, \theta_y^T, \theta_b^T\right)^T$ denotes the parameter vector, with θ_t representing the parameters for the event time outcome, θ_y the parameters for the longitudinal outcomes and θ_b the unique parameters of the random-effects covariance matrix, y_i is the $n_i \times 1$ vector of longitudinal responses of the i-th subject, and $p(\cdot)$ denotes an appropriate probability density function. In particular, $p\left\{Y_i\left(t_{ij}\right) | b_i; \theta_y\right\}$ is the univariate normal density for the longitudinal responses.

Under the assumptions for joint models, the log-likelihood contribution for the i-th subject can be formulated as follows

$$\log p\left(T_i, \Delta_i, Y_i; \theta\right) = \log \int p\left(T_i, \Delta_i | b_i; \theta_t, \beta_t\right) \left[\prod_j p\left\{Y_i\left(t_{ij}\right) | b_i; \theta_y\right\}\right] p\left(b_i; \theta_b\right) db_i \quad (5.5)$$

where $p\left(b_i; \theta_b\right)$ is the multivariate normal density for the random effects and the likelihood of the survival part is given by

$$p\left(T_i, \Delta_i | b_i; \theta_t, \beta_t\right) = \left\{\lambda_i\left(T_i | \mathcal{M}_i(t); \theta_t, \beta\right)\right\}^{\Delta_i} S_i\left(T_i | \mathcal{M}_i(t); \theta_t, \beta\right), \quad (5.6)$$

with $\lambda_i(\cdot)$ given by (5.2), and

$$\begin{aligned} S_i\left(t | \mathcal{M}_i(t); W_i, \theta_t, \beta\right) &= P\left(T_i > t | \mathcal{M}_i(t); W_i, \theta_t, \beta\right)\\ &= \exp\left\{-\int_0^t \lambda_i\left(s | \mathcal{M}_i(s); \theta_t, \beta\right) ds\right\} && (5.7)\end{aligned}$$

It should be noted that the maximization of the log-likelihood function given by (5.5) with respect to θ is a computationally challenging task. In this work, we used the Expectation-Maximization (`EM`) algorithm to maximaze the approximated log-likelihood, which is a traditional method that deals with the random effects as being 'missing data'. See Ferrer *et al.* (2016) [56] for more details on the generalization of the likelihood of the joint distribution for multi-state models.

In this chapter, we are actually interested in estimating the following conditional transition probabilities $p^i_{00}(s,t|Y_i)$, $p^i_{02}(s,t|Y_i)$ and $p^i_{11}(s,t|Y_i)$. Baseline covariates are not considered. Assuming the landmarking approach, we are interested in the prediction of the conditional transition probabilities for a new subject, i, with a set of longitudinal measurements $\mathcal{Y}_i(s) = \{Y_i(s), 0 \leq u \leq s\}$. These quantities can be expressed through the following conditional probabilities:

$$\begin{aligned} p^i_{00}(s,t|Y_i(t)) &= P(Z > t | Z > s, \mathcal{Y}_i(s), \mathcal{S}_0) \\ p^i_{02}(s,t|Y_i(t)) &= P(T \leq t | Z > s, \mathcal{Y}_i(s), \mathcal{S}_0) \\ p^i_{11}(s,t|Y_i(t)) &= P(T > t | Z \leq s, T > s, \mathcal{Y}_i(s), \mathcal{S}_1) \end{aligned} \tag{5.8}$$

where $t > s$, $\mathcal{Y}_i(s) = \{Y_i(s), 0 \leq u \leq s\}$, $\mathcal{S}_0$ and $\mathcal{S}_1$ denote the landmark samples on which the joint model was fitted. To this end, we propose to estimate these conditional probabilities by predicting survival probabilities for a new subject i taking the advantage of the landmark approach that reduces the estimation of these quantities to the estimation of survival functions, following the same procedures given by Rizopoulos (2010), for calculating expected survivals.

To be specific, let's consider the conditional probability $p^i_{11}(s,t|Y_i(t))$ for the individuals who remain in the intermediate state (State 1). In this particular case, with the necessary adaptations in terms of notation, this transition probability can be reduce to the following expression

$$p^i_{11}(s,t|Y_i(t)) = \int \frac{S_i\left\{t|\mathcal{M}_i(t,b_i,\theta);\theta\right\}}{S_i\left\{s|\mathcal{M}_i(s,b_i,\theta);\theta\right\}} p\left(b_i|Z \leq s, T > s, \mathcal{Y}_i(s);\theta\right) db_i \tag{5.9}$$

where $S_i(\cdot)$, denotes the survival function, and furthermore we explicitly note that the longitudinal history $\mathcal{M}_i(\cdot)$, as approximated by the linear mixed-effects model, is a function of both the random effects and the parameters. Finally, to estimate the transition probability $\widehat{p}_{11}^{i}(s,t|Y_i(t))$, we can derive a Monte Carlo, based on Bayesian formulation of the problem, using the simulation scheme described in Rizopolous (2010, 2012) [94][98].

5.3 Application to a data set

In this section, we investigate the performance of the proposed estimators of the conditional transition probabilities through two data sets that contain longitudinal and multi-state data in order to estimate the joint multi-state model described in the simulation study (see Section 4 of Ferrer *et al.* (2016) [56]). The data can be download at `https://github.com/LoicFerrer/JMstateModel/tree/master/Example`.

More specifically, we aim to compare the proposed estimator based on a joint modeling landmark approach (labeled as `JMLM`) with the estimators based on a Cox's model fitted marginally to the corresponding transition with the baseline hazard function estimated by the Breslow's method (labeled as `BRES`). For completeness purposes, we also included the original landmark estimator, which does not account for the effect of the covariate (labeled as `LM`). Note that the estimator based on a Cox model does not take into account the complete longitudinal history of the covariate up to time s, but the value of the longitudinal covariate at time s. It assumes that a patient's risk can be fully explained by the current covariate value. This simplifying assumption is commonly made in medical literature. Note also that the proposed `JMLM` estimator not only accounts for the complete longitudinal history of the covariate up to time s but also for the two event times. This means that two individuals with the same longitudinal history can have different `JMLM` estimates of the conditional transition probabilities.

The data set comprises a total of 500 individuals. 164 of them have developed an illness (State 1, in this chapter) and among these 99 died (State 2). 157 died without having the illness, passing directly from the initial state to the absorbing state. The rest of the subjects (179) remained disease-free (State 0) up to the end of the follow-up. Thus, it was observed nearly 36% of censored observations in State 1 and 40% in State 2. Most of the values of the longitudinal marker vary from -0.681 to 1.217 (interquartile range), with a minimum of -12.753 and a maximum of 7.985.

In order to compare the behavior of the three estimators considered in this section, we evaluate the conditional transition probabilities for a fixed value of $s = 8$. This value was used to built the two landmark data sets that are based on the individuals occupying State 0 ($\mathcal{S}_0 = \{i : \widetilde{Z}_i > 8\}$) and State 1 ($\mathcal{S}_1 = \{i : \widetilde{Z}_i \leq 8, \widetilde{T}_i > 8\}$). The first subsample of data will be used to estimate the conditional transition probabilities $p^i_{00}(8, t|Y_i(t))$ whereas the second subsample is used to estimate $p^i_{11}(8, t|Y_i(t))$.

Figure 5.1 shows two plots to summarize the progression of the longitudinal marker, for the individuals observed at time $s = 8$ in the initial state (State 0, subset $\mathcal{S}_0$) (first row, left column) and for those that belong to the subsample of individuals in State 1 (first row, right column). A total of 151 subjects were observed in State 0, at time $s = 8$, while this number decreases to 96 that are observed in State 1 at the same time value. The grey lines represent the values of the longitudinal marker along time for each individual, against the line with red color that indicates the average trend of the progression. For both subsets, the red line suggests, on average, that the marker progression decreases at a fast rate up to nearly $s = 1$, after entry in study. Individuals in State 0 at time $s = 8$ (i.e. that belong to $\mathcal{S}_0$) show, on average, an initial decrease in the marker values until a value near 2 and then is roughly constant. On the other hand, individuals in State 1 at time $s = 8$ (i.e. that belong to $\mathcal{S}_1$) show, on average, a fast decreasing rate for lower values and then a fast increase until time 8. Almost all the individuals that belong to $\mathcal{S}_1$ have an increase marker progression whereas the evolution in $\mathcal{S}_0$ is much more heterogenous as we can see by its wide band of marker values for higher time values near 8.

To illustrate the usefulness of the proposed methods, we have considered four subjects with different progression of the longitudinal marker for each subsamples (S_0 and S_1). Their longitudinal values can be seen in the second row of Figure 5.1. These subjects were chosen in order to analyze the ability of the estimators to deal with different trends of the longitudinal variable. For the first subsample, $\mathcal{S}_0$, we selected the individuals with the following numbers and trends: subject 228 with a high decreasing rate of the marker progression; subject 181 with an almost constant trend of the marker with values close to 0; subject 192 with a bathtub shape progression, first decreasing and then increasing; and subject 421 with an increasing rate (Figure 5.1, left column of the second row). We experience some difficulties in finding subjects with a decreasing rate in the longitudinal marker. To this second subset, $\mathcal{S}_1$, we have chosen the subject 56 with a small decreasing rate (with a smaller number of longitudinal measures when compared to the others three subjects); subjects 20 and 74 with relatively slow increasing rates, almost constant, after the first moments since entry in study; and, finally, subject 8 with a fast increasing rate and high marker values for all times (Figure 5.1, right column of the second row).

Figure 5.2 shows the estimated curves of the conditional transition probabilities $\hat{p}^i_{00}(8, t \mid$

$Y_i(t)$), for all subjects observed in State 0 at time $s = 8$, using the proposed JMLM estimator (left column of the first row). The BRES estimator is shown on the right column of the first row. For completeness purposes, the unconditional landmark estimator LM (black dash line) and the curve that represents the global average for all the subjects (red line) are also displayed in both plots. These plots reveal higher variability of estimated curves based on the BRES estimator when compared with those provided through the JMLM estimator. With regard to the progression of the longitudinal marker in time, it can be seen that the estimated curves of the BRES and LM estimators are quite similar whereas the curves of the JMLM estimator present higher values for time values near $s = 8$. The less variability of the JMLM estimator is more evident for lower values of t, revealing narrower bands, given by the curves, that become wider with greater lag times $t - s$, in agree of the average progression of the marker. The plots of the estimated curves of the selected subjects are displayed on the second row. The estimated curves based on the JMLM estimator (left column) and the BRES estimator (right column), revealed that a decreasing progression of the longitudinal marker leads to higher values of the estimated survival curves as occurs with subject 228, for both JMLM and BRES estimators (black line). On the other hand, an increasing rate of the evolution of the marker leads to lower values of the estimated curve that become more evident as the longitudinal markers are higher. Such behavior can be seen when observing the estimated curve for subject 421 for both JMLM and BRES estimators (blue line). However, the analysis of the estimated curves of subjects 192 and 181 revealed interesting differences between the two estimators that allow to highlight the importance of the proposed method (JMLM) by considering the trend of the longitudinal marker in the estimation. In fact, since subject 192 has a higher increasing rate than subject 181, as a consequence, for both estimators, the survival curve of the subject 192 (green line) is closer to the blue line (associated to subject 421) than those given by subject 181 (red line). This can be explained by the fact that the BRES estimator is takes only into account the last value of the marker before time 8, and these values are quite similar for subjects 421 and 192 (around 2). This can explain the similarity of the two curves (with blue and green lines). Such behavior is not present when the curves are estimated through the JMLM because it considers the evolution of the longitudinal process, explaining why the curve for subject 192 (green line) is now closer of that with a red line (of subject 181) than the curve of subject 421 (blue line).

Similar plots for the estimated curves of the conditional transition probabilities $\hat{p}^i_{11}(8, t \mid$

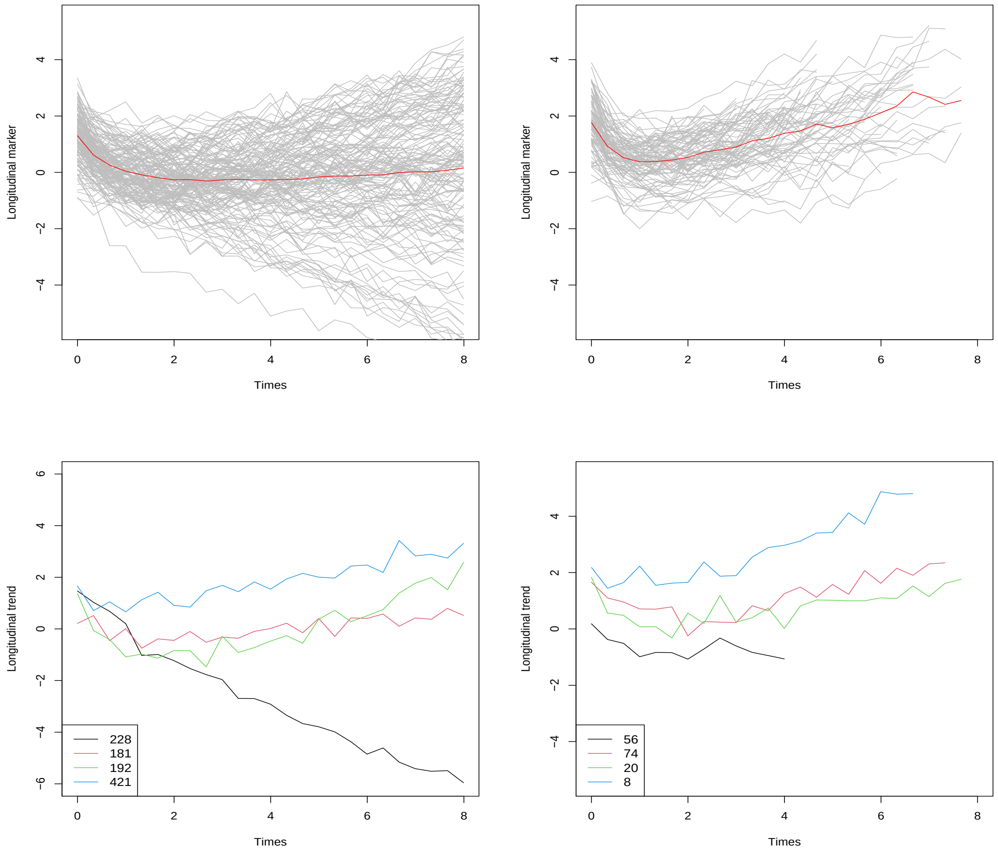


FIGURE 5.1: Spaghetti plot for marker values for the subsets $\mathcal{S}_0$ and $\mathcal{S}_1$ (first line) and the marker progression of selected individuals belong to the same $\mathcal{S}_0$ and $\mathcal{S}_1$ subsets (second line).

Y_i) are shown in Figure 5.3. Again, it can be seen higher variability among the estimated curves based on the BRES method when compared with those provided through the JMLM estimator. Interestingly, the variability seems smaller when compared to that shown in Figure 5.2, being closer to the unconditional landmark estimator, labeled as LM, with some discrepancies for lower values of t. Similarly, as in Figure 5.2, subjects with a higher increasing rates for the longitudinal values reveal estimated survival curves that decrease faster for both JMLM and BRES estimators (see subject 8, represented with a blue line). On the other hand, subjects with a decreasing rates for the longitudinal variable, have higher values of the estimated curves as can be seen when observing subject 56 (represented by a black line). Finally, subjects 74 and 20, with intermediate marker progressions with slow increasing rates (almost constant), attained estimated curves that are close to the curve obtained through the LM estimator.

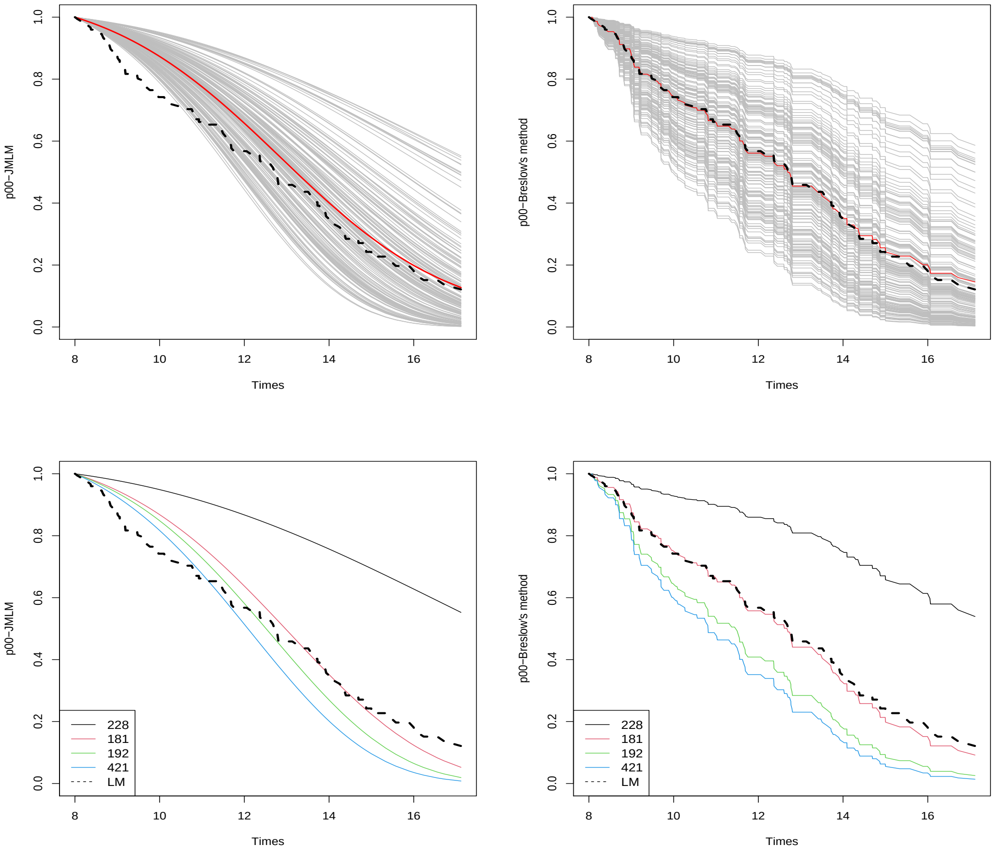


FIGURE 5.2: Spaghetti plot for transition probabilities $\hat{p}^{i}_{00}(8,t \mid Y_i)$, for all individuals belong to the subset $\mathcal{S}_0$, using the proposed JMLM estimator (First row, left column) and the BRES estimator (First row, right column). Transition probabilities $\hat{p}^{i}_{00}(8,t \mid Y_i)$ using the JMLM estimator (Second row, left column) and the BRES estimator (Second row, right column) for selected individuals belong to the subset $\mathcal{S}_0$.

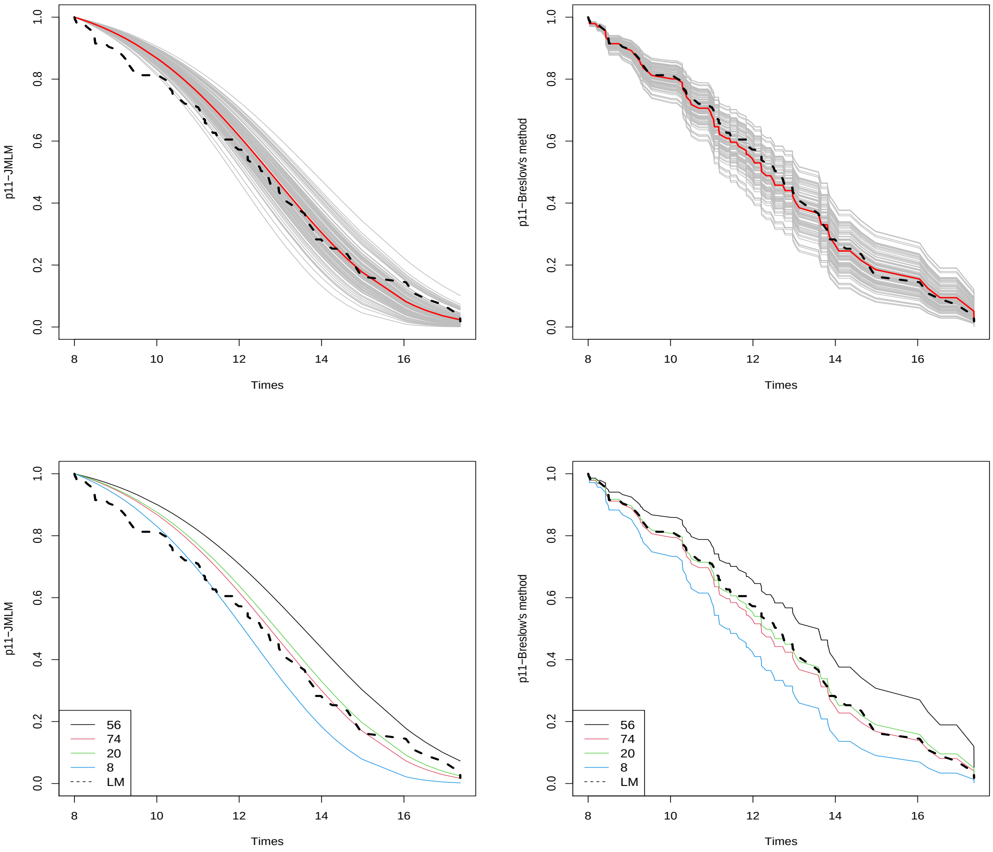


FIGURE 5.3: Spaghetti plot for transition probabilities $\hat{p}^{i}_{11}(8, t \mid Y_i)$, for all individuals belong to the subset $\mathcal{S}_1$, using the proposed JMLM estimator (First row, left column) and the BRES estimator (First row, right column). Transition probabilities $\hat{p}^{i}_{11}(8, t \mid Y_i)$ using the JMLM estimator (Second row, left column) and the BRES estimator (Second row, right column) for selected individuals belong to the subset $\mathcal{S}_1$.

5.4 Discussion

The estimation of transition probabilities is a major goal in multi-state models since they allow for long term prediction of the process. In the literature there exist a set of approaches to relate the effect of the individual characteristics given by a covariate (or a vector of covariates) for the transition probabilities among states. A common method is to decouple the whole and, for each transition, considering Cox's regression models for modeling the transition intensities whose baseline hazard function are estimated by the Breslow's method. This method is particularly well-suited to deal with multiple covariates, but it is restricted to a single value for each covariate and it does not take into consideration individual longitudinal trends of biomarkers. In this chapter, we revisit the problem of estimation of these quantities and introduce new proposals of estimators given a continuous covariate measured repeatedly over time.

Joint modeling of longitudinal and survival data are becoming increasingly popular in clinical research. In these models, as the longitudinal markers are measured with errors, they are not considered as time-dependent covariates unlike the Cox model that assumes that the exact values of the explanatory variables are known for all the individuals at risk at each event time. In this chapter we proposed a new method, called joint modeling landmark, for estimating the transition probabilities conditionally on covariates observed with repeated measures. This approach is based on the use of specific samples of data, consisting of subjects occupying a given state at a particular time. This procedure allows the adaptation of existing methods for joint modeling of longitudinal and survival data to estimate the transition probabilities for each individual while taking into account the repeated measures of a longitudinal variable.

To illustrate the ability of the new estimators, we have used two data sets involving the longitudinal and progressive illness-death models provided as supplementary material by Ferrer *et al.* (2016) [56]). Results seem to confirm the good performance of the proposed estimator with accurate estimated conditional transition probabilities. This method also revealed for a particular example more sensibility to reflect the evolution of the longitudinal measures when compared to the Breslow-based method, which only makes use of a single value of the covariate.

In further research developments, we intend to extend our approach to simulation studies to more complex multi-state models than the progressive illness-death model, as well as to include longitudinal submodels with more than one covariate and corresponding interactions. We also plan to illustrate the application of the proposed method to a real data set involving oncological studies.

Chapter 6

Goodness-of-fit test statistics for the Markov condition

The inference in multi-state models is traditionally performed under a Markov assumption that claims that past and future of the process are independent given the present state. This assumption has an important role in the estimation of the transition probabilities. When the multi-state model is Markovian, the Aalen-Johansen estimator gives consistent estimators of the transition probabilities but this is no longer the case when the process is non-Markovian. Usually, this assumption is checked including covariates depending on the history. Since the landmark methods of the transition probabilities are free of the Markov assumption, they can also be used to introduce such tests by measuring their discrepancy to Markovian estimators. In this chapter, we introduce tests for the Markov assumption and compare them with the usual approach based on the analysis of covariates depending on history through simulations. The methods are also compared with more recent and competitive approaches. Three real data examples are included for illustration of the proposed methods.

The contents of this chapter are mainly based on the paper published in *Computational Statistics* by Soutinho and Meira-Machado (2021) [108].

6.1 Introduction

Multi-state models are the most suitable models for the description of complex longitudinal survival data involving several events of interest. The inference for transition intensities often includes regression analysis which usually involves the modeling of each transition intensity separately. A traditional choice is to model each transition intensity using a proportional hazards model assuming the process to be Markovian. However, it has been quoted that the Markov assumption is violated in some applications (Andersen and Esbjerj and Sorensen (2000) [20]; Andersen and Keiding (2002) [17]). In such cases, if interest is on multi-state regression, one alternative approach is to use a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. Semi-Markov models are also called "clock reset" models because each time the patient enters a new state, time is reset to 0. In terms of transition probabilities, the Markov assumption also allows the construction of simple estimators, since individuals with different past histories become comparable (Aalen and Johansen (1978) [61]). Unfortunately, when this assumption is violated, the use of the so-called Aalen-Johansen estimators for transition probabilities can induce bias, and thus may not be recommended. To tackle this, substitute estimators for the Aalen-Johansen estimator for a non-Markov process were introduced by de Uña-Álvarez and Cadarso-Suárez (2006) [68], Allignol *et al.* (2014) [109], Uña-Álvarez and Meira-Machado (2015) [69] and Titman (2015) [110].

Therefore, to perform inference for transition intensities or for the transition probabilities it is essential to check if the Markov assumption is tenable. This assumption is usually checked by including covariates depending on the history (Kay (1986) [22]; Andersen and Esbjerj and Sorensen (2000) [20]; Andersen and Keiding (2002) [17]). Alternative methods, based on a local Kendall's τ, measuring the future-past association along time, were proposed by Rodríguez-Girondo and de Uña-Álvarez (2012, 2016) [111] [112]. These methods can be used for three-state progressive and illness-death models but the extension of these tests to general multi-state models is not straightforward and thus, flexible methods that may be used in general models are required. A very recent work by Titman and Putter (2020) [113] considers new approaches to check this assumption. In one of these approaches a general test is developed by considering summaries from families of log-rank statistics where patients are grouped by the state occupied at different times. Chiou *et al.* (2018) [114] also consider an equivalent problem for testing Markoviaty (in the progressive illness-model) but involving tests for dependent truncation.

The organization of this chapter is as follows. Section 6.2 gives an introduction to the methodological background and introduces tests for checking the markov assumption. In Section 6.3, we evaluate the performance of the proposed methods and compare them with competive methods through simulations studies. In Section 6.4, the use of the proposed methods is illustrated by the analysis of an illness-death model describing the disease process of breast and colon cancer patients. Liver cirrhosis data is used to illustrate the application of the proposed methods to more general models. Main conclusions and discussion are reported in Section 6.5. To provide the biomedical researchers with an easy-to-use tool to compute these proposed methods we develop an R package called `markovMSM` which are available available at the CRAN repository at `https://cran.r-project.org/web/packages/markovMSM`. Details on the usage of its functions can be obtained with the corresponding help pages. The main funcionalities of the package are presented in Section A.3.

6.2 Tests for the markov assumption

Traditionally, the Markov condition is verified by modeling particular transition intensities on aspects of the history of the process using a proportional hazards model (Kay (1986) [22]). In the progressive illness-death model, for example, with state space $\{1, 2, 3\}$, we can examine whether the time spent in the initial state is important on the transition from the disease state (the intermediate state) to death (the absorbing state) or not. For doing that, let $\lambda_{23}(t)$ denote the hazard function of T for those individuals going from State 2 to State 3, and let Z denote the sojourn time in State 1. Fitting a Cox model $\lambda_{23}(t \mid Z) = \lambda_{23,0}(t)\exp(\beta Z)$, where $\lambda_{23,0}$ is the baseline hazard and β a regression parameter, we now need to test the null hypothesis, $H_0 : \beta = 0$, against the general alternative, $H_1 : \beta \neq 0$. This would assess if the transition rate from the disease state into death is unaffected by the time spent in the initial state. It is worth to remember that the semi-parametric Cox proportional hazard model is based on the assumption of proportional hazards and that it assumes a linear effect on the hazard for the covariate. Both may fail in practice, and consequently this approach may be unable to detect the lack of Markovianity.

Since the landmark methods (`LM`) for estimating the transition probabilities proposed by de Uña-Álvarez and Meira-Machado (2015) [69], and the landmark Aalen-Johansen estimators (`LMAJ`) by Putter and Spitoni (2018) [71], are free of the Markov assumption, they can also be used to introduce local tests for Markovianity by measuring their discrepancy to Markovian Aalen-Johansen estimators (`AJ`), for a fixed value $s > 0$. Though the two landmark methods behave similarly, the `LMAJ` can be used in general multi-state models which can be considered an advantage. These ideas were recently used by Titman and Putter (2020) [113] to introduce a general tests based on summaries from families of log-rank statistics where patients are grouped by the state occupied at a given (landmark) time.

In this chapter we also introduce a local test based on the areas under the two curves, AUC, (i.e., the curves of the estimated transition probabilities) that can be used for a general multi-state model. We propose the use of the following test statistic based on direct nonparametric estimates of the transition probabilities,

$$U = \int_s^\tau \left(\widehat{p}_{hj}^{\,\texttt{LMAJ}}(s,u) - \widehat{p}_{hj}^{\,\texttt{AJ}}(s,u) \right) du,$$

where τ is the upper bound of the support of T. The test statistics can be seen as the difference between the area under the estimated transition probability curve for the non-markov `LMAJ` estimator and the `AJ` estimator. Intuitively, the test statistics should be close to zero if the process is Markov. The Markov assumption becomes less likely as the test statistic get further away from zero in either direction. Because of censoring, both estimators (`LMAJ` and `AJ`) may reveal high variability in the right tail which may inflate the test statistic. In addition to this issue, since landmarking is based on reduced data, the maximum point for which the `LMAJ` transition probability estimate is strictly defined may be lower than the maximum point for `AJ`. To overcome these problems, we suggest that in the computation of U one should use the minimum between the upper bound for which `LMAJ` is defined and the 90% percentile of the total time for the upper limit in the integral that defines the test statistic. In the progressive illness-death model, besides the transition probability $\widehat{p}_{23}(s,t)$, also $\widehat{p}_{12}(s,t)$ can be used to test the Markov assumption. For general multi-state models, one can use transitions depending on history (i.e., $p_{hj}(s,t)$ depending on subject specific arrival time at state $h > 1$). In fact, if the goal is to decide which estimator is the most appropriate to estimate a specific transition probability $p_{hj}(s,t)$, then the test statistic should be the one based on that same transition probability.

Note that if the null hypothesis of Markovianity holds, the value of U should be close to zero. To approximate the distributions of the test statistic, bootstrap methods with a large number of resamples are used. We generate M bootstrap samples and for each sample the test statistic $U^\star$ is calculated. Then, according to large sample asymptotic distribution theory, when M, the number of replicates goes to infinity, we have the following statistic distributed approximately as a standard normal distribution with a mean of 0 and variance of 1: $V = (U - 0)/\sigma^{\star}_{(U^{\star})} \sim N(0,1)$. The null hypothesis will be rejected if $V > v_{(1-\alpha/2)}$ or $V < v_{(\alpha/2)}$, where $v_{(\alpha/2)}$ and $v_{(1-\alpha/2)}$ denote the $\alpha/2$ and $1-\alpha/2$ percentiles, respectively, of a normal distribution with a mean of 0 and variance of 1.

In this chapter we also propose a global test which can be achieved by combining the results obtained from local tests over different times. The testing procedure used here involves the following steps:

Step 1: Using the original sample of the illness-death model, obtain the percentiles 5, 10, 20, 30 and 40 of the sojourn time in State 1. For general multi-state models, we recommend the use of the same percentiles of the subject specific arrival time at the corresponding state.

Step 2: For each of the values s obtained in Step 1, obtain the probability values for the local method as explained before.

Step 3: Obtain the mean of the probability values for each closest pairs; i.e., the mean of the probability values of the following pairs of percentiles: $(5,10)$, $(10,20)$, $(20,30)$ and $(30,40)$.

Step 4: Get the minimum between the four probability values obtained in Step 3.

Step 1 considers a global test based on local tests computed at low percentiles of subject specific arrival times at the corresponding state. This is based on our experience that the failure of Markovianity often occurs for small transition times. Besides the hypothesis tests proposed above, in Section 6.4 we also propose graphical local tests that can be used to check the Markov assumption in the illness-death model as well as for more complex multi-state models, possibly with reversible transition between states. These graphical tests can be used to validate the default values proposed in Step 1 or to propose alternative values for which a discrepancy between the two methods (`LMAJ` and `AJ`) is more evident. The procedure described in Step 3 can be used to ensure that there is a discrepancy between the two estimated curves in a large range of time values.

To provide the biomedical researchers with an easy-to-use tool to compute the proposed methods, we have developed a R package which details on the usage are available as supplementary material [A.3]. The package allows users to choose different percentiles for the sojourn time in State 1 (Step 1).

6.3 Simulation studies

In this section, we report results from simulation studies, where the aim is to compare the finite sample performance of the proposed methods to test the Markov assumption in a progressive illness-death model. Due to computing time issues the simulations shown here only address this model. However, an application of the proposed methods to a more complex multi-state model is presented in Section 6.4 from a real data set. To simulate the data in the progressive illness-death model, we assume that all individuals are in the initial state (State 1) at time $t = 0$, and that these individuals may follow two possible paths: passing through the intermediate state (State 2), at some specific time; or going directly to the absorbing state (State 3). Transition times for those leaving the initial state are generated from the cause-specific hazards given by $\lambda_{12}(t) = 0.29/(t+1)$ and $\lambda_{13}(t) = 0.024 \times t$, where $t > 0$ denotes the time since the start point. To study the Markov assumption, three different scenarios are considered corresponding to different hazards that are used to generate death times for individuals passing through the intermediate state: $\lambda^1_{23}(t) = 0.05$, $\lambda^2_{23}(t) = 0.25(t_{12}+1)^{-0.8}$ and $\lambda^3_{23}(t) = 0.04\log(t+1)$, where t_{12} is the transition time to the intermediate event. The first scenario is Markov since the hazard is independent of time, whereas the second is semi-Markov and the third is non-Markov. Censoring times were generated from uniform distributions. Two samples size were considered for each scenario ($n = 250$ and $n = 500$).

We also consider a fourth scenario in which the traditional test, based on the Cox proportional hazard model may fail. In this scenario, the transition times are generated from the following cause-specific hazards given by $\lambda_{12}(t) = 1/(2-t)$, $\lambda_{13}(t) = 2/(3-2t)$ for $0 \leq t < 2$ and $0 \leq t < 1.5$, respectively. To generate death times for individuals passing through the intermediate state we consider $\lambda_{23}(t) = \exp\left(-(t_{12}-1)^2\right)$. This simulated scenario is the same as that described in Rodríguez-Girondo and Uña-Álvarez (2016) [112]. Note that this scenario is non-Markov, because of the dependence on the transition time to the intermediate state but in this case a misspecification of the Cox model is expected because of the shape of the hazard $\lambda_{23}(t)$ with a parabolic influence of the predictor.

Table 6.1 reports the rejection proportions of the proposed tests for the first three scenarios with sample sizes $n = 250$ and $n = 500$. Random censoring was simulated using uniform distributions $U[0, 60]$ and $U[0, 30]$. The first censoring distribution led to medium censoring percentages (between 41% and 47%) whereas these percentages increase in the

TABLE 6.1: Rejection proportions for nominal level of 5% of the local tests for fixed values $s = 1$, $s = 2$, $s = 4$, $s = 6$ and $s = 8$ (AUC(s) and LR(s)). Rejection proportions for the global tests (AUC and Cox) are also included. Censoring times uniformly distributed between 0 and 30, and between 0 and 60.

										Global	
Scenario	Trans. Prob.	n	C	Method	1	2	4	6	8	AUC / LR	Cox
Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.055	0.055	0.064	0.073	0.062	0.073	0.046
		500	$U[0,30]$	AUC(s)	0.066	0.057	0.069	0.072	0.076	0.057	0.045
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.051	0.047	0.054	0.051	0.056	0.043	0.046
		500	$U[0,30]$	LR(s)	0.036	0.048	0.054	0.052	0.057	0.052	0.045
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.055	0.043	0.049	0.046	0.033	0.076	0.046
		500	$U[0,30]$	AUC(s)	0.060	0.052	0.061	0.065	0.055	0.056	0.045
Semi-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.765	0.762	0.611	0.437	0.286	0.845	0.757
		500	$U[0,30]$	AUC(s)	0.964	0.961	0.881	0.701	0.530	0.992	0.977
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.872	0.891	0.739	0.520	0.296	0.960	0.757
		500	$U[0,30]$	LR(s)	0.996	0.999	0.976	0.862	0.635	1.000	0.977
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.759	0.744	0.536	0.316	0.131	0.862	0.757
		500	$U[0,30]$	AUC(s)	0.967	0.955	0.855	0.648	0.449	0.993	0.977
non-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,30]$	AUC(s)	0.172	0.284	0.308	0.292	0.258	0.354	0.382
		500	$U[0,30]$	AUC(s)	0.336	0.458	0.508	0.502	0.468	0.602	0.701
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	LR(s)	0.225	0.268	0.267	0.241	0.191	0.414	0.382
		500	$U[0,30]$	LR(s)	0.369	0.464	0.515	0.479	0.384	0.696	0.701
	$\widehat{p}_{23}(s,t)$	250	$U[0,30]$	AUC(s)	0.172	0.240	0.226	0.176	0.114	0.302	0.382
		500	$U[0,30]$	AUC(s)	0.348	0.452	0.474	0.420	0.332	0.574	0.701
Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.048	0.038	0.048	0.050	0.072	0.066	0.058
		500	$U[0,60]$	AUC(s)	0.052	0.052	0.050	0.042	0.070	0.062	0.038
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.055	0.055	0.061	0.053	0.054	0.043	0.058
		500	$U[0,60]$	LR(s)	0.064	0.067	0.053	0.054	0.052	0.046	0.038
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.048	0.036	0.042	0.044	0.068	0.062	0.058
		500	$U[0,60]$	AUC(s)	0.050	0.054	0.050	0.032	0.068	0.062	0.038
Semi-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.918	0.946	0.84	0.736	0.600	0.980	0.926
		500	$U[0,60]$	AUC(s)	0.998	1.000	0.982	0.940	0.876	1.000	0.940
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.961	0.970	0.943	0.847	0.708	0.998	0.926
		500	$U[0,60]$	LR(s)	1.000	1.000	0.999	0.999	0.961	1.000	0.940
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.918	0.928	0.812	0.664	0.490	0.982	0.926
		500	$U[0,60]$	AUC(s)	0.996	1.000	0.982	0.942	0.848	1.000	0.940
non-Markov	$\widehat{p}_{12}(s,t)$	250	$U[0,60]$	AUC(s)	0.282	0.382	0.442	0.410	0.382	0.504	0.368
		500	$U[0,60]$	AUC(s)	0.474	0.652	0.724	0.724	0.656	0.754	0.692
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	LR(s)	0.260	0.341	0.431	0.412	0.376	0.546	0.368
		500	$U[0,60]$	LR(s)	0.504	0.650	0.739	0.711	0.663	0.861	0.692
	$\widehat{p}_{23}(s,t)$	250	$U[0,60]$	AUC(s)	0.276	0.344	0.404	0.330	0.288	0.472	0.368
		500	$U[0,60]$	AUC(s)	0.506	0.648	0.692	0.704	0.656	0.758	0.692

second censoring distribution (between 45% and 62%). Four tests are considered in this table: (i) local test based on the area under the transition probabilities $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, denoted by AUC(s); (ii) local test proposed by Titman and Putter (2020) [113], based on the log-rank, for the transition probability $\widehat{p}_{23}(s,t)$, denoted by LR(s); (iii) global test based on the area under the transition probabilities (AUC) or the global test base on the log-rank statistics (LR) (Titman and Putter (2020) [113]); (iv) global test based on the Cox model (Cox). The global test LR is based on the mean value of the log-rank statistics as described in Titman and Putter (2020) [113]. The local tests were evaluated at five fixed values $s = 1$, $s = 2$, $s = 4$, $s = 6$ and $s = 8$. Results in this table were obtained by the empirical rejection proportions from 1000 trials at the significant level of 0.05.

Results show that, for the semi-Markov and non-Markov scenarios, the power of the tests is higher for lower censoring percentages, increasing with the sample size. The bootstrap test based on the areas under the curves (of the transition probabilities) (AUC) and the local test based on log-rank statistics both reveal their capacity to identify the differences between curves in the semi-Markov scenario showing higher rejection probabilities for lower values of s. Note that in this scenario, departures between the two curves (obtained from `AJ` and `LMAJ` methods) are expected to decrease as the difference $t-s$ increase. In non-Markov scenario, departures between the two curves (obtained for the transition probabilities $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$ from `AJ` and `LMAJ` methods) denote a great improvement when considering a sample size of n = 500, but with rejection probabilities below 0.50 for all s, with the exception for censoring uniform distribution $U[0,60]$. Both local tests also obtain low rejection proportions (near the nominal level of 5%) when the data is generated from a Markov scenario. Note that we expect rejection proportions about 0.05 in this case. The results based on the log-rank statistic also confirm the good accuracy of this method in agreement with the conclusions shown in Titman and Putter (2020) [113]. In general for all scenarios, sample sizes and censoring distributions, results between the log-rank test and the local AUC test are quite similar being able to distinguish the inequality between `AJ` and `LMAJ` curves in semi-Markov and non-Markov scenarios, while providing low rejection proportions when the process is indeed Markovian. When comparing the results for the two local tests based on different transition probabilities, $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, it can be seen that they provide similar values but slightly higher when based on the computation of the transition probability $\widehat{p}_{12}(s,t)$. This behavior may be explained by the number of observations from which the transition probability is computed, those in State 1 at time s for $\widehat{p}_{12}(s,t)$, and those in State 2 at the same time for $\widehat{p}_{23}(s,t)$. For completeness purposes, Table 6.1 also shows the results from the three global tests. These global tests present satisfactory results in all scenarios, reporting rejection proportions of about 5% for the Markov scenario, and high levels of rejection proportions for the semi-Markov and non-Markov scenarios. These results are in accordance with those obtained using a local test based on the area under the curves of the estimated transition probabilities. As expected, in general, the performance of the proposed methods is improved for scenarios with less censoring percentages (i.e., for censoring times following an uniform distribution $U[0,60]$). This improvement is not so obvious for the method based on the Cox model. We can also notice that the global log-rank and the AUC global tests behave similarly in all cases. Some of these patterns, for censoring uniform distribution $U[0,30]$, can be clearly seen in Figure 6.1.

Table 6.2 reports the rejection proportions of the four proposed tests for the forth scenario, non-Markovian with an hazard based on a quadratic predictor. Random censoring was simulated from uniform distributions $U[0,\tau_G]$ for τ_G equal to 8.1 and 4.6. The model

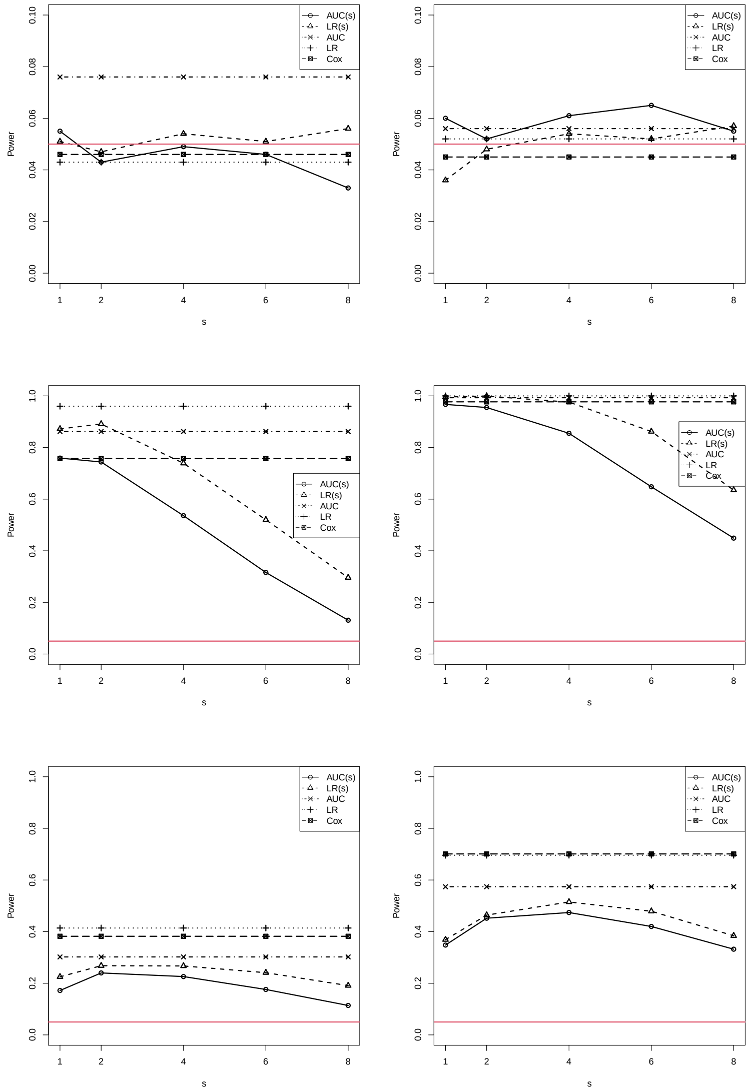


FIGURE 6.1: Rejection probabilities for testing the null hypothesis of the Markov condition for the three tests for nominal level 5%. Markov, semi-Markov and non-Markov scenarios (upper, middle, and lower panels, respectively), for $n = 250$ and $n = 500$ (left and right panels, respectively). Results for the transition probability $\hat{p}_{23}(s, t)$. Censoring times uniformly distributed between 0 and 30.

with $\tau_G = 8.1$ results in 12% censoring on the first gap time and in 24% for the total time. The model with $\tau_G = 4.6$ increases these censoring levels to 20% and about 40%, respectively. In this case, the global method based on the Cox proportional model has a bad performance which can be explained by failure of the linear specification of the Cox model. It can also be seen that the power of this test does not increase substantially with the sample size, as it happens in semi-Markov and non-Markov scenarios shown in Table 6.1. Results shown in Table 6.2, reveal that the tests (local and global) based on the area under the curves have a good performance, revealing reasonable levels of rejection

TABLE 6.2: Rejection proportions for nominal level of 5% of the local tests for fixed values $s = 0.2$, $s = 0.6$, $s = 1$, $s = 1.2$, $s = 1.4$ and $s = 1.6$ (AUC(s) and LR(s)). Rejection proportions for the global tests (AUC and Cox) are also included. Non-Markovian scenario, hazard with a quadratic predictor.

									Global	
Scenario	Trans. Prob.	n	Method	0.2	0.6	1	1.4	1.6	AUC/LR	Cox
Non-Markov	$\widehat{p}_{12}(s,t)$	250	AUC(s)	0.270	0.260	0.042	0.186	0.256	0.360	0.074
quadratic		500	AUC(s)	0.492	0.504	0.094	0.278	0.468	0.708	0.094
predictor	$\widehat{p}_{23}(s,t)$	250	LR(s)	0.348	0.283	0.053	0.169	0.264	0.439	0.074
		500	LR(s)	0.638	0.542	0.065	0.299	0.489	0.815	0.094
$C \sim U[0, 8.1]$	$\widehat{p}_{23}(s,t)$	250	AUC(s)	0.324	0.314	0.064	0.128	0.162	0.430	0.074
		500	AUC(s)	0.538	0.532	0.010	0.228	0.376	0.742	0.094
Non-Markov	$\widehat{p}_{12}(s,t)$	250	AUC(s)	0.250	0.276	0.072	0.092	0.186	0.410	0.092
quadratic		500	AUC(s)	0.422	0.455	0.112	0.122	0.256	0.638	0.107
predictor	$\widehat{p}_{23}(s,t)$	250	LR(s)	0.294	0.257	0.063	0.115	0.161	0.305	0.092
		500	LR(s)	0.535	0.449	0.059	0.172	0.287	0.647	0.107
$C \sim U[0, 4.6]$	$\widehat{p}_{23}(s,t)$	250	AUC(s)	0.238	0.294	0.080	0.062	0.098	0.420	0.092
		500	AUC(s)	0.416	0.430	0.114	0.094	0.168	0.642	0.107

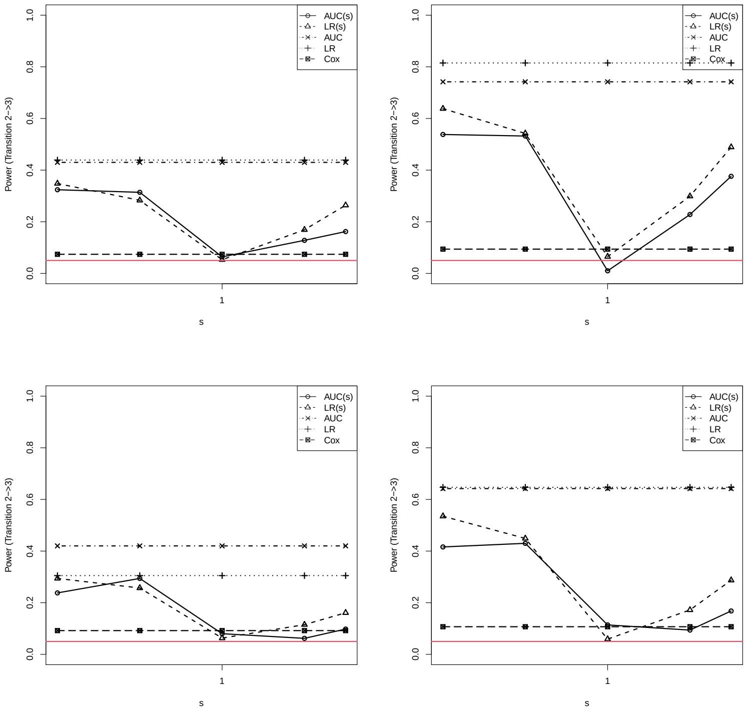


FIGURE 6.2: Rejection probabilities for testing the null hypothesis of the non-Markov condition for the three tests for nominal level 5%. Non-Markovian scenario, hazard with a quadratic predictor. Results based on different censoring percentages ($C \sim U[0, 8.1]$ - upper, $C \sim U[0, 4.6]$ - bottom), for $n = 250$ and $n = 500$ (left and right panels, respectively). Results for the transition probability $\hat{p}_{23}(s,t)$.

proportions of Markovianity. It can be seen that the power of these tests increases with the sample size. Results in terms of power performance for non-Markovian scenario, hazard with a quadratic predictor are shown in Figure 6.2. The plots show the rejection probabilities for the transition probability $\hat{p}_{23}(s,t)$ as a function of s. Simulation results also confirm the similarity of the local and global tests between the log-rank and the AUC test for both scenarios.

Rodríguez-Girondo and Uña-Álvarez (2016) [112] also introduced methods for checking the Markov assumption for the progressive illness-death model. The performance of their methods was studied through simulation studies. Among the methods for simulating data, their model 2 is the one that we aim to reproduce in our scenario 4, making some comparisons possible. As in their case, our simulations reveal the inability of the Cox model to identify the failure of the Markovianity with proportion rejections varying between 5% and 10%. As in our case, the methods proposed in Rodríguez-Girondo and Uña-Álvarez (2016) [112] revealed an increased power of the global tests as the sample size increases and with a decrease in the censoring percentage. Among the proposed tests, the wC_n method, based on the local Kendall's tau τ_i, appears to be the one with better accuracy to distinguish the non-markovianity of the process either for subjects who pass directly from State 1 to State 2 or for those that have passed through the intermediate state. Comparing the results of the AUC global test, reported in our Table 6.2, to the proposed wC_n method, namely for individuals that experienced a transition through the intermediate state, we can observe higher rejection proportions for the AUC test for all samples sizes (n=250 and n=500) and censoring parameters (4.6 and 8.1). It is worth remember that the extension of the methods proposed in Rodríguez-Girondo and Uña-Álvarez (2016) [112] to general models is not straightforward, while our methods (based on the AUC) can be applied to general multi-state models as illustrated in our third real data example.

6.4 Real data analysis

In this section, we illustrate the proposed methods using data from three clinical trial studies. We first use data from a colon cancer study from a large clinical trial on Duke's stage III patients (Moertel *et al.*, 1995) [115]), the second one is from a clinical trial on breast cancer and the last one from a data set of liver cirrhosis patients subjected to a prednisome treatment (Andersen *et al.*, 1993) [13].

Surgical resection is the best treatment option for cancer patients and the most powerful tool for assessing prognosis following potentially curative surgery. In a large percentage of the patients with such cancers, the diagnosis is made at a sufficiently early stage when all apparent disease tissue can be surgically removed. Unfortunately, some of these patients have residual cancer, which leads to recurrence of the disease and death (in some cases). Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of mortality. Usually, this mortality is higher in cases of early recurrences. The effect of a recurrence in a survival model is traditionally studied using extensions of the Cox proportional hazards model (Cox (1972 [6]); Genser and Wernecke (2005) [116]). Multi-state models can also be successfully used to model such data (Pérez-Ocón *et al.* (2001) [117]; Putter, Fiocco and Geskus (2007) [18]; Meira-Machado *et al.* (2009) [15]; Meira-Machado (2016) [73]; Meira-Machado and Sestelo (2019) [16]). In both real data examples from cancer studies, data can be viewed as arising from a progressive illness-death model with states 'Alive and disease-free', 'Alive with Recurrence' and 'Dead'. Below, the Markov assumption is carefully analyzed comparing the proposed methods with the traditional approach.

6.4.1 Colon cancer study

In this study, 929 patients affected by colon cancer were followed from the date of a curative surgery for colorectal cancer until censoring or death from colon cancer. From this total, 468 developed a recurrence and among these 414 died; 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up.

Figure 6.3 reports estimated transition probabilities for fixed values of $s = 365, 730$, 1095 and 1460 days (1, 2, 3 and 4 years, respectively), along time, for the transition probabilities $\widehat{p}_{12}(s,t)$ (left hand side) and $\widehat{p}_{23}(s,t)$ (right hand side). As expected, these plots reveal that the landmark estimators (`LM` and `LMAJ`) have more variability than the Aalen-Johansen estimator (`AJ`). This is a obvious consequence of the subsampling approach which will be more evident for some specific values of s and higher values of t. Plots shown in the top of the figure (for s equal to one year) show departures between the two Markov-free estimators (`LM` and `LMAJ`) and the Aalen-Johansen estimator (`AJ`). Note that for the mortality transition from State 2 to State 3 the two (Markov-free) landmark estimators are equivalent. Deviations from the two approaches (Markovian and Markov-free), as those shown for s equal to one year, may be explained by the failure of the Markov assumption. On the other hand, the corresponding plots for the remaining values of s show that all methods behave quite similar.

Plots shown in the first row of Figure 6.4 compare the Aalen-Johansen estimator (`AJ`) and the landmark non-Markovian estimator (`LMAJ`) for $p_{12}(s = 365, t)$, $p_{13}(s = 365, t)$ and $p_{23}(s = 365, t)$. A small deviation can be seen in these plots with respect to the straight line $y = x$. The plot on the second row presents the estimated transition probabilities $\widehat{p}_{23}(s = 365, t)$ from the landmark Aalen-Johansen estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line), revealing some discrepancies between the two approaches in the estimation of this transition probability. These plots provide a graphical test of the Markov assumption which reveal some evidence on the lack of Markovianity of the underlying process beyond one year after surgery.

For further illustration, in Figure 6.5 we display the discrepancy between the Aalen-Johansen estimator (Markovian) and the landmark non-Markovian estimator (`LMAJ`), for $p_{12}(s,t)$ and $p_{22}(s,t)$, for $s = 365$, $s = 730$, $s = 1095$ and $s = 1460$, measured through $D_{hj} = \widehat{p}_{hj}^{\,\mathtt{AJ}}(s,t) - \widehat{p}_{hj}^{\,\mathtt{LMAJ}}(s,t)$, $h = 1,2$, $j = h+1$. The 95% pointwise confidence limits were obtained using simple bootstrap. This plot reveals clear differences between the two methods in large intervals for $s = 365$. The differences are observed by the deviation of the plot with respect to the straight line $y = 0$, from which one gets some evidence on the lack of Markovianity of the underlying process beyond one year after surgery. On the other hand the plots depicted for other values of s do not reveal evidence against the Markov assumption. In summary, these plots show that there is some evidence, at least for $s = 365$, that the application of the Aalen-Johansen method is not recommended here, due to possible biases. They also reveal a possible failure of the Markov assumption. It is worth mention that deviations of the plots with respect to the straight line $y = 0$ in the right tail (higher values of t) should not be overvalued since they often occur due to the limited number of individuals at these times. Note that the findings observed in Figure 6.4 are not in agreement with the results obtained through the 'global' test for Markovianity based on the Cox model (using time to recurrence as a time-dependent covariate). This test reported a coefficient of negative sign for the recurrence time, according to an increased risk of death shortly after relapse (p-value = 0.154) revealing no evidence against the Markov model for the colon data.

Results reported in Table 6.3 are in agreement with those obtained from the graphical inspection shown in Figure 6.5, revealing a failure of the Markov assumption only for non-null lower values of s. They show that, the test based on the difference of the area under the two curves lead to a probability value of 0.002 and 0.003, respectively for $\widehat{p}_{12}(s,t)$ and $\widehat{p}_{23}(s,t)$, for $s = 365$. Low probability values (less than 5%) were also obtained for $s = 90$ and $s = 180$ too. These findings were also confirmed by the local tests based on

TABLE 6.3: Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Colon cancer data.

								Global	
Trans. Prob.	Method	90	180	365	730	1095	1460	AUC / LR	Cox
$\hat{p}_{12}(s,t)$	AUC(s)	0.012	0.007	0.002	0.154	0.135	0.857	0.014	0.154
$\hat{p}_{23}(s,t)$	LR(s)	0.006	0.026	0.036	0.685	0.981	0.509	0.018	0.154
	AUC(s)	0.003	0.004	0.003	0.155	0.118	0.714	0.013	0.154

TABLE 6.4: Probability values of the local test for $s = 365$ days by treatment for AUC local test. Rejection proportions for the test based on the Cox model also included. Colon cancer data.

Trans. Prob.	Treatment	Method	s=365	Cox
$\hat{p}_{12}(s,t)$	Obs	AUC(s)	0.0002	
	Lev		0.7192	
	Lev+5FU		0.1116	
$\hat{p}_{23}(s,t)$	Obs	AUC(s)	0.0008	0.062
	Lev		0.3013	0.401
	Lev+5FU		0.1562	0.712

the log-rank statistic. The global test we propose (based on the areas under the transition probabilities) are also in agreement with our findings, reporting a probability value lower than 0.014 against the Markov condition. The local tests based on the log-rank statistic also confirmed small probability values mainly for s up to 365. Either the AUC and the log-rank global tests confirm the failure of the Markovianity of the process.

Often multi-state models include covariates and it may be the omission of covariate effects that induces apparent non-Markovianity. The methods proposed in this chapter can also deal with this problem since discrete covariates can be included in the estimation of the transition probabilities $p_{hj}(s,t)$ by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. As shown in Table 6.4 treatment (Obs(ervation), Lev(amisole), Lev(amisole)+5-FU) revealed a strong effect on the 2→3 transition intensities and a greater effect on 1→ 2. Results reported in Table 6.4 also show that the test for Markovianity based on the Cox model reported a p-value of 0.062 (regression coefficient: -0.000528) for the Observation group.

6.4.2 Breast cancer data

In this section we use data from the second trial in which a total of 720 women with primary node positive breast cancer were recruited in the period between July 1984 and December 1989. The data which was also used by Sauerbrei and Royston (1999) [118] considers 686 patients who had complete data for the two event times (time to recurrence and time to death). In this study, patients were followed from the date of breast cancer diagnosis until censoring or dying from breast cancer. From the total of 686 women, 299 developed a recurrence and 171 died.

As for the analysis of the colon cancer data, we start to present on Figure 6.6 the estimated transition probabilities for fixed values of $s = 365$, 730, 1095 and 1460 days, along time, for the transition probabilities $\widehat{p}_{12}(s,t)$ (left hand side) and $\widehat{p}_{23}(s,t)$ (right hand side). In this case, differences between the estimated curves of the Aalen-Johansen (`AJ`) and the Landmark estimator (`LMAJ`) are not evident. The discrepancy of the two estimators with the 95% pointwise confidence limits is also displayed in Figure 6.7 for $D_{hj} = \widehat{p}_{hj}^{\texttt{AJ}}(s,t) - \widehat{p}_{hj}^{\texttt{LMAJ}}(s,t), h = 1,2, j = h+1$. In this case, there are no clear evidences of a deviance of the plot with respect to the straight line $y = 0$, at least in large intervals. In summary, these plots do not show evidence against the use of the Aalen-Johansen estimator and therefore, against the Markov assumption. These findings are in agreement with the results obtained through the three 'global' tests for Markovianity in Table 6.5. The test based on the Cox model which reported a negative coefficient sign for the recurrence time, according to an increased risk of death shortly after relapse (p-value = 0.121) revealing no evidence against the Markov model for the breast cancer data. Higher probability values were obtained from the global test based on the area under the transition probabilities and log-rank statistics. The two local tests confirm this fact too.

TABLE 6.5: Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Breast cancer data.

							Global	
Trans. Prob.	Method	180	365	730	1095	1460	AUC / LR	Cox
$\hat{p}_{12}(s,t)$	AUC(s)	0.543	0.306	0.232	0.247	0.241	0.230	0.121
$\hat{p}_{23}(s,t)$	LR(s)	0.926	0.647	0.246	0.163	0.922	0.580	0.121
	AUC(s)	0.955	0.603	0.269	0.428	0.577	0.280	0.121

6.4.3 Liver cirrhosis data

In this section we consider a data set of liver cirrhosis patients who were included in a randomized clinical trial at several hospitals in Copenhagen between 1962 and 1974. The study aimed to evaluate whether a treatment based on prednisone prolongs survival for patients with cirrhosis (Andersen *et al.* (1993) [13]). Let State 1 correspond to 'normal prothrombin level', State 2 to 'low (or abnormal) prothrombin level', and the State 3 to 'dead'. The movement of the patients among these three states can be modeled through the reversible multi-state model shown in Figure 6.8. From the total of 488 patients with liver cirrhosis initially enrolled in the study, 292 died, from which 104 experienced a direct transition from State 1 to the absorbing state, and in 188 patients an abnormal prothrombin level was detected at any time. There were also 314 patients that had movements from abnormal prothrombin levels towards normal levels and 274 from the normal prothrombin level to the intermediate state. Most transition times are below 1460 days, with a maximum of 4892 days.

Following the same procedure of the previous real data set analysis, we started comparing the estimated curves of the LMAJ and the AJ estimators for the transitions probabilities $\widehat{p}_{12}(s,t)$, $\widehat{p}_{21}(s,t)$ and $\widehat{p}_{23}(s,t)$, for fixed values of s= 180, 365, 730 and 1095 days, with the purpose to identify a possible failure of the Markov assumption. These times were chosen to cover the first years of the study corresponding to the most cases with transitions. In fact, after 4 years for all transitions the number of individuals decrease with potential consequences for the estimates under the landmark approach as refered previously in case of small size samples. The plots with the estimated curves at those points are shown in Figure 6.9. Plots shown in the first column reveal some departures between LMAJ and AJ estimators of $p_{12}(s,t)$, but only for lower values of s. The deviation between the two estimators seem to be more evident when comparing the estimated curves of $p_{21}(s,t)$ (second column), while this is not so evident when comparing the estimated curves of the transition probability $p_{23}(s,t)$ (third column). As referred above, apparent deviation between the two estimated curves, at least at some lowers values of s, may due to the lack of Markov condition.

The discrepancy of the two estimators, computed using $D_{hj} = \widehat{p}_{hj}^{\,\mathtt{AJ}}(s,t) - \widehat{p}_{hj}^{\,\mathtt{LMAJ}}(s,t)$ with the 95% pointwise confidence limits is also displayed in Figure 6.10. Some of these plots reveal some evidence of a deviance of the plot with respect to the straight line $y = 0$, revealing a possible failure of the Markov condition. Some of these findings are in agreement with the results reported in Table 6.6, which shows the rejection proportions, for $\widehat{p}_{12}(s,t)$, $\widehat{p}_{21}(s,t)$ and $\widehat{p}_{23}(s,t)$ of the proposed tests for checking the Markov assumption. Results were obtained by the empirical rejection proportions from 250 trials at the significant level of 0.05. Interestingly, the proposed local test was able to detect a failure of the Markov condition for $s = 365$ for the mortality transition of patients with abnormal prothrombin level. For the remaining time points of s, the test obtained lower rejection probabilities which are in agreement with the results obtained in all global tests. For the transition from State 1 to State 2, the proposed local test only reveal the failure of the Markov condition for $s = 180$. For the transition 2 to 1, besides $s = 180$, the local test also revealed a failure of the Markov condition for $s = 365$. These evidences (of failure of the Markov condition) for these two transitions are confirmed by the results of the proposed

TABLE 6.6: Probability values of the local test for several fixed values of s (measured in days). Rejection proportions for the global tests also included. Liver cirrhosis data.

							Global	
Trans. Prob.	Method	180	365	730	1095	1460	AUC / LR	Cox
$\hat{p}_{12}(s,t)$	AUC(s)	0.002	0.158	0.134	0.639	0.793	<0.001	0.002
$\hat{p}_{21}(s,t)$	AUC(s)	<0.001	<0.001	0.156	0.253	0.237	0.001	<0.001
$\hat{p}_{23}(s,t)$	LR(s)	0.699	0.336	0.594	0.641	0.034	0.298	0.999
$\hat{p}_{23}(s,t)$	AUC(s)	0.317	0.030	0.677	0.367	0.195	0.258	0.999

global test based on the AUC and the test based on the Cox model.

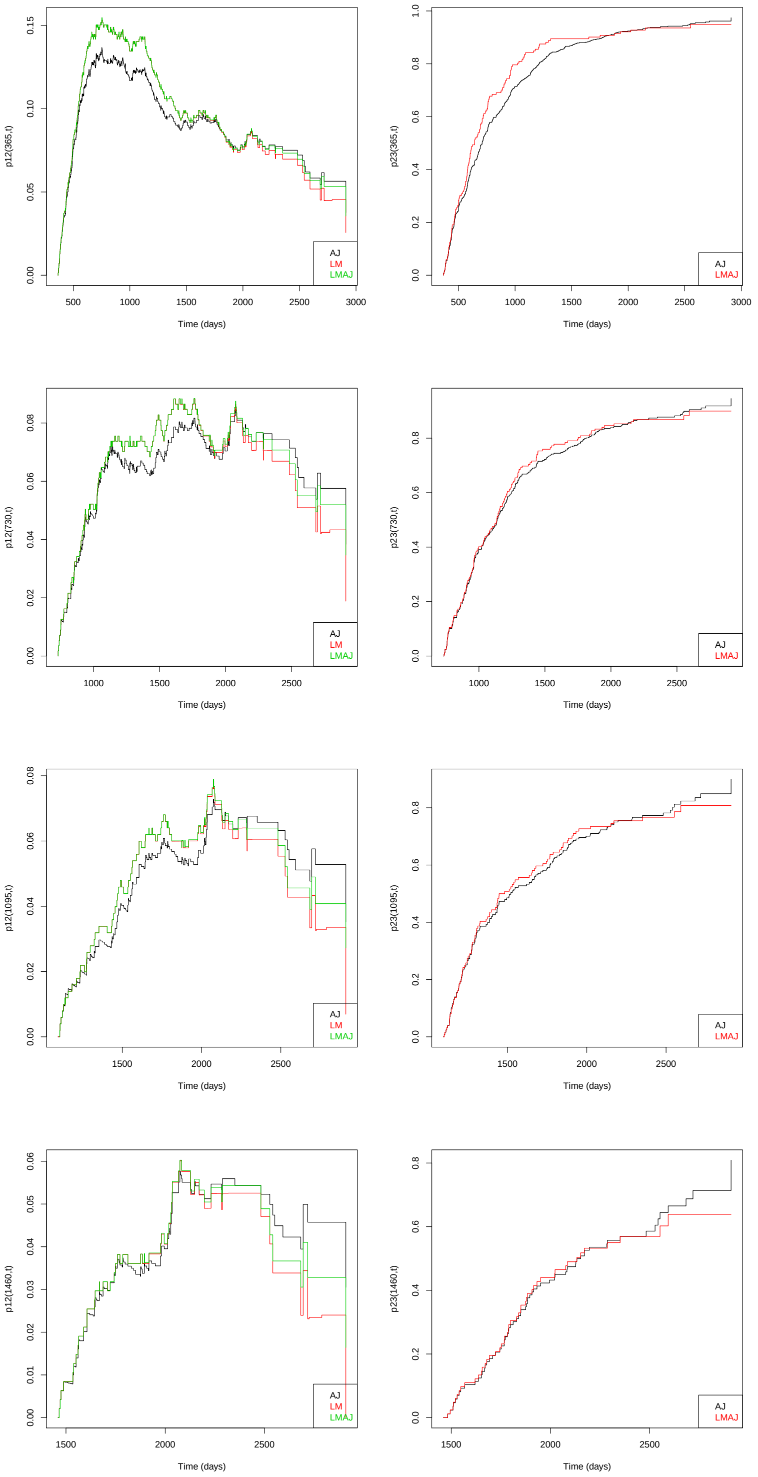


FIGURE 6.3: Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark and landmark Aalen-Johansen), for s equal to 1, 2, 3 and 4 years since entry in study. Colon cancer data.

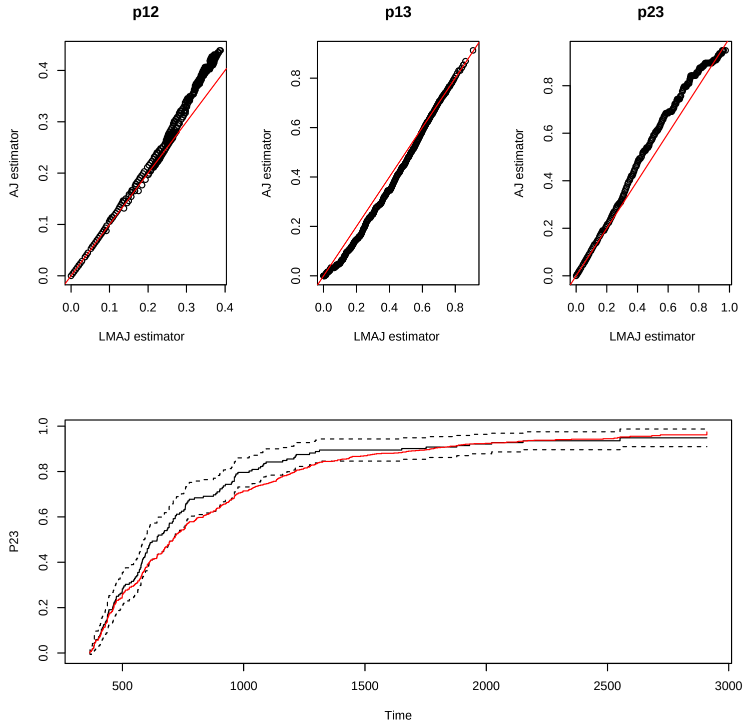


Figure 6.4: Graphical test for the Markov condition, $s = 365$ (First row). Transition probabilities of $\widehat{p}_{23}(s = 365, t)$ from the landmark Aalen-Johansen estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line) (Second row). Colon cancer study.

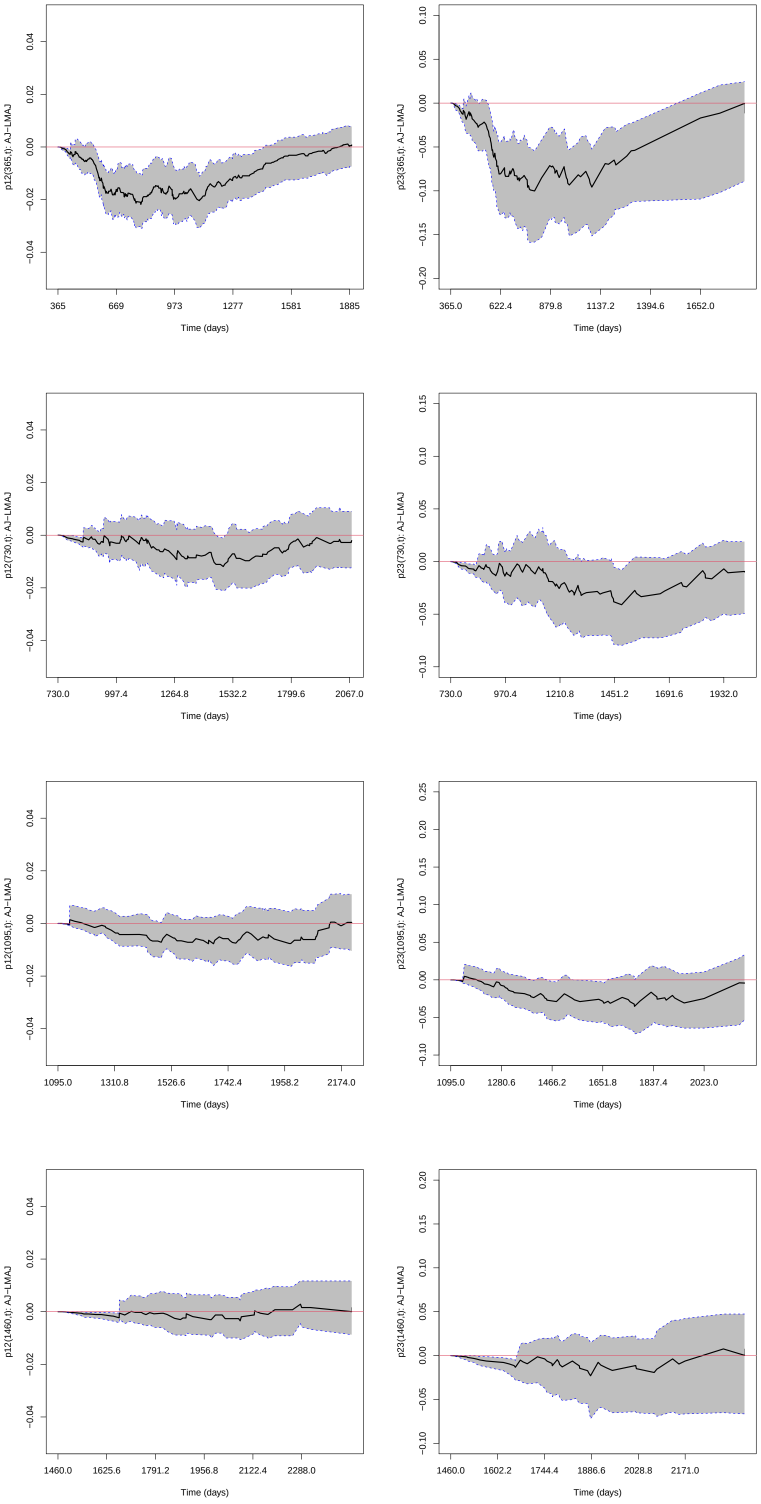


FIGURE 6.5: Local graphical test for the Markov condition, for *s* equal to 1, 2, 3 and 4 years since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (`LM`). Colon cancer data.

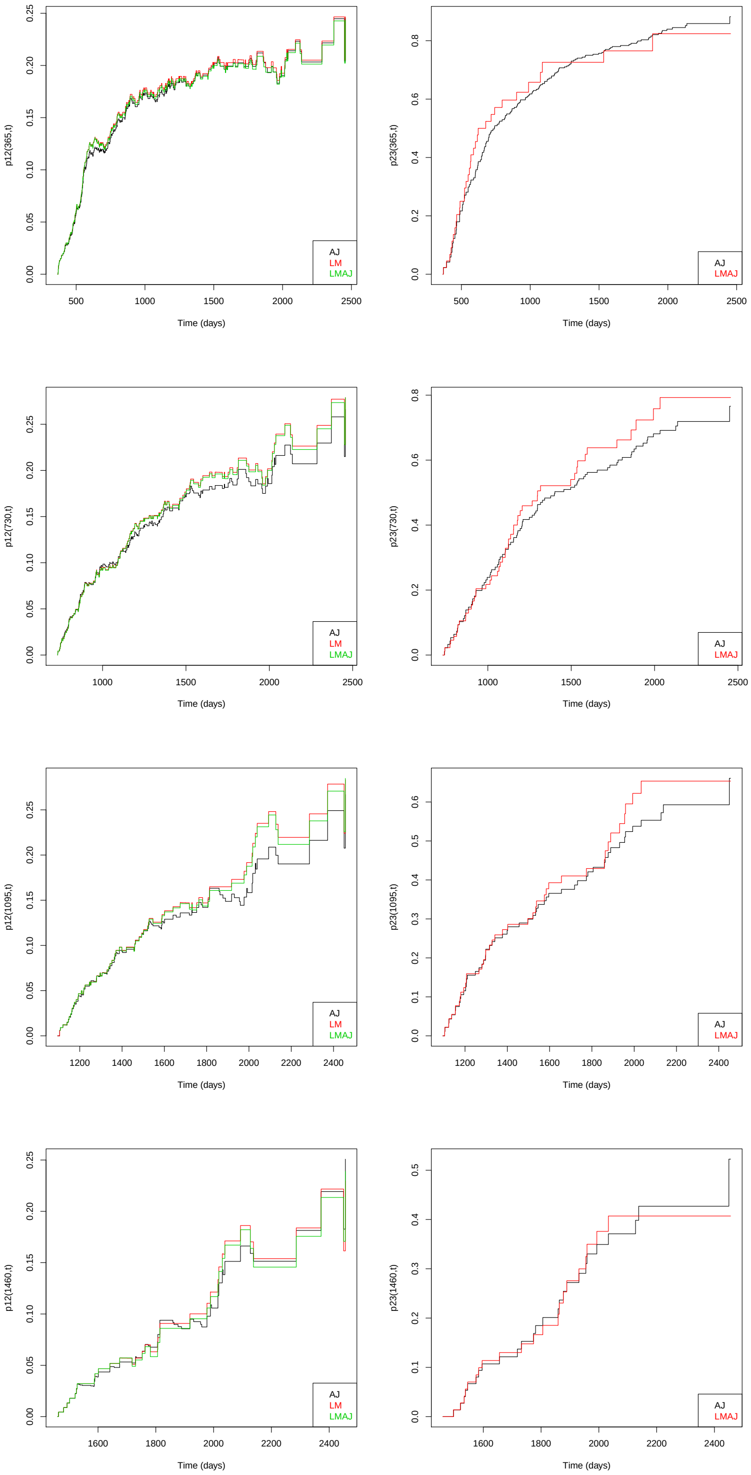


FIGURE 6.6: Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark and landmark Aalen-Johansen), for s equal to 1, 2, 3 and 4 years since entry in study. Breast cancer data.

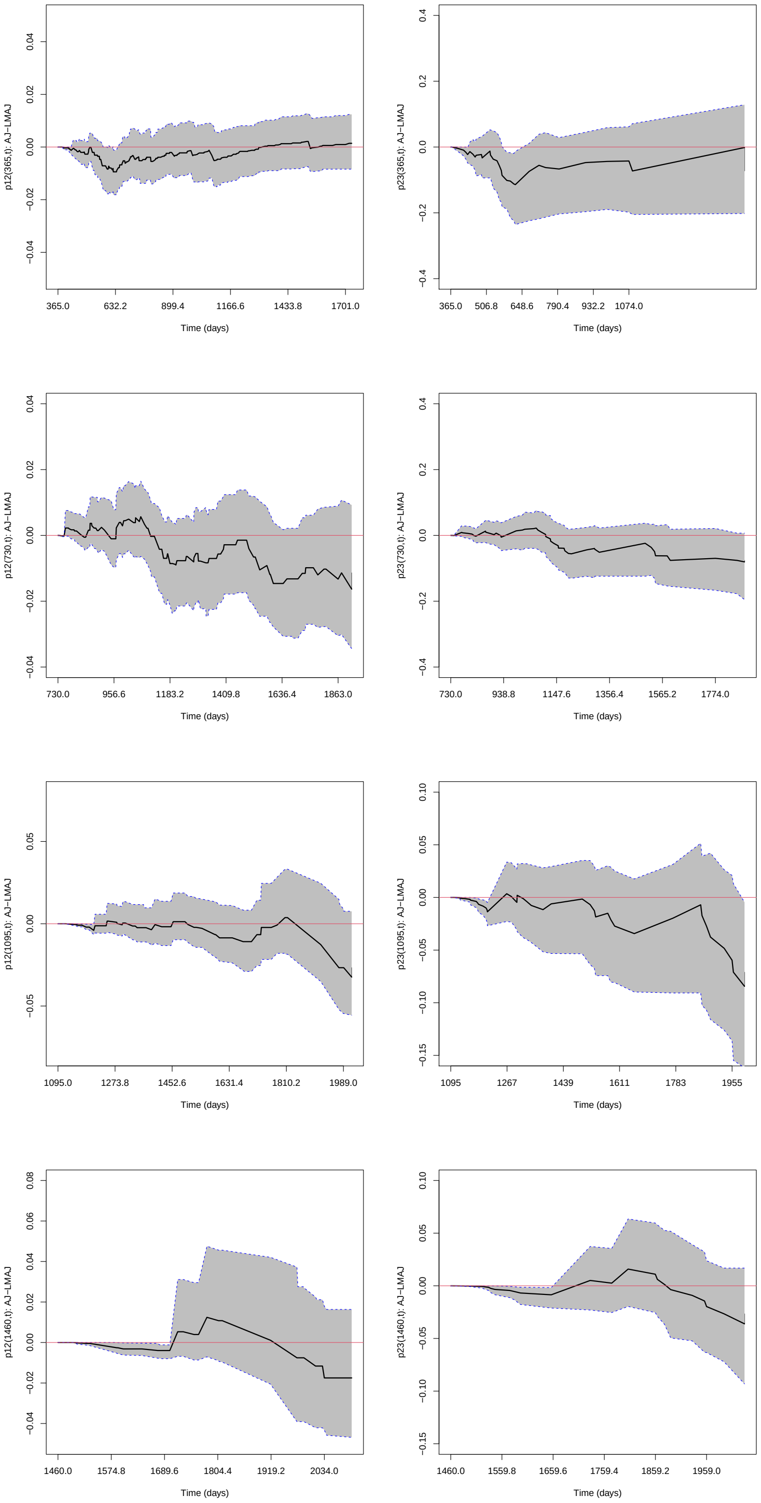


FIGURE 6.7: Local graphical test for the Markov condition, for s equal to 1, 2, 3 and 4 years since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (LM). Breast cancer data.

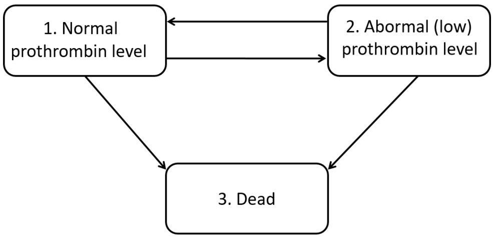

1. Normal prothrombin level
2. Abormal (low) prothrombin level
3. Dead

FIGURE 6.8: The reversible illness-death model for patients with liver cirrhosis.

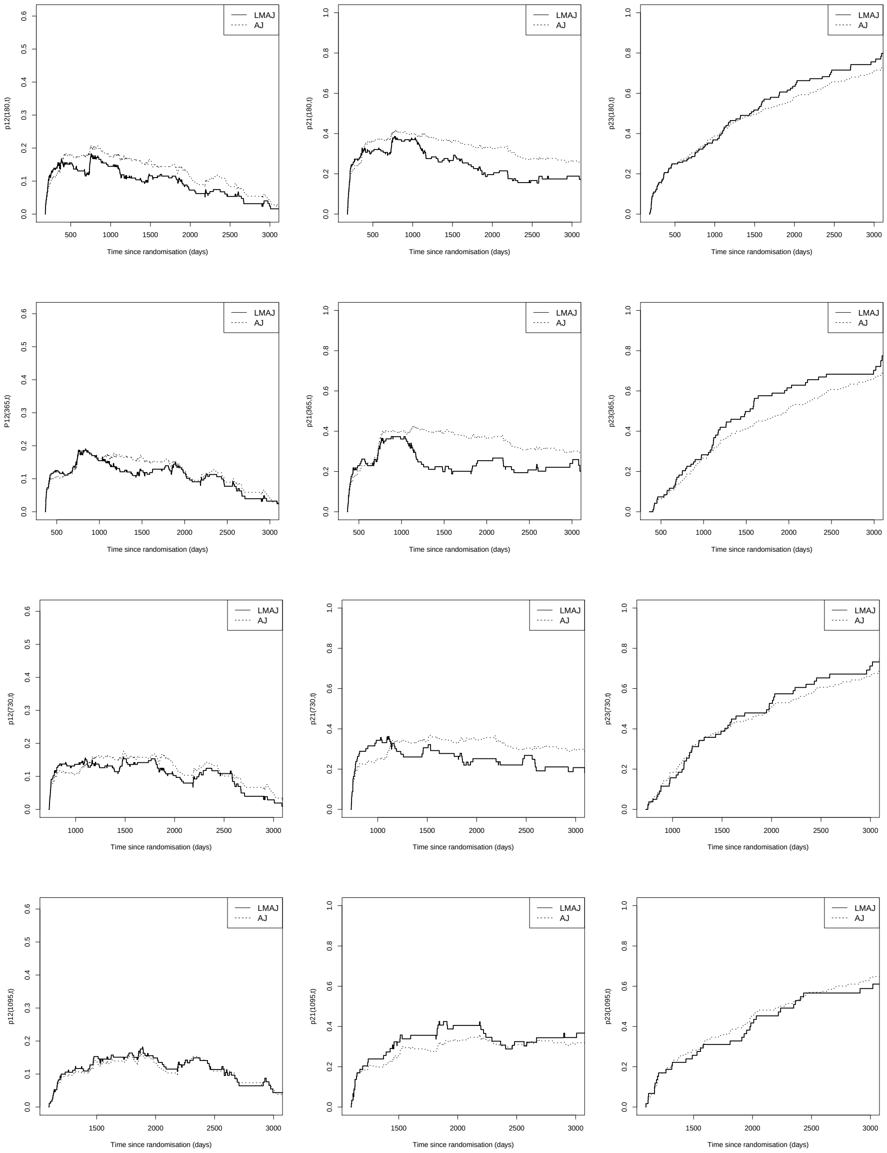


FIGURE 6.9: Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark Aalen-Johansen), for some *s* equal to 180, 365, 730 and 1095 days since entry in study. Liver cirrhosis data.

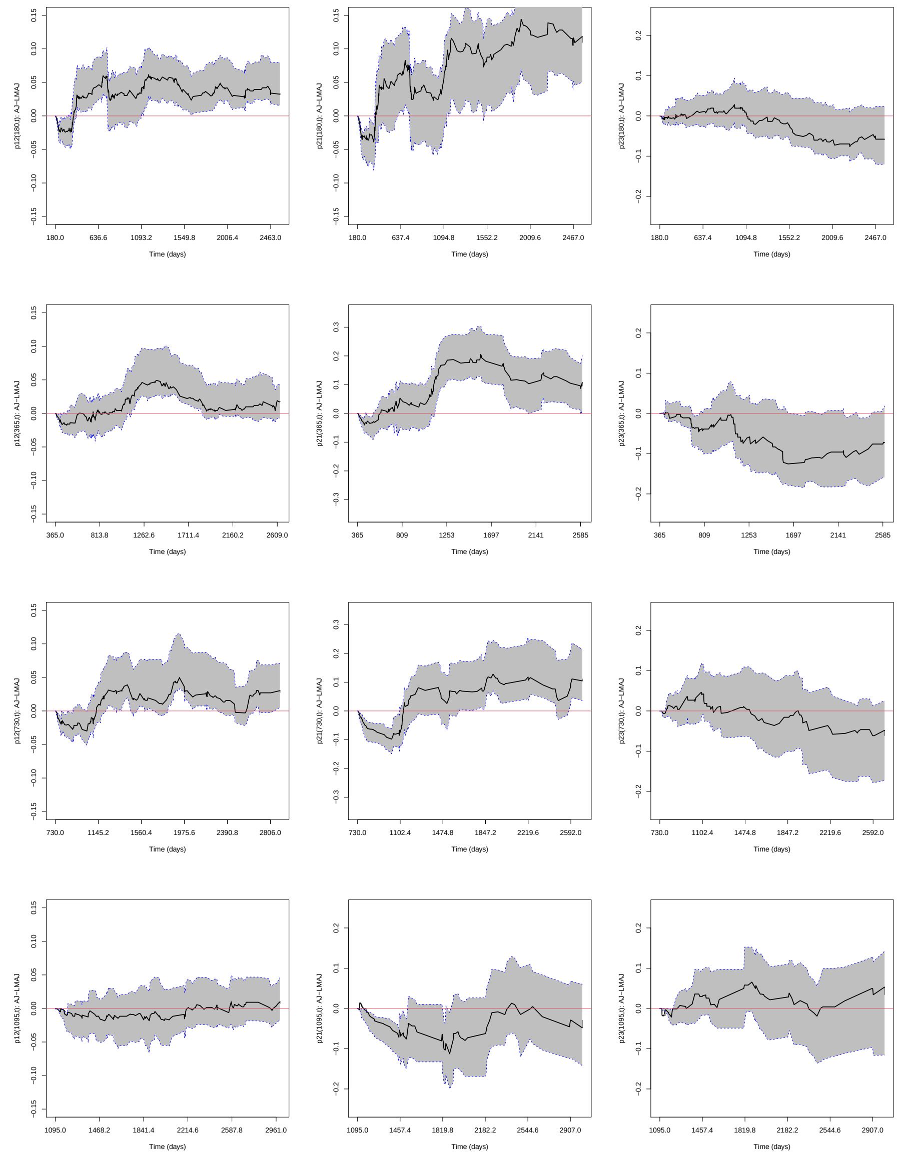


FIGURE 6.10: Local graphical test for the Markov condition, for s equal to 180, 365, 730 and 1095 days since entry in study since entry in study. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (LMAJ). Liver cirrhosis data.

6.5 Discussion

The Markov assumption is commonly used to analyze multi-state survival data. Therefore, goodness-of-fit tests for the Markov assumption are crucial in these models. Traditionally, this assumption is tested including covariates depending on the history on the modeling process. The comparison between estimated transition probabilities is the basis to introduce two formal local tests for the Markov assumption. The new methods are based on measuring the discrepancy of the Aalen-Johansen estimator which gives consistent estimators in Markov processes, and recent approaches that do not rely on this assumption. A log-rank test is used on specific transitions to check if the Markov assumption holds. A second method is proposed in this chapter in which the test statistic is based on the difference of the areas under the two curves. We note that alternative test statistics could also have been considered such as those based on the absolute differences or squared differences between the Aalen-Johansen and the landmark estimators that would lead to a kolmogorov smirnov or a Cramer-von Mises-type test statistic, respectively.

Simulation results reveals that the two methods perform similarly revealing high power to detect a failure of the Markov condition. The simulation results and the results obtained through real medical data analysis suggest that the second approach may be a good alternative to the existing methods. The use of the graphical local tests based on the discrepancy between estimated curves of the transition probabilities, proposed here, are recommended to confirm the conclusions obtained from the application of this formal local test. In general, the two curves may cross at mid time points when the process is indeed Markovian (and the two curves are similar). If the process is not Markovian, then it is expected that the two curves only cross at earlier time points or at higher time points (at the right tail). Nevertheless, it is wise to start the analysis with a graphical test in particular to identify possible situations in which the process is indeed not Markovian and the two curves cross at mid time points. In such cases the usage of a different test statistics (e.g. based on a squared difference) should be also analyzed in future research investigation.

The use of local tests is recommended whenever the interest is focused on the estimation of the transition probabilities and, in particular, to decide which estimator is the most appropriate to use: the Aalen-Johansen estimator or a robust estimator. The use of the proposed local test is advised for each transition probability $p_{hj}(s,t)$ $(h > 1)$, and the use of the robust Markov-free estimator when faced of evidences against Markovianity. This procedure may be followed for a general multi-state model.

A global test, such as the test proposed here, might be preferable for regression purposes. To this end, a common simplifying strategy is to decouple the whole process into various survival models by fitting separate intensities to all permitted transitions using semiparametric Cox proportional hazard regression models, while making appropriate adjustments to the risk set. The most common models are characterized through one of the two model assumptions that can be made about the dependence of the transition intensities and time. The transition intensities may be modeled using separated Cox models assuming the process to be Markovian (also known as the clock forward modeling approach). When the test rejects the Markov assumption, a possible alternative is to use a semi-Markov Cox model in which the future of the process does not depend on the current time but rather on the duration in the current state. Both models can be easily implemented using standard software such as the R packages `survidm` or `mstate`. To decide the appropriate modeling approach, the global test should be used to all transitions depending on history.

The global test proposed is obtained through the combination of the results from local tests over different times. Simulation results show that the proposed global test may be much more powerful than the standard parametric method based on the proportional hazard specification which relies on a prior model specification that may fail in practice. The proposed methods can be used in general multi-state models.

Discrete covariates can be included in the proposed methods by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. To account for the effect of continuous covariates, one can consider estimators of the transition probabilities conditional on covariates. One standard method is to consider estimators based on a Cox's model fitted marginally to each type of transitions, with the corresponding baseline hazard function estimated by the Breslow's method.

All the proposed methods in this chapter were implemented using the R language and release as a package, called `markovMSM`, which is available at the CRAN repository at `https://cran.r-project.org/web/packages/markovMSM` (Soutinho and Meira-Machado (2021) [119]). A detailed description of the main funcionalities is also presented as supplementary material [A.3] to this thesis and makes part of a paper submitted for publication [B].

Chapter 7

survidm: An R package for Inference and Prediction in an Illness-Death Model

Multi-state models are a useful way of describing a process in which an individual moves through a number of finite states in continuous time. The illness-death model plays a central role in the theory and practice of these models, describing the dynamics of healthy subjects who may move to an intermediate ″diseased″ state before entering into a terminal absorbing state. In these models, one important goal is the modeling of transition rates which is usually done by studying the relationship between covariates and disease evolution. However, biomedical researchers are also interested in reporting other interpretable results in a simple and summarized manner. These include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, and the sojourn time distributions. The development of `survidm` package has been motivated by recent contributions that provide answers to all these topics. The current version of the package provides seven different approaches to estimate the transition probabilities, two methods for the sojourn distributions and two approaches for the cumulative incidence functions. In addition, these probabilities can also be estimated conditionally on covariate measures. The package also allows the user to perform multi-state regression where the estimation of the covariate effects is achieved using Cox regression in which different effects of the covariates are assumed for different transitions.

The contents of this chapter are mainly based on the paper published in *R Journal* by Soutinho, Sestelo and Meira-Machado (2021) [120].

7.1 Introduction

Several researchers have recently developed software for multi-state survival analysis. A comprehensive list of the available packages in the Comprehensive R Archive Network (CRAN) can be seen in the CRAN task view 'Survival Analysis' (Allignol and Latouche (2019) [57]). In R, several packages provide functions for estimating the transition probabilities (e.g., the package `p3state.msm` (Meira-Machado and Roca-Pardiñas (2011) [121]), `TPmsm` (Araújo, Roca-Pardiñas and Meira-Machado (2014) [122]), `etm` (Allignol, Schumacher and Beyersmann (2011) [123]), `mstate` (de Wreede, Fiocco and Putter (2011) [124]) and `TP.idm` (Balboa and de Uña-Álvarez (2018) [125]), but none implements all the methods addressed by `survidm` which includes all newly developed methods based on the subsampling approach (see de Uña-Álvarez and Meira-Machado (2015) [69] and references therein). In addition, not all allow the users to obtain estimates of the transition probabilities conditional to covariates. The `cmprsk` and the `timereg` R packages can be used to estimate the cumulative incidence functions in a competing risks model. The package `survival` (via `survfit` and `coxph` functions) can also be used for competing risks data. The `msSurv` can be used to estimate the state occupation probabilities and the sojourn distributions for multi-state models subject to right-censoring (possibly state-dependent) and left-truncation. The package also provides matrices of transition probabilities between any two states. However, none of the available software provides an encompassing package which can be used to estimate all these quantities. Finally, the use of different packages to estimate these quantities separately is rather difficult because each of the current programs requests its own data structure.

This chapter introduces `survidm` (version 1.3.2, available from the Comprehensive R Archive Network at `https://cran.r-project.org/web/packages/survidm/`), a software application for R which performs inference in a progressive illness-death model. It describes the capabilities of the program for estimating semiparametric regression models and for implementing nonparametric estimators for all quantities mentioned above. In the remainder of this chapter we provide a brief introduction of the methodological background which most of them were already described in previous chapters (7.2). Following, a detailed description of the package is presented and its usage is illustrated through the analysis of a real data set (7.3). Finally, the last section contains the main conclusions of the package are presented in Section 7.4.

7.2 Methodology background

The mathematical background underlying the `survidm` package is briefly introduced in this section. A more detailed explanation of the concepts can be found in the previously chapters of this thesis. To be specific, the importance of the intensity rates in multi-state models to identify the effect of the different predictors into the outcome and the way of these quantities are modeled can be found in sections 1.3, 3.2 and 6.2. In the package `survidm` the inference of the transitions intensities is restricted to two semiparametric multi-state models (Cox models assuming the process to be Markovian and semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state), but other models are possible for the analysis of multi-state survival data. For example, time-homogeneous Markov models and model with piecewise constant intensities are implemented in the msm R package (Jackson (2011) [126]). Aalen additive model (Aalen, Borgan and Fekjaer (2001) [127]) and accelerated failure time models (Wei (1992) [128]) are another class of regression models that can be an alternative to the Cox proportional hazards model. The estimation of the transition probabilities also plays a very important role in the inference in multi-state models since allows predictions of the clinical prognosis of a patient across the evolution of the disease. The definition of this conditional probabilities and different methods for estimating the transition probabilities, with also a historical perspective, can be seen in the sections 1.3, 3.2, 3.3 and 4.2.

To estimate both these quantities is the major importance to check the Markov assumption, the reasons for checking and consequences in case of failure of this assumption are presented in the sections 1.3, 3.2, 3.3, 3.6, 4.1, 4.6 and 6.1. Global and local test are also proposed in Section 6.2.

Another quantity of interest in multi-state modeling is the cause-specific cumulative incidence function, as defined by Kalbfleisch and Prentice (1980) [50]. In the illness-death model the cumulative incidence of the illness is of particular interest. It represents the probability of an individual being or having been diseased at time t. One possible estimator, proposed by Geskus (2011) [129], is obtained by applying the Nelson-Aalen estimator and the product-limit estimator of survival. This estimator can also be expressed in terms of the Kaplan-Meier weights of the distribution of Z, the sojourn time in State 0, as introduced in the paper by Meira-Machado and Sestelo (2019) [16]. A modification of this estimator based on presmoothing can be introduced to reduce its variability. Both methods are implemented in the `survidm` package. Estimation methods for the cumulative incidence function conditionally on covariate measures based on local constant (Nadaraya-Watson) regression are also implemented in the package.

The estimation of the marginal distributions in multi-state modeling is an interesting topic too. In the context of the illness-death model, if the independence assumption between the censoring variable C and the vector of times (Z, T) is assumed, the marginal distribution of the sojourn time in State 0, Z, can be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{Z}_i, \Delta_{1i})$'s. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{T}_i, \Delta_i)$'s. However, the estimation of the marginal distribution of the sojourn time in State 1 is not such a simple issue. Nonparametric estimates for this marginal distribution allowing for state and path-dependent censoring were proposed by Satten and Datta (2002) [130].

7.3 survidm in practice

This section introduces an overview of how the package is structured. This software enables both numerical and graphical outputs to be displayed for all methods described in the previous section. It is intended to be used with the R statistical program (R Core Team (2019) [86]). The package is composed of 17 functions that allow users to obtain estimates for all proposed methods. Details on the usage of the functions (described in Table 7.1) can be obtained with the corresponding help pages.

It should be noted that to implement the methods described in the methodology section, one needs the following variables: `time1`, `event1`, `Stime`, and `event`. Covariates can also be included. The variable `time1` represents the sojourn time in State 0 and `Stime` the total time, whereas `event1` and `event` are the respective censoring indicators. This means that `event1` will take the value 1 if the subject leaves State 0 and 0 otherwise; `event` takes value 1 if the subject reaches State 2 and 0 otherwise.

For illustration, we apply the proposed methods to data from a large clinical trial on Duke's stage III patients affected by colon cancer, that underwent a curative surgery for colorectal cancer [74]. This data set is freely available as part of the R `survival` package. The data is also available as part of the R package `survidm`. Besides the two

Function	Description
survIDM	Create a survIDM object.
coxidm	Fits proportional hazards regression models for each transition.
tprob	Estimation of the transition probabilities.
CIF	Estimation of the cumulative incidence functions.
sojourn	Nonparametric estimation of the sojourn distribution in the intermediate state.
autoplot.survIDM	Visualization of survIDM objects with ggplot2 and plotly graphics.
plot.survIDM	Plot for an object of class survIDM.
print.survIDM	Print for an object of class survIDM.
summary.survIDM	Summary for an object of class survIDM.
nevents	Counts the number of observed transitions in the multi-state model.
markov.test	Performs a test for the Markov assumption.
KM	Computes the Kaplan-Meier product-limit of survival.
PKM	Computes the presmoothed Kaplan-Meier product-limit of survival.
Beran	Computes the conditional survival probability of the response, given the covariate under random censoring.
KMW	Returns a vector with the Kaplan-Meier weights.
PKMW	Returns a vector with the presmoothed Kaplan-Meier weights.
LLW	Returns a vector with the local linear weights.
NWW	Returns a vector with the Nadaraya-Watson weights.

TABLE 7.1: Summary of functions in the survidm package.

event times (disease-free survival time and death time) and the corresponding indicator statuses, a vector of covariates including rx (treatment: Obs(ervation), Lev(amisole), Lev(amisole)+5FU), sex (1 - male), age (years), nodes (number of lymph nodes with detectable cancer), surge (time from surgery to registration: 0 = short, 1 = long), adhere (adherence to nearby organs) are also available. The covariate ‘recurrence’ is the only time-dependent covariate, while the other covariates included are fixed. Recurrence can be considered as an intermediate transient state and modeled using the progressive illness-death model with transient states ‘alive and disease-free’ and ‘alive with recurrence’, and the absorbing state ‘dead’. In the following, we will demonstrate the package capabilities using this data. Below is an excerpt of the data.frame with one row per individual. Individuals were chosen in order to represent all possible combinations of movements among the three states.

```
> library("survidm")
> data(colonIDM)
> colonIDM[c(1:2,16,21),1:7]
```

```
   time1 event1 Stime event      rx sex age
1    968      1  1521     1 Lev+5FU   1  43
2   3087      0  3087     0 Lev+5FU   1  63
16  1323      1  3214     0     Obs   1  68
21  2789      1  2789     1     Obs   1  64
```

Individual represented in the first line experienced a recurrence of the tumor and have died. In such cases, `event1` and `event` are equal to 1 and `time1` are different of `Stime`. Individual represented in line 2 remain alive and without recurrence at the end of follow-up (`event1` = 0 and `event` = 0). Individual represented in line 16 of the original data set, with `event1 = 1` and `event = 0`, corresponds to an individual with an observed recurrence that remains alive at the end of the follow-up. Note that in this case, the disease-free survival time is equal to the death time (`time1` = `Stime`). Finally, individual represented in line 21 of the original data set has died without observing a recurrence. We note that `event1 = 1` and `event = 0` correspond to individuals with an observed recurrence that remain alive at the end of the follow-up.

Of the total of 929 patients, 468 developed a recurrence, and among these 414 died, 38 patients died without developing a recurrence. A summary of the data with the number of the undergoing transitions can be obtained through the `nevents` function. The colums of the data set must include at least the four columns named `time1`, `event1`, `Stime`, and `event` according to the requirements of the `survIDM` function presented in the help file. Parameter `state.names` enables to change the default values of states, 'healthy', 'illness', and 'death'.

```
> nevents(with(colonIDM, survIDM(time1, event1, Stime, event)),
          state.names = c("healthy", "recurrence", "death"))
           healthy recurrence death
healthy        423        468    38
recurrence       0         54   414
death            0          0   452
```

7.3.1 Regression models for transitions intensities

To relate the individual characteristics to the intensity rates, semiparametric multi-state regression models are used. Specifically, separated Cox models assuming the process to be Markovian (i.e., the transition intensities only depend on the history of the process through the current state) or using a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. Therefore, of practical interest to determine whether the Markov property holds within a particular data set to determine whether a Markov model or a semi-Markov model is more appropriate.

7.3.2 The Markov assumption

The Markov assumption may be checked, among others, by including covariates depending on the history. For the progressive illness-death model, the Markov assumption is only relevant for mortality transition after recurrence. We can examine whether the time spent in the initial state ”Alive and disease-free” (i.e., the past) is important in the transition from the recurrence state to death (i.e., the future). For doing that, let Z be the time spent in State 0 and t the current time. Fitting a model $\alpha_{12}(t;Z) = \alpha_{12,0}(t)exp\{\beta Z\}$, we now need to test the null hypothesis, $H_0 : \beta = 0$, against the general alternative, $H_1 : \beta \neq 0$. This would assess the assumption that the transition rate from the disease state into death is unaffected by the time spent in the previous state.

```
> library(survival)
> fit <- coxph(Surv(time1, Stime, event) ~ time1, data = colonIDM,
         subset=c(time1 < Stime))
> fit
           coef  exp(coef)   se(coef)      z     p
time1 -0.0002475  0.9997526  0.0001737 -1.424 0.154

Likelihood ratio test=2.04  on 1 df, p=0.1533
n= 468, number of events= 414
```

Following this procedure, we verified that the effect of time spent in State 0 reported a p-value of 0.154 (regression coefficient: - 0.0002475), revealing no evidence against the Markov model for the colon data. Results from this test can also be obtained through the function `markov.test`, which has an output fairly similar to those obtained from `coxph` function.

```
> mk <- markov.test(survIDM(time1, event1, Stime, event) ~ 1,
        data = colonIDM)
> mk
```

Since there is no evidence on the lack of Markovianity, a multi-state Markov regression model based on the Cox model can be fitted through the following input command:

```
> fit.cmm <- coxidm(survIDM(time1, event1, Stime, event) ~ rx + sex + age +
                    nodes + surg + adhere, data = colonIDM)

> summary(fit.cmm)

Cox Markov Model: transition 0 -> 1

                  coef exp(coef) lower 0.95 upper 0.95     Pr(>|z|)
rxLev     -0.061251858 0.9405863  0.7596976  1.1645457 5.740592e-01
rxLev+5FU -0.515170844 0.5973985  0.4713678  0.7571264 2.031682e-05
sex       -0.149177218 0.8614164  0.7160077  1.0363552 1.137849e-01
age       -0.004669254 0.9953416  0.9876802  1.0030625 2.362711e-01
nodes      0.083943790 1.0875678  1.0686993  1.1067694 5.418662e-21
surg       0.251798521 1.2863368  1.0509673  1.5744186 1.460249e-02
adhere     0.296839791 1.3455997  1.0551768  1.7159575 1.671466e-02

Cox Markov Model: transition 0 -> 2

                 coef exp(coef) lower 0.95 upper 0.95     Pr(>|z|)
rxLev     -0.29152482 0.7471235  0.3271685   1.706135 4.889711e-01
rxLev+5FU -0.11211853 0.8939383  0.4220165   1.893589 7.697006e-01
sex        0.39293182 1.4813174  0.7641923   2.871399 2.445966e-01
age        0.08422764 1.0878765  1.0476871   1.129608 1.157046e-05
nodes      0.07538428 1.0782984  0.9895116   1.175052 8.552937e-02
surg       0.41564547 1.5153485  0.7703441   2.980851 2.285509e-01
adhere     0.05435239 1.0558566  0.4377875   2.546517 9.036879e-01



Cox Markov Model: transition 1 -> 2

                  coef exp(coef) lower 0.95 upper 0.95     Pr(>|z|)
rxLev      0.068953592  1.071386  0.8533466   1.345138 5.525534e-01
```

```
rxLev+5FU   0.327043851  1.386862  1.0741245   1.790656 1.212756e-02
sex         0.214094887  1.238740  1.0138220   1.513557 3.623833e-02
age         0.009342474  1.009386  1.0014760   1.017359 1.994502e-02
nodes       0.046061552  1.047139  1.0249376   1.069821 2.522475e-05
surg       -0.012258877  0.987816  0.7944594   1.228232 9.121722e-01
adhere      0.137708158  1.147641  0.8851963   1.487895 2.985854e-01
```

The transition intensities characterize the hazard for movement from one state to another, revealing how the different covariates affect the various permitted transitions. The results indicate that none of the covariates were found to have a strong effect on all three transitions. Save for covariates *age* and *sex*, all the remaining predictors were considered important for recurrence transition. Interestingly, *age* displayed a strong linear effect on mortality transition without recurrence, whereas all the other covariates failed to show relevant association on this transition. Finally, save for covariates *surg* and *adhere*, all the remaining predictors were considered important for the mortality transition after recurrence. The `coxidm` function also returns the analysis of the deviance for each Cox model. In this case, only an overall p-value is presented for categorical variables. To obtain the outputs, we have to indicate `type`='anova' in `summary` function.

```

> summary(fit.cmm,type = 'anova')

Cox Markov Model: transition 0 -> 1

        loglik   Chisq Df Pr(>|Chi|)
NULL   -2954.2
rx     -2941.8 24.6964  2  4.338e-06 ***
sex    -2941.0  1.6402  1    0.20030
age    -2939.8  2.3435  1    0.12581
nodes  -2909.0 61.6050  1  4.198e-15 ***
surg   -2906.2  5.7134  1    0.01684 *
adhere -2903.5  5.3740  1    0.02044 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Cox Markov Model: transition 0 -> 2
```

```

        loglik   Chisq Df Pr(>|Chi|)
NULL   -231.79
rx     -231.54  0.4938  2     0.7812
sex    -231.04  1.0065  1     0.3158
age    -219.26 23.5445  1  1.221e-06 ***
nodes  -218.04  2.4536  1     0.1173
surg   -217.35  1.3830  1     0.2396
adhere -217.34  0.0145  1     0.9043
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1


Cox Markov Model: transition 1 -> 2


        loglik   Chisq Df Pr(>|Chi|)
NULL   -1897.5
rx     -1895.0  4.8864  2  0.0868804 .
sex    -1892.8  4.3995  1  0.0359501 *
age    -1890.8  4.0650  1  0.0437799 *
nodes  -1883.4 14.7205  1  0.0001247 ***
surg   -1883.4  0.0090  1  0.9242629
adhere -1882.9  1.0505  1  0.3054007
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The effect of the continuous covariates on the log hazards is often assumed to have a linear functional form in all intensities. To introduce flexibility into the Cox Markov model, several smoothing methods may be applied, but P-splines (Eilers and Marx (1996) [12]) are being most frequently considered in this context. Results showed of a strong nonlinear effect for nodes (checked through a formal test) when using a Cox model on the recurrence transition. Figure 7.1 returns a centered set of predictions on a log hazard scale. The average predicted value is zero with a mean value of nodes as the reference (see the vignette ’Splines, plots, and interactions’ in Therneau (2021) [131]). The main curve depicts the smooth curve for nodes on a log hazard scale, indicating that the risk of recurrence increases rapidly until about 6 nodes. The apparent decrease after 23 nodes is not significant due to the wide confidence intervals.

```
> library(ggplot2)
> library(plotly)

> fit2.cmm <- coxidm(survIDM(time1, event1, Stime, event) ~ rx + sex + age +
                  pspline(nodes) + surg + adhere, data = colonIDM)


> d<-data.frame(x=fit2.cmm$term01$nodes$x, y=fit2.cmm$term01$nodes$y,
              y1=fit2.cmm$term01$nodes$y-1.96*fit2.cmm$term01$nodes$se,
              y2=fit2.cmm$term01$nodes$y+1.96*fit2.cmm$term01$nodes$se)

> nonlinear<-ggplot(d, aes(x,y))+theme_bw()+labs(x = "nodes") +
                 labs(y = "Partial for pspline(nodes)")+
                 geom_ribbon(aes(ymin=y1,ymax=y2),fill='gray92',alpha=0.9)+
                 geom_line(aes(x,y))+
                 geom_line(color=1,size=1)

> ggplotly(nonlinear)
```

The proportional hazards assumption can be tested formally using the summary function. The output can be obtained putting type='ph' in summary function.

```
> summary(fit2.cmm, type = 'ph')

Cox Markov Model: transition 0 -> 1
Test the Proportional Hazards Assumption


                  chisq   df    p
rx             4.12e-01 2.00 0.81
sex            2.10e+00 1.00 0.15
age            9.37e-04 1.00 0.98
pspline(nodes) 7.60e+00 3.95 0.10
surg           1.97e+00 1.00 0.16
adhere         6.13e-01 1.00 0.43
GLOBAL         1.30e+01 9.94 0.22
```

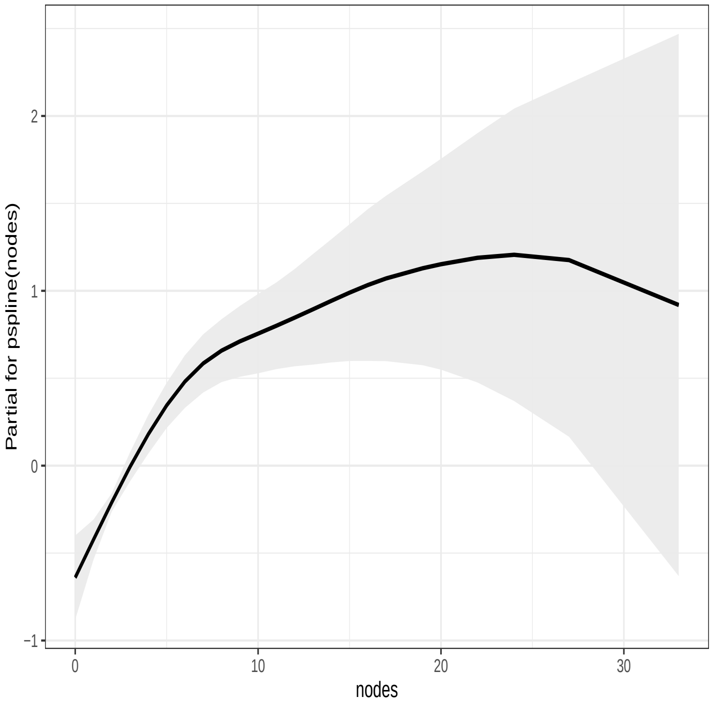


FIGURE 7.1: Predicted values of the smooth log hazard based on penalized splines (black line) with pointwise 95% confidence intervals obtained from the partial residuals for nodes (recurrence intensity), using the colon cancer data.

```


Cox Markov Model: transition 0 -> 2
Test the Proportional Hazards Assumption

                chisq   df    p
rx             1.6292 2.00 0.44
sex            0.0668 1.00 0.80
age            0.8396 1.00 0.36
pspline(nodes) 0.7859 4.00 0.94
surg           0.4955 1.00 0.48
adhere         2.3606 1.00 0.12
GLOBAL         6.1424 9.99 0.80


Cox Markov Model: transition 1 -> 2
Test the Proportional Hazards Assumption
```

```
                   chisq    df    p
rx               5.03913  1.99 0.08
sex              0.02204  1.00 0.88
age              0.73628  1.00 0.39
pspline(nodes)  4.25500  4.09 0.39
surg             2.02427  1.00 0.15
adhere           0.00177  1.00 0.97
GLOBAL          13.19170 10.08 0.22
```

A semi-Markov model could be obtained by including the argument `semiMarkov = TRUE` in the `coxidm` function.

7.3.3 Occupation probabilities and transition probabilities

The occupation probabilities and the transition probabilities are key quantities of interest in multi-state models. They offer interpretable results in a simple and summarized manner.

Estimates and plots of the transition probabilities can be obtained using the `tprob` function. The default method is the Aalen-Johansen estimator (`AJ`) which assumes the process to be Markovian. The presmoothed version of the Aalen-Johansen estimator (`PAJ`) also assumes the process to be Markovian while the remaining methods (`LIDA`, `LM`, `PLM`, `LMAJ`, and `PLMAJ`) are free of the Markov condition.

When one is confident of the Markov assumption, the Aalen-Johansen is preferred over the non-Markovian estimators since it reports a smaller variance in estimation. Estimates and plot for the Aalen-Johansen method can be obtained through the following input commands:

```
> tpAJ <- tprob(survIDM(time1, event1, Stime, event) ~ 1, s = 365,
                method = "AJ", conf = TRUE, data = colonIDM)

> summary(tpAJ, times=365*2:6)

Estimation of pij(s=365,t)

    t        00        01        02        11        12
  730 0.7966309 0.1300071 0.0733620 0.4686360 0.5313640
 1095 0.7192603 0.1224599 0.1582799 0.2533822 0.7466178
 1460 0.6805333 0.0884287 0.2310380 0.1335300 0.8664700
 1825 0.6444157 0.0859123 0.2696720 0.0932851 0.9067149
 2190 0.6131533 0.0774912 0.3093556 0.0632835 0.9367165

2.5%

    t        00        01        02        11        12
  730 0.7673408 0.1093487 0.0589350 0.4105298 0.4728114
 1095 0.6867036 0.1026150 0.1354061 0.2105314 0.7011204
 1460 0.6468259 0.0714743 0.2030840 0.1047501 0.8346547
 1825 0.6098804 0.0688614 0.2396632 0.0708282 0.8813846
 2190 0.5780541 0.0612090 0.2777007 0.0464018 0.9172849

97.5%

    t        00        01        02        11        12
  730 0.8270390 0.1545683 0.0913208 0.5349666 0.5971676
 1095 0.7533604 0.1461425 0.1850177 0.3049547 0.7950677
 1460 0.7159973 0.1094050 0.2628397 0.1702170 0.8994981
 1825 0.6809066 0.1071852 0.3034384 0.1228620 0.9327733
 2190 0.6503836 0.0981045 0.3446188 0.0863070 0.9565597

> autoplot(tpAJ)
```

Besides being consistent regardless the Markov condition, the landmark non-Markov estimators (`LM`, `PLM`, `LMAJ`, and `PLMAJ`) can be preferable in many situations due to their greater accuracy. When comparing the original nonparametric landmark estimator (`LM`) and the Aalen-Johansen estimator, some discrepancies are observed for $t = 730$ and $t = 1095$ (2 and 3 years, respectively). In addition to the aforementioned discrepancy between the two estimates, the plots for the two methods (Figure 7.2) also show that the confidence bands are narrower in the case of the Aalen-Johansen, revealing less variability for this method.

```
> tpLM <- tprob(survIDM(time1, event1, Stime, event) ~ 1, s = 365,
               method = "LM", conf = TRUE, data = colonIDM)

> summary(tpLM, times=365*2:6)
```

```

Estimation of pij(s=365,t)

    t        00         01        02         11        12
  730 0.7966309 0.14750103 0.0558681 0.38815789 0.6118421
 1095 0.7192603 0.14320925 0.1375305 0.15789474 0.8421053
 1460 0.6805333 0.09446864 0.2249981 0.10526316 0.8947368
 1825 0.6444157 0.08583643 0.2697479 0.09210526 0.9078947
 2190 0.6131533 0.07465238 0.3121944 0.06432749 0.9356725

2.5%

    t        00         01        02         11        12
  730 0.7673274 0.12294665 0.0411836 0.31792669 0.5390734
 1095 0.6866872 0.12033558 0.1142137 0.10937624 0.7860868
 1460 0.6468058 0.07447488 0.1960521 0.06621973 0.8472552
 1825 0.6098421 0.06804756 0.2387239 0.05591405 0.8630680
 2190 0.5777125 0.05742370 0.2791810 0.03480413 0.8969820

97.5%

    t        00         01         02        11        12
  730 0.8270534 0.17695930 0.07578852 0.4739034 0.6944337
 1095 0.7533784 0.17043081 0.16560740 0.2279357 0.9021157
 1460 0.7160195 0.11982998 0.25821767 0.1673268 0.9448794
 1825 0.6809493 0.10827565 0.30480372 0.1517218 0.9550498
 2190 0.6507682 0.09705015 0.34911161 0.1188947 0.9760319

> autoplot(tpLM)
```

Since the landmark estimators of the transition probabilities are free of the Markov assumption, they can also be used to introduce such tests (at least in the scope of the illness-death model) by measuring their discrepancy to Markovian estimators. The function `markov.test` performs a local graphical test for the Markov condition. This graphical test is based on a PP-plot which compares the estimations reported by the Aalen-Johansen transition probabilities to their non-Markov counterparts. The corresponding plot for a

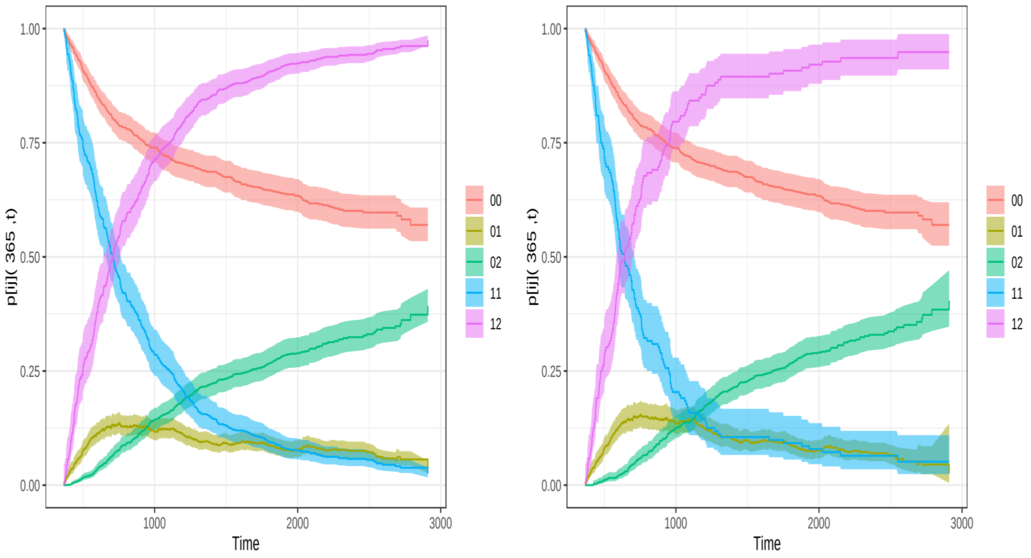


FIGURE 7.2: Transition probability estimates using the `AJ` (left hand side) and `LM` (right hand side) method, using the colon cancer data.

local test of Markovianity ($s = 365$) can be obtained through the following input command:

```
> mk <- markov.test(survIDM(time1, event1, Stime, event) ~ 1, s = 365,
        data = colonIDM)
> autoplot(mk)
```

The plot shown in Figure 7.3 compares the Aalen-Johansen estimator and the landmark non-Markovian estimator for $p_{01}(s,t)$, $p_{02}(s,t)$, and $p_{12}(s,t)$, for $s = 365$. Existing deviations of the plots with respect to the straight line $y = x$ reveals some evidence on the lack of Markovianity of the underlying process beyond one year after surgery. For further illustration, this figure jointly displays the landmark non-Markovian estimator and the Aalen-Johansen estimator for $p_{12}(s = 365, t)$. In this, plot the differences between both estimators are clearly seen. Thus, in principle, the application of the Aalen-Johansen method is not recommended here, due to possible biases.

The variability of the nonparametric landmark estimator (`LM`) may be successfully reduced using presmoothing ideas (Dikta (1998) [65]; Cao *et al.* (2005) [66]). The presmoothed landmark estimator is implemented in the same function through the method `PLM`. The same ideas can be used to reduce the variability of the Markovian Aalen-Johansen estimator and the (non-Markov) Landmark Aalen-Johansen estimator through methods `PAJ` and `PLMAJ`, respectively.

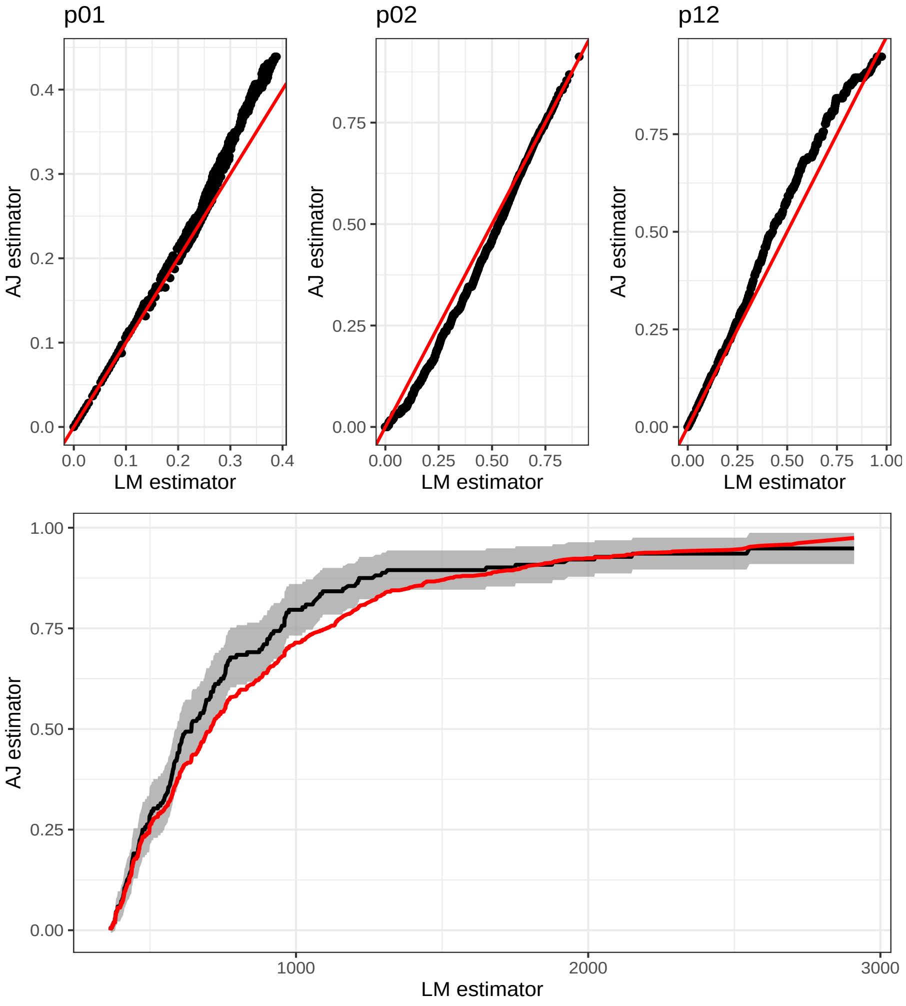


FIGURE 7.3: Graphical test for the Markov condition, $s = 365$. The second row shows the landmark (Markov-free) estimator with 95% pointwise confidence limits (black lines) and Aalen-Johansen estimator (red line) for the transition probability $p_{12}(365, t)$, using the colon cancer data.

The package `survidm` also allows for the computation of the above quantities conditional on covariates that are observed for an individual before the individual makes a particular transition of interest. For continuous covariates, one possible and flexible non-parametric approach is to consider local smoothing by means of kernel weights based on local constant (Nadaraya-Watson: `NW`) regression. This estimator is implemented in our package through function `tprob` using the `method = IPCW`. Below are the input commands to obtain the estimates of the transition probabilities at time $s = 365$ for an individual of 48 years old. For the bandwidth in the estimator, we use `dpik` function which is available from the R `KernSmooth` package. This is the data-based bandwidth selector of Wand & Jones (1997) [81].

```
> tpIPCW.age <- tprob(survIDM(time1, event1, Stime, event) ~ age, s = 365,
              method = "IPCW", z.value = 48, conf = FALSE, data = colonIDM,
              bw = "dpik", window = "gaussian", method.weights = "NW")

> summary(tpIPCW.age, time=365*2:6)

Estimation of pij(s=365,t)

    t        00        01         02         11        12
  730 0.7662208 0.1921290 0.04165012 0.28946129 0.7105387
 1095 0.7308496 0.1688189 0.10033149 0.12631010 0.8736899
 1460 0.6980293 0.1088373 0.19313342 0.05905711 0.9409429
 1825 0.6310625 0.1186104 0.25032706 0.05903929 0.9409607
 2190 0.6157095 0.1051797 0.27911080 0.04035816 0.9596418

> autoplot(tpIPCW.age)
```

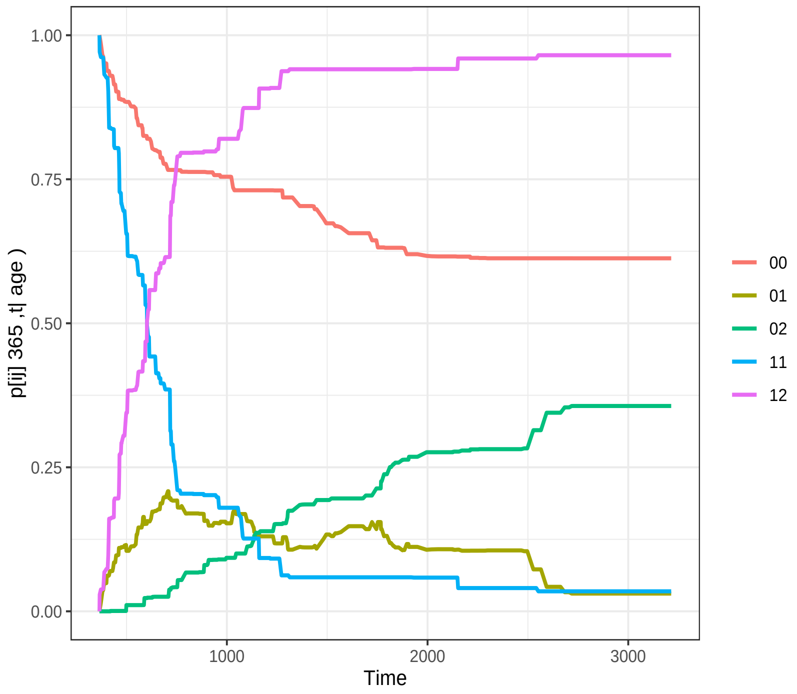


FIGURE 7.4: Conditional transition probabilities given that the subject is alive and disease-free at $s = 365$ days for a 48-years-old patient, using the colon cancer data.

The curves depicted in Figure 7.4, which are purely nonparametric, enable flexible modeling of the data providing flexible and robust effects of the covariate that can be used at least as a preliminary attempt, providing insights on the data being analyzed. Such methods can be used to capture nonstandard data features that may not be detected through parametric or semiparametric proposals. A general problem in multivariate nonparametric regression estimation is the so-called curse of dimensionality. In higher dimensions, the observations are sparsely distributed even for large sample sizes. Consequently, estimators based on local averaging (like those based on kernel smoothing) perform unsatisfactorily in this situation.

An alternative method is to consider estimators based on Cox's regression model (Cox (1972) [6]) fitted marginally to each transition with the corresponding baseline hazard function estimated by Breslow's method (Breslow (1972) [23]). The following input commands illustrate the use of the tprob function in this context:

```
> tp.breslow.age <- tprob(survIDM(time1, event1, Stime, event) ~ age, s = 365,
              method = "breslow", z.value = 48, conf = FALSE, data = colonIDM)

> summary(tp.breslow.age, time=365*2:6)

Estimation of pij(s=365,t)

    t        00          01          02          11         12
  730 0.7970855 0.15020199 0.05271253 0.37528949 0.6247105
 1095 0.7198657 0.14999685 0.13013746 0.14814634 0.8518537
 1460 0.6826444 0.10384005 0.21351550 0.09843946 0.9015605
 1825 0.6451532 0.09850122 0.25634562 0.08617378 0.9138262
 2190 0.6139465 0.08891388 0.29713961 0.06066618 0.9393338
```

Note that if the argument z.value is missing, then the tprob function computes the predicted conditional transition probabilities at the average values of the covariate. The Breslow method (based on the Cox regression model) is particularly well-suited to the setting with multiple covariates:

```
> tp.breslow <- tprob(survIDM(time1, event1, Stime, event) ~ rx + age + nodes, s
    = 365,
              method = "breslow", z.value = c('Obs', 50, 10), conf = FALSE, data
    = colonIDM)

```

```
> summary(tp.breslow, time=365*2:6)

Estimation of pij(s=365,t)

    t        00         01        02          11        12
  730 0.6423398 0.24905912 0.1086010 0.30017412 0.6998259
 1095 0.5222992 0.21890332 0.2587975 0.09465150 0.9053485
 1460 0.4680828 0.12787851 0.4040387 0.05433167 0.9456683
 1825 0.4181094 0.10712224 0.4747684 0.04519157 0.9548084
 2190 0.3762996 0.08424903 0.5394514 0.02685212 0.9731479
```

7.3.4 Cumulative Incidence Function

Another quantity of interest in multi-state modeling is the cause-specific cumulative incidence of the illness (recurrence). Function `CIF` can be used to obtain the nonparametric estimator of Geskus (2011) [129] (default method), which is equivalent to the classical Aalen-Johansen estimator. The corresponding presmoothed version (Meira-Machado and Sestelo, 2018) is also implemented through the argument `presmooth = TRUE`:

```
> cif <- CIF(survIDM(time1, event1, Stime, event) ~ 1, data = colonIDM,
           conf = TRUE)
> summary(cif, time=365*1:6)

Estimation of CIF(t)
    t       CIF
  365 0.2378902
  730 0.3844412
 1095 0.4372663
 1460 0.4620841
 1825 0.4859813
 2190 0.5032043

2.5%

    t       CIF
  365 0.2088267
  730 0.3509039
 1095 0.4038141
 1460 0.4296740
 1825 0.4540347
 2190 0.4697608

97.5%

    t       CIF
  365 0.2616792
  730 0.4103338
 1095 0.4666664
 1460 0.4900876
 1825 0.5161684
 2190 0.5319749

> autoplot(cif, ylim=c(0, 0.6), confcol = 2)
```

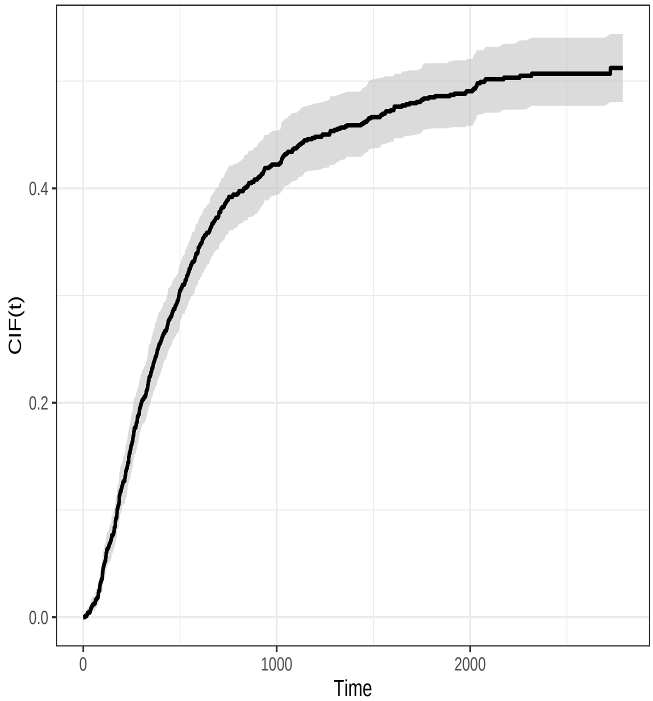


FIGURE 7.5: Cumulative incidence function in the recurrence state with 95% bootstrap confidence bands, using the colon cancer data.

Figure 7.5 depicts the estimates of cumulative incidence function for the recurrent state together with a 95% pointwise confidence bands based on simple bootstrap that resamples each datum with probability $1/n$. From this plot, it can be seen that individuals have a probability of recurrence higher than 50%. This cumulative probability is about 43% at three years after surgery.

Figure 7.6 depicts the estimates of the (conditional) cumulative incidence function for patients with 1 and 9 lymph nodes with detectable cancer. Curves depicted in this figure, which are purely nonparametric, indicate that patients with 9 lymph nodes with detectable cancer have a considerably higher probability of recurrence. The corresponding input commands are shown below:

```
> cif.1.nodes <- CIF(survIDM(time1, event1, Stime, event) ~ nodes, data =
    colonIDM,
           conf = FALSE, z.value = 1)
> cif.9.nodes <- CIF(survIDM(time1, event1, Stime, event) ~ nodes, data =
    colonIDM,
                    conf = FALSE, z.value = 9)

> d<-as.data.frame(cbind(rep(cif.1.nodes$est[,1],2),c(cif.1.nodes$est[,2],
                  cif.9.nodes$est[,2]), c(rep("1 nodes", length(cif.1.nodes$est
    [,1])),
                  rep("9 nodes", length(cif.1.nodes$est[,2])))))
> names(d)<-c('time','cif','type')

>  cif<-ggplot(d, aes(x=as.numeric(time), y=as.numeric(cif),group=factor(type),
                   color=factor(type)))+theme_bw()+labs(x = 'Time (days)',
                   y = 'CIF(t|nodes)')
> cif+geom_step(size=1)+ theme(legend.title=element_blank())
```

7.3.5 Sojourn distribution

Another interesting quantity is the sojourn time in each state. Estimates for the distribution function of the sojourn time in the recurrence state can be obtained using the estimator by Satten and Datta (2002) [130] through function sojourn.

```
> soj <- sojourn(survIDM(time1, event1, Stime, event) ~ 1,
              data = colonIDM, method = "Satten-Datta", conf = FALSE)
> summary(soj, time=365*1:6)
```

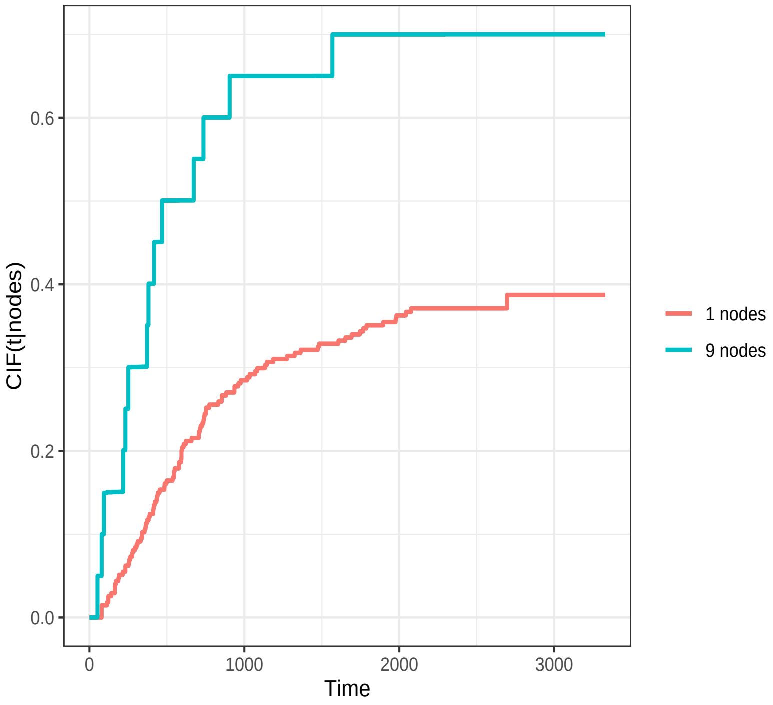


FIGURE 7.6: Conditional cumulative incidence function for the colon cancer data for `nodes = 1` and `nodes = 9`, using the colon cancer data.

```

Estimation of sojourn(t)

    t   sojourn
  365 0.4852424
  730 0.7723636
 1095 0.8755021
 1460 0.8983714
 1825 0.9102335
 2190 0.9220849
```

The estimates for the distribution function of the sojourn time in the recurrence state, corresponding to the time between entry in recurrence and death, reveal that the distribution function increases to a value near 49% and 78% for a time of one and two years, respectively, revealing a high risk of death shortly after relapse.

The methods for implementing some of the proposed methods can be computationally demanding. In particular, the use of bootstrap resampling techniques is time-consuming process because it is necessary to estimate the model a great number of times. In such cases, we recommend the use of parallelization (`cluster = TRUE`). This should considerably increase performance on multi-core/ multi-threading machines.

7.4 Discussion

There has been several recent contributions for the inference in the context of multi-state models. Many of these contributions were made for the illness-death model. One important and perhaps undervalued aspect of multi-state models is the possibility to apply them to obtain predictions of the clinical prognosis. This is usually achieved using estimates of the transition probabilities and survival estimates. However, there are several other quantities that could also be used in the analysis of these data, such as the state occupation probabilities, the sojourn time distributions, and the cumulative incidence functions. To provide the biomedical researchers with an easy-to-use tool for obtaining predictive estimates for all these quantities, we develop an R package called `survidm`. This package can be used to implement several nonparametric and semiparametric estimators for the transition probabilities. In addition, estimators have also implemented that account for the influence of covariates. Bootstrap confidence bands are provided for all methods. The software can also be used to perform multi-state regression (using type-specific Cox models).

Chapter 8

MSM.app: a Web-Based Tool for the Analysis of Multi-state Survival Data

The development of applications for obtaining interpretable results in a simple and summarized manner in multi-state models is a research field with great potential, namely in terms of using open source tools that can be easily implemented in biomedical applications. This chapter introduces `MSM.app`, an interactive web application using the `Shiny` package for the `R` language. This web application consists of three parts representing different aspects of the survival analysis and its extension to complex multi-state models. The first one allows to perform the survival analysis from mainly of most common functions of the survival `R` packages. The second enables to obtain some of the main goals of a multi-state analysis, such as the inference of regression models and the estimation of transition probabilities through the `survidm` and `mstate` `R` packages. Finally, `MSM.app` also includes local and global statistical tests to check the Markov assumption for multi-state using the `markovMSM` package.

The `MSM.app` allows one to perform traditional survival analysis in terms of estimating survival curves, comparing multiple curves, or inferring in regression models. It can also be used to obtain the results from some newly developed methods, such as the estimation of transition probabilities or the recent methods for checking the Markov assumption for multi-state models. For all of these aspects, the user can easily compare, for instance, statistical measures related to the precision of estimates or the validity of regression models. The application comprises a set of dynamic web forms, tables, and graphics, making use of the capabilities of the `Shiny` package, which enables any user, regardless of their previous knowledge of informatics, to perform a dynamic analysis involving the most important topics in multi-state models.

8.1 Introduction

Data visualization has grown in importance and popularity across a wide variety of fields over recent years (Kirk (2012) [132]; Yau (2013) [133]). In this regard, the development of a new variety of web tools has followed this increase, enabling researchers to gather and publish information to spur interest in the research community (Govan (2016) [134]). This type of web application includes databases, algorithms, services, and software tools, which provide some of the following advantages: availability; installation of software packages is unnecessary; optimization of computational resources; updates and access to the latest versions can easily be done; and the possibility of real-time validation of previous analysis (Lánczky and Győrffy (2021) [135]).

Since the mid-2000s, R (R Core Team, 2019 [86]) has become one of the most important programming languages to work in academia and businesses due to its abilities in data processing, statistics, and visualization, which has enabled it to be the fastest growing language for training methods at universities (Muenchen (2017) [136]). According to IEEE, it is also the main competitor of Python concerning data science, consistently rising to the top of the list of languages for most jobs (Varma and Virmani (2017) [137]). In recent years, new innovations have been developed to provide interactive web tools, such as the htmlwidgets package, that are friendlier by allowing us to wrap Javascript web visualization libraries in R code (Walker (2016) [138]).

The appearance of the shiny R package has also simplified the way to display outputs from R language. This type of application interactively shows outputs that can be viewed on the localhost or via the internet (Wojciechowski, Hopkins and Upton (2015) [139]). The shiny package also includes a number of graphical user interfaces for controlling the app's appearance and behavior (similar to how the HTTP wrapper works). In a sense, all one needs is a basic understanding of shiny input and output functions and some ideas on how to customize the user interface using in-built wrappers that allow the creation of a simple and intuitive user interface with dynamic filters and real-time explanatory analysis. If a developer wants to customize the user interface, Shiny can also integrate additional CSS and Javascript libraries within the web application (Seal and Wild, 2016; Varma and Virmani, 2017). Because all codes in shiny tools use only the R language and no knowledge of Javascript or HTTP is required, it has become simple for programmers to create and deploy web applications on Windows and Linux servers (Powers, Kopp and Martinez (2016) [140]; Dunning *et al*. (2017) [141]; Murrell and Potter (2014) [142]).

This chapter introduces MSM.app application, which combines shiny package, the survival R package (Therneau *et al.* (2021) [131]) , and the multi-estate R packages mstate (Putter *et al.* (2020) [143]) and survidm (Soutinho, Sestelo and Meira-Machado (2021) [120]). Other contributions for survival analysis and multi-state models in the R language can be seen in 'CRAN Task View: Survival Analysis'. Among the functionalities of the proposed web tool, MSM.app allows one to conduct a traditional survival analysis regarding the following topics: (i) estimation of survival functions (using the classical Kaplan-Meier estimator); (ii) comparison of survival functions between groups; and (iii) use of semiparametric and parametric regression models to study the relationship between explanatory variables and survival time. The MSM.app can also be used for the analysis of multi-state survival data. In fact, multi-state models (Putter, Fiocco and Geskus (2007) [18]; Meira-Machado *et al.* (2009) [15]; Meira-Machado and Sestelo (2019) [16]) can be seen as a generalization of survival analysis in which survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. Besides studying the effects of covariates in the course of the illness, two additional goals in multi-state models are the estimation of the transition probabilities and the cumulative incidence functions, which may be influenced by the fact that the process is Markovian or not. All these issues are considered in the MSM.app web tool. Recent reviews on these topics may be found in the papers by Meira-Machado and Sestelo (2019) [16] and Soutinho and Meira-Machado (2021) [108].

Among the available web tools for the analysis of multi-state survival data are MSDshiny, MSM-shiny and MSMplus. The MSDshiny application provides a useful and streamlined way to plan and power clinical trials with multi-state outcomes. Among the possibilities of analysis, this application provides a view of the multi-state structure, treatment effects, and can be used to perform simulations (Peterson (2019) [144]). The MSM-shiny application uses a CSV file containing multi-state data and provides the modeling and comparison of transition hazard models and the prediction of occupation probabilities (Lacy (2021) [145]). A recent contribution is MSMplus which provides a flexible visualization of the transition probabilities, transition intensities, or probability of visiting a particular state. In this regard, the user must upload two files: one that contains the structural and descriptive information of the multi-state model; and a second file that provides the various predictions from the model. This can be done using JSON files derived from functions developed in Stata or R. The CSV files can also be used as input files from a specific structure to be accepted for the application (Skourlis *et al.* (2021) [146]). It is our belief that a friendly web application for the analysis of survival data with one or more events of interest is still missing. MSM.app's goal is to provide a comprehensive web application that allows users to collect and publish dynamic data analysis involving the most important topics in survival analysis and multi-state models.

This chapter is organized as follows. In Section 8.2 a brief description of the background underlying to the development of the shiny application is presented. Section 8.3 describes all the pages that comprise the web tool which are available at the Shiny Apps repository https://gsoutinho.shinyapps.io/appmsm/. Finally, the main conclusions are reported in Section 8.4.

8.2 About Shiny applications architecture

Shiny is a package developed for the R language by RStudio that has become a popular way to create and deploy web server applications (Walker (2016) [138]). Through the combination of the structure of the shiny framework with R codes, one can build interactive web applications to obtain results or graphics without the user's prior knowledge of R, HTML, CSS or JavaScript (Chang (2017) [147]; Kaushik (2016) [148]). Shiny applications also provide integration with other R packages, JavaScript libraries, or CSS customization, and they are under the GPL-2 open source license (Seal and Wild (2016) [149]). The structure of shiny applications is built upon two components: the user-interface scripts for the layout of the application where the outputs are displayed (*ui.R*); and the other given by the server scripts with the instructions of the application (*server.R*) (Govan (2016) [134].

A shiny application can be run on a localhost where R and shiny package are installed. This is the easiest way to share when the users are familiar with R, and it is only necessary to execute the instructions to run the app (Varma and Virmani (2017) [137]). When using the RStudio services, shiny applications can be accessed in three ways by the internet: ShinyApps.io, Shiny Server, and Shiny Server Pro (Wojciechowski, Hopkins and Upto (2015) [139]). Shinyapps.io is a self-service platform that makes it easy to share shiny applications on the web. The service runs in the cloud on shared servers that are operated by RStudio. Each application is self-contained and operates on either data that is uploaded with the application, or data that the code pulls from third-party data stores, such as databases or web services (RStudio (2021) [150]). The Shiny server is totally free and open source, stable and well featured (Beeley (2013) [151]). Shiny Server Pro is a commercial version with enhanced security (possibility of allowing confidential web sharing of proprietary material) and additional features (Wojciechowski, Hopkins and Upton (2015) [139]).

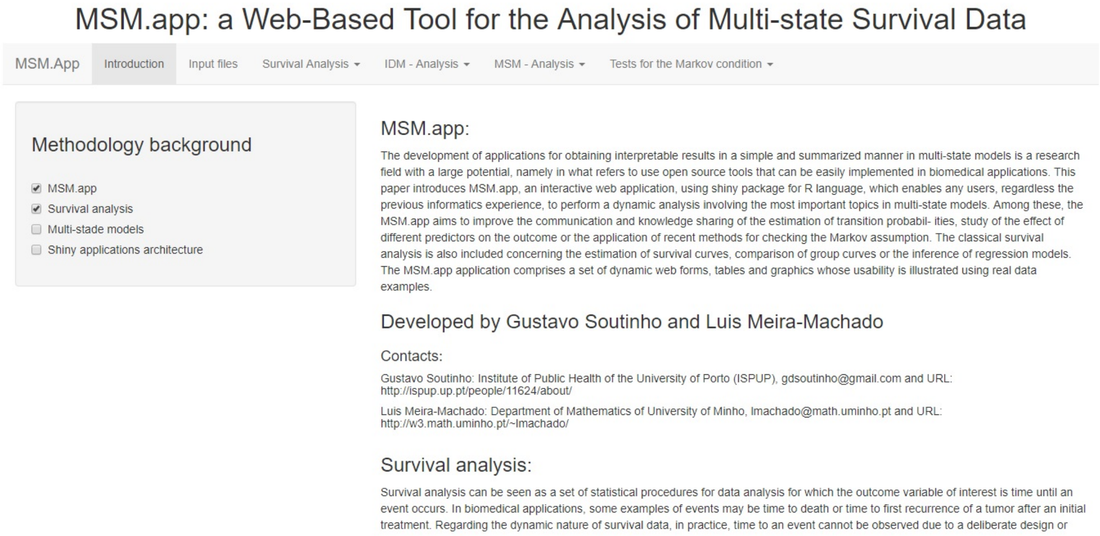
The communication between the client and server is done over the normal TCP connection. The data traffic that is needed for many web applications between the browser and the server is facilitated over the websockets protocol. This protocol operates separately from the handshake mechanism between the client and server and is done over the HTTP protocol (Seal and Wild (2016) [149]).

8.3 The MSM.app web application in practice

The MSM.app is a web application that can be used by any user, independently of their background knowledge of informatics, through a user-friendly interface that interactively provides web forms, reports, tables, and graphics. In the following subsections, all the functionalities of the application for obtaining results from survival and multi-state models are described using three real data examples: The first one involves data from survival in patients with Acute Myelogenous Leukemia (Miller (1997) [152]). The second corresponds to data from a clinical trial on colon cancer, which can be modeled using the progressive illness-death model (Moertel *et al.* (1990) [74]). Finally, extensions to progressive processes beyond the three-state illness-death model are discussed using data from the European Group for Blood and Marrow Transplantation (EBMT) (Putter, Fiocco and Geskus (2007) [18]).

8.3.1 About the Introduction web page

This page introduces an overview of the theoretical concepts addressed in MSM.app which are presented through check boxes. These HTML elements allow one to show or hide the contents described previously in Section 2: "survival analysis", "Multi-state models" and "Shiny applications architecture" as well as the main goals of the web tool in MSM.app (Figure 8.1).



MSM.app: a Web-Based Tool for the Analysis of Multi-state Survival Data

Methodology background

☑ MSM.app
☑ Survival analysis
☐ Multi-stade models
☐ Shiny applications architecture

MSM.app:

The development of applications for obtaining interpretable results in a simple and summarized manner in multi-state models is a research field with a large potential, namely in what refers to use open source tools that can be easily implemented in biomedical applications. This paper introduces MSM.app, an interactive web application, using shiny package for R language, which enables any users, regardless the previous informatics experience, to perform a dynamic analysis involving the most important topics in multi-state models. Among these, the MSM.app aims to improve the communication and knowledge sharing of the estimation of transition probabil- ities, study of the effect of different predictors on the outcome or the application of recent methods for checking the Markov assumption. The classical survival analysis is also included concerning the estimation of survival curves, comparison of group curves or the inference of regression models. The MSM.app application comprises a set of dynamic web forms, tables and graphics whose usability is illustrated using real data examples.

Developed by Gustavo Soutinho and Luis Meira-Machado

Contacts:

Gustavo Soutinho: Institute of Public Health of the University of Porto (ISPUP), gdsoutinho@gmail.com and URL: http://ispup.up.pt/people/11624/about/

Luis Meira-Machado: Department of Mathematics of University of Minho, lmachado@math.uminho.pt and URL: http://w3.math.uminho.pt/~lmachado/

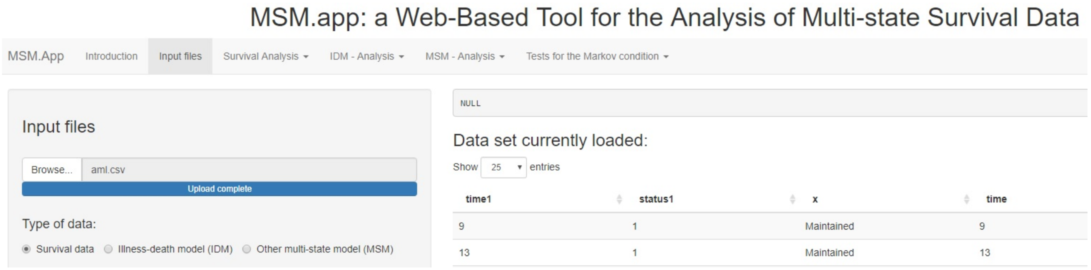
Survival analysis:

Survival analysis can be seen as a set of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. In biomedical applications, some examples of events may be time to death or time to first recurrence of a tumor after an initial treatment. Regarding the dynamic nature of survival data, in practice, time to an event cannot be observed due to a deliberate design or

FIGURE 8.1: Introduction web page: Through four check boxes, the user can see a brief description of the mathematical background underlying `MSM.app`. The main goals of the web application are also included.

8.3.2 The input file page

This page allows the user to select the necessary data sets for implementing the analysis. To this regard, there is a selection box to choose the data set to be used. When uploading the file, we must choose which type of analysis corresponds to the data set. This is accomplished by selecting one of three options ("survival data", "illness-death model", or "multi-state model"), the latter of which should be chosen if there are more than three states or reversible transitions. The selection is based on radio button elements (Figure 8.2). In terms of format, the `MSM.app` only requires a `CSV` file as input. This is a delimited file that uses a comma to separate the values and where each line of the file is a data record. Files that use a semicolon to separate the values are also accepted.



MSM.app: a Web-Based Tool for the Analysis of Multi-state Survival Data

Input files

aml.csv

Type of data:

● Survival data ○ Illness-death model (IDM) ○ Other multi-state model (MSM)

NULL

Data set currently loaded:

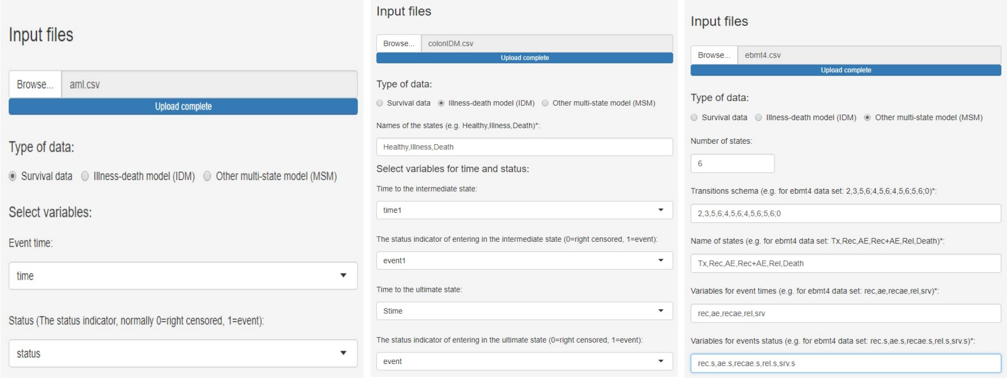
Show 25 entries

time1	status1	x	time
9	1	Maintained	9
13	1	Maintained	13

FIGURE 8.2: The input file page with the data table with some results for the *aml* data set and three radio buttons for each type of model.

Once the type of data is selected, a new web form appears below the radio buttons. For survival analysis, we must identify the "event time" and "status indicator" variables

(Figure 8.3, left hand side). In the case of an illness-death model, we have to first indicate the three state names of the model in the text box "names of the states", and then, select, respectively, the "time to the intermediate state", "the status indicator of entering the intermediate state", "time to the ultimate state" and "the status indicator of entering the ultimate state" variables (Figure 8.3, center). Finally, for more complex models (MSM), the options are "number of states", "transitions schema", 'name of states", "variables for event times" and "variables for event status" given by the text boxes `HTML` elements (Figure 8.3, right hand side).



Input files

aml.csv

Type of data:

● Survival data ○ Illness-death model (IDM) ○ Other multi-state model (MSM)

Select variables:

Event time: time

Status (The status indicator, normally 0=right censored, 1=event): status

Input files

colonIDM.csv

Type of data:

○ Survival data ● Illness-death model (IDM) ○ Other multi-state model (MSM)

Names of the states (e.g. Healthy,Illness,Death)*: Healthy,Illness,Death

Select variables for time and status:

Time to the intermediate state: time1

The status indicator of entering in the intermediate state (0=right censored, 1=event): event1

Time to the ultimate state: Stime

The status indicator of entering in the ultimate state (0=right censored, 1=event): event

Input files

ebmt4.csv

Type of data:

○ Survival data ○ Illness-death model (IDM) ● Other multi-state model (MSM)

Number of states: 6

Transitions schema (e.g. for ebmt4 data set: 2,3,5,6;4,5,6;4,5,6;5,6;0)*: 2,3,5,6;4,5,6;4,5,6;5,6;0

Name of states (e.g. for ebmt4 data set: Tx,Rec,AE,Rec+AE,Rel,Death)*: Tx,Rec,AE,Rec+AE,Rel,Death

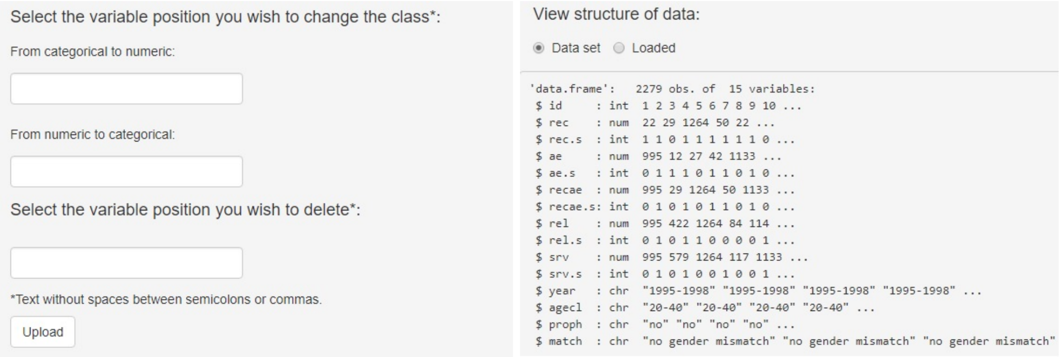
Variables for event times (e.g. for ebmt4 data set: rec,ae,recae,rel,srv)*: rec,ae,recae,rel,srv

Variables for events status (e.g. for ebmt4 data set: rec.s,ae.s,recae.s,rel.s,srv.s)*: rec.s,ae.s,recae.s,rel.s,srv.s

FIGURE 8.3: The event time and status variables for the survival analysis from the *aml* data set (left); indication of the two event times and their corresponding status for the illness-death model given by the *colonIDM* data (center); and a description of the MSM model through the number and the state names, the transition schema, and the event times and corresponding status for the *ebmt4* data (right).

For ease in statistical processing, most data sets have categorical variables that are assigned by numeric indices. However, these types of variables should require special attention since they cannot be entered into regression models just as they are. Instead, they need to be recoded into a set of binary variables that can then be entered into the model. This can be done on this page. It is also possible to delete some variables that are unnecessary to the analysis. All this can be done through three text boxes by typing the order index of each variable by the order in the data set (Figure 8.4, left hand side). After clicking on the 'upload' button, a data table appears on the right hand side of the page which can be dynamically changed using filters or by searching for specific words in the table (Figure 8.2). At any moment, we can also check the name and the type of variables that comprise the data set, as well as the last one that has been uploaded from the 'view structure of data' (Figure 8.4, right hand side).

For better understanding of the structure of the data, the `MSM`.app also provides three examples of data sets each one representing a type of analysis. These `CSV` files are available



Select the variable position you wish to change the class*:

From categorical to numeric:

From numeric to categorical:

Select the variable position you wish to delete*:

*Text without spaces between semicolons or commas.

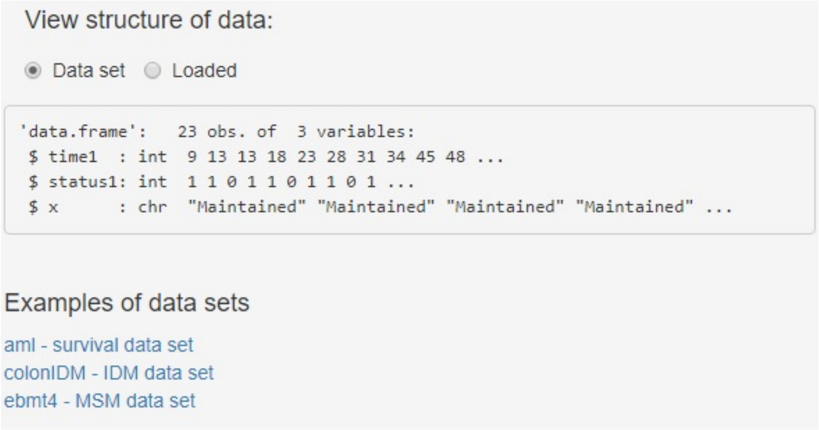
View structure of data:

● Data set ○ Loaded

```
'data.frame':	2279 obs. of  15 variables:
 $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ rec    : num  22 29 1264 50 22 ...
 $ rec.s  : int  1 1 0 1 1 1 1 1 1 0 ...
 $ ae     : num  995 12 27 42 1133 ...
 $ ae.s   : int  0 1 1 1 0 1 1 0 1 0 ...
 $ recae  : num  995 29 1264 50 1133 ...
 $ recae.s: int  0 1 0 1 0 1 1 0 1 0 ...
 $ rel    : num  995 422 1264 84 114 ...
 $ rel.s  : int  0 1 0 1 1 0 0 0 0 1 ...
 $ srv    : num  995 579 1264 117 1133 ...
 $ srv.s  : int  0 1 0 1 0 0 1 0 0 1 ...
 $ year   : chr  "1995-1998" "1995-1998" "1995-1998" "1995-1998" ...
 $ agecl  : chr  "20-40" "20-40" "20-40" "20-40" ...
 $ proph  : chr  "no" "no" "no" "no" ...
 $ match  : chr  "no gender mismatch" "no gender mismatch" "no gender mismatch"
```

FIGURE 8.4: Input file page: A partial view of the web form to select the index of variables to change the classes or delete them from the data set (left). Structure of the variables for the *ebmt4* data set (Right).

through the links on left side of the page at bottom of "View structure of data" (Figure 8.5).



View structure of data:

● Data set ○ Loaded

```
'data.frame':	23 obs. of  3 variables:
 $ time1  : int  9 13 13 18 23 28 31 34 45 48 ...
 $ status1: int  1 1 0 1 1 0 1 1 0 1 ...
 $ x      : chr  "Maintained" "Maintained" "Maintained" "Maintained" ...
```

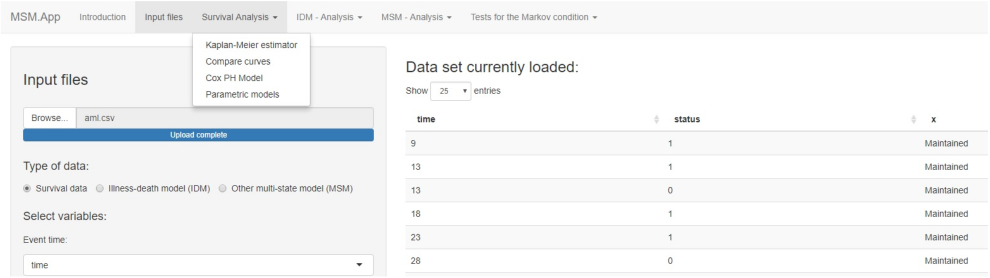
Examples of data sets

aml - survival data set
colonIDM - IDM data set
ebmt4 - MSM data set

FIGURE 8.5: Three examples of data set representing each of type of analysis.

8.3.3 Survival analysis pages

From the "survival analysis" button, we can carry out the classical methods for survival analysis (Figure 8.6). The Kaplan-Meier estimator can be used to estimate overall survival or the survival of different groups, while the log-rank (or the Gehan-Wilcoxon) test can be used to compare curves from two or more groups. The Cox model or the accelerated failure time model (AFT) can be used to test multiple predictors at once. To this end, we use the several functions of `survival` *R* package (Therneau *et al.* (2021) [131]).



Input files

aml.csv

Type of data:

● Survival data ○ Illness-death model (IDM) ○ Other multi-state model (MSM)

Select variables:

Event time:

time ▾

Data set currently loaded:

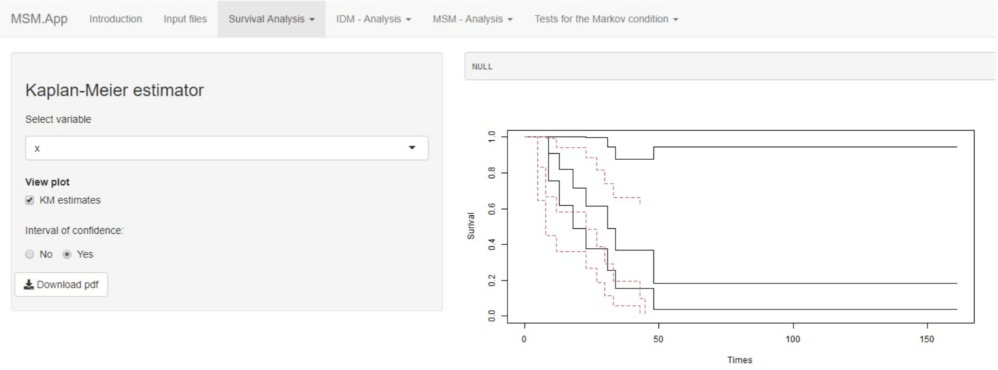
Show 25 ▾ entries

time	status	x
9	1	Maintained
13	1	Maintained
13	0	Maintained
18	1	Maintained
23	1	Maintained
28	0	Maintained

FIGURE 8.6: Output with the *tabPanels* shiny elements for the pages to obtain the survival analysis.

8.3.3.1 Kaplan-Meier estimator

Non-parametric estimation of the survival function is traditionally performed using the Kaplan-Meier estimator. This can be done from this page, where the outputs of the summary of estimates and plots are displayed on the right side. The estimates of the survival curves can be desegregated for different groups of categorical variables using the drop list “select variable”. By default, the check box for obtaining plots is disabled. When we click on it, the graph of survival is presented on the right side. By default, the interval of confidence for the plots is not selected. Plots with the confidence intervals can be obtained by choosing ”yes” in the corresponding radio button. Finally, it is also possible to export the plots of the survival curves to the PDF format by clicking the “Download pdf” button (Figure 8.7).



Kaplan-Meier estimator

Select variable

x ▾

View plot

☑ KM estimates

Interval of confidence:

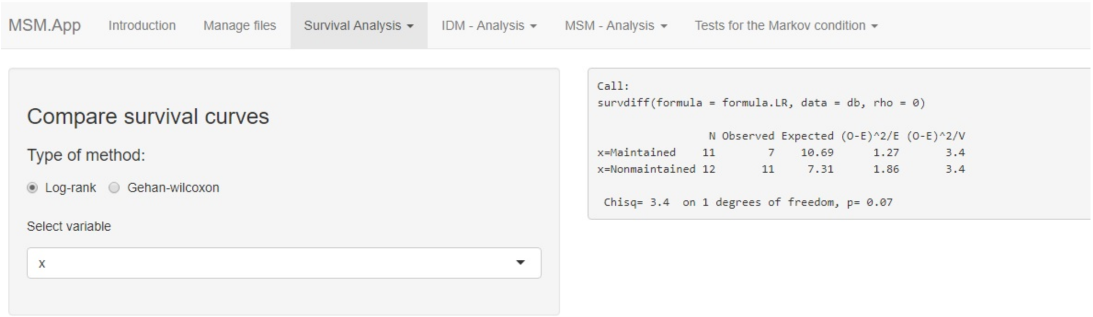
○ No ● Yes

```
NULL
```

FIGURE 8.7: The output of the survival estimation for the categorical covariate “x” of the *aml* data set using the Kaplan-Meier estimator. Survival curves for each group with confidence intervals are also shown at the bottom.

8.3.3.2 Compare survival curves

Statistical tests can be used to compare survival rates between groups. The null hypothesis states that there is no difference in survival between groups. These tests can be performed on this page. The log-rank test is the most commonly-used statistical test that is built using equal weights for all time points. When the survival curves cross at early time points, then one can identify groups that have higher risk at early time points while others have higher risk at late time points, making it necessary to use methods that give more weight to deaths at early time points. The Gehan-Wilcoxon test is one of these tests. Both tests are available in the `MSM.App`. On the left side of the page we have two radio buttons for the type of method to be used and a drop list `HTML` element with all the categorical variables which are automatically loaded. The outputs with the results of the tests are shown at all times that the user chooses the test or the specific covariate (Figure 8.8).



Compare survival curves

Type of method:

● Log-rank ○ Gehan-wilcoxon

Select variable

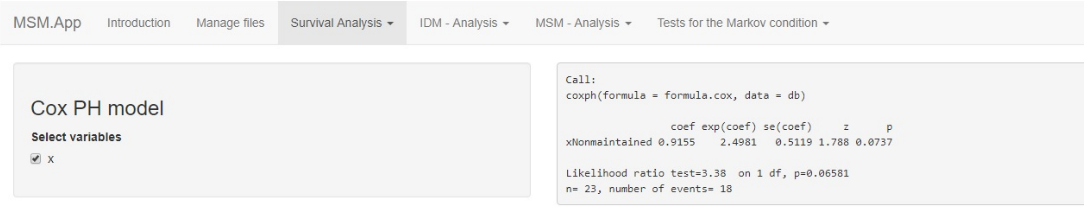
x

```
Call:
survdiff(formula = formula.LR, data = db, rho = 0)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27       3.4
x=Nonmaintained 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.07
```

FIGURE 8.8: The output of the Log-rank test for the categorical covariate "x" of the *aml* data set for testing the null hypothesis of no difference in survival between the two groups.

8.3.3.3 Cox PH models

The next step in the survival analysis is to simultaneously evaluate the effects of several factors on survival. To this end, on this page, we use the semiparametric Cox proportional hazards model (Cox (1972) [6]). Following the same idea as the previous pages, the outputs for the fitted models are shown on the right side of the page. All possible covariates to be included in the models are automatically shown in check boxes `HTML` elements. The output for the Cox model is updated as the user selects the variables to be included in the model (Figure 8.9).



Cox PH model

Select variables

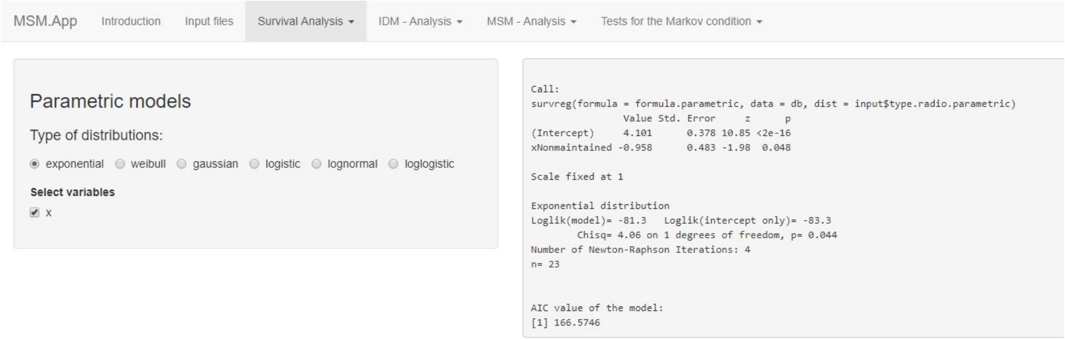
☑ x

```
Call:
coxph(formula = formula.cox, data = db)

                  coef exp(coef) se(coef)     z      p
xNonmaintained 0.9155    2.4981   0.5119 1.788 0.0737

Likelihood ratio test=3.38  on 1 df, p=0.06581
n= 23, number of events= 18
```

FIGURE 8.9: Results of the Cox model for the variable “x” of the *aml* data set.

8.3.3.4 Parametric models

Parametric survival models can be fitted in a similar way as for the Cox proportional hazards model. However, in this case, besides selecting the set of covariates from the check boxes, users also have to indicate the type of distribution to be used in the regression models. Six possibilities are available on the left side of the page through radio buttons. Using the interactivity of the shiny package, the outputs of the fitted models are updated by choosing the corresponding variables and the type of distribution for the survival times. The outputs also provide the Akaike Information Criterion (AIC) values for each model, making it easier to compare the model fit when using different distributions (Figure 8.10).



Parametric models

Type of distributions:

● exponential ○ weibull ○ gaussian ○ logistic ○ lognormal ○ loglogistic

Select variables

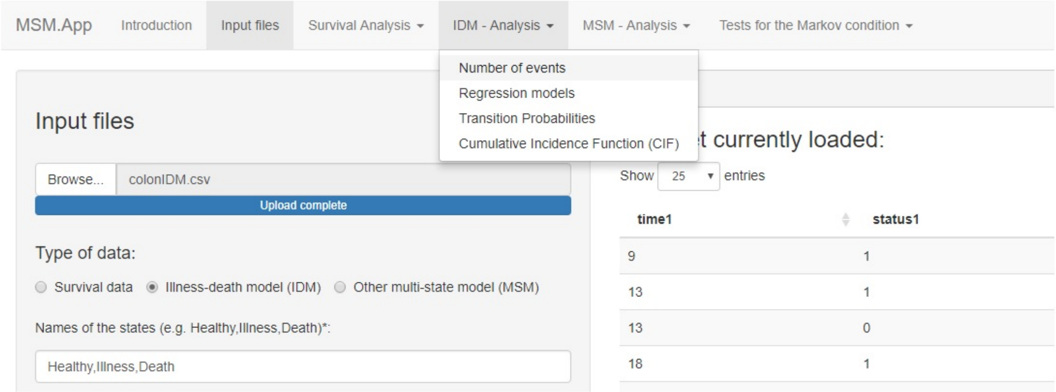
☑ x

```
Call:
survreg(formula = formula.parametric, data = db, dist = input$type.radio.parametric)
                Value Std. Error     z      p
(Intercept)     4.101      0.378 10.85 <2e-16
xNonmaintained -0.958      0.483 -1.98  0.048

Scale fixed at 1

Exponential distribution
Loglik(model)= -81.3   Loglik(intercept only)= -83.3
        Chisq= 4.06 on 1 degrees of freedom, p= 0.044
Number of Newton-Raphson Iterations: 4
n= 23


AIC value of the model:
[1] 166.5746
```

FIGURE 8.10: The outputs of the parametric models page with the results of the fitted model for the categorical variable “x” of the *aml* data set using the exponential distribution.

8.3.4 IDM-Analysis pages

In this section, we describe the steps for obtaining some of the most important goals for analyzing data from a progressive illness-death model given by two events, three states, and three transitions (Figure 8.11). Besides the occupation probabilities and the transition probabilities, the cumulative incidence functions are important predictive probabilities. Regression models for each of the three transition intensities can also be used to evaluate the effect of prognostic factors for each transition. To obtain the results, we have used several functions from the `survidm` R package (Soutinho, Sestelo and Meira-Machado (2021) [120]).



Input files

colonIDM.csv

Type of data:

○ Survival data ● Illness-death model (IDM) ○ Other multi-state model (MSM)

Names of the states (e.g. Healthy,Illness,Death)*:

Healthy,Illness,Death

t currently loaded:

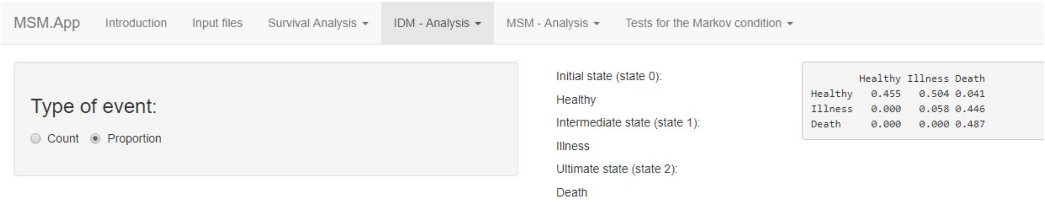
Show 25 entries

time1	status1
9	1
13	1
13	0
18	1

FIGURE 8.11: TabPanels shiny elements for the pages to obtain the illness-death model analysis.

8.3.4.1 Number of events

This page shows the number of individuals or items undergoing each transition among the three states of the illness-death model. On the left side of the page, we can change the output from “count” to “proportion” of transitions (Figure 8.12).



Type of event:

○ Count ● Proportion

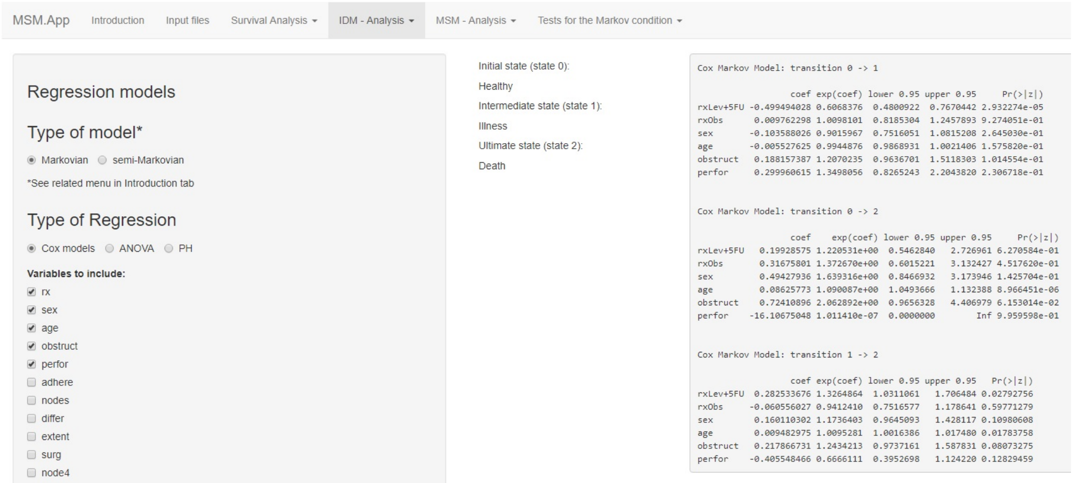
Initial state (state 0):
Healthy
Intermediate state (state 1):
Illness
Ultimate state (state 2):
Death

```
         Healthy Illness Death
Healthy    0.455   0.504 0.041
Illness    0.000   0.058 0.446
Death      0.000   0.000 0.487
```

FIGURE 8.12: The number of transitions among the three states of the *colonIDM* data set.

8.3.4.2 Regression models

Covariates may be incorporated into models through transition intensities to explain differences among individuals in the course of the illness. A common simplification strategy that allows one to relate the individual characteristics to the intensity rates is to decouple the multi-state process into various survival models by fitting separate intensities to all permitted transitions using semiparametric Cox proportional hazard regression models (Cox (1972) [6]). This is done with the introduction of appropriate adjustments to the risk set. This approach depends on the dependence of the transition intensities and time. To be specific, if the past and future are independent given the present state, which is known as the Markov assumption. When this assumption cannot be assumed, one alternative approach is to use a semi-Markov model in which the future of the process does not depend on the current time but rather on the duration in the current state. For the illness-death model, the Markov assumption is only relevant for the transition leaving the intermediate "disease" state, and the difference in the two approaches is also at that transition. Since the Markov assumption is of major importance to this end, first the user has to indicate if the process is Markovian or semi-Markovian through the radio buttons on the web form on the left side of the page. To check this assumption, a separate page was built and details about it are given below. In terms of regression, besides the traditional method for inference using Cox models, this page also provides the outputs of ANOVA tests and the *p*-values of the tests for nonlinearity. This can also be done by selecting the radio buttons on the web form. Finally, all possible covariates to be included in the model are associated with check boxes. As with other previous pages, as the user chooses some of the selected options in the web form, the output of the fitted models is also automatically updated on the right side of the page (Figures 8.13, 8.14 and 8.15).



Regression models

Type of model*

● Markovian ○ semi-Markovian

*See related menu in Introduction tab

Type of Regression

● Cox models ○ ANOVA ○ PH

Variables to include:

☑ rx
☑ sex
☑ age
☑ obstruct
☑ perfor
□ adhere
□ nodes
□ differ
□ extent
□ surg
□ node4

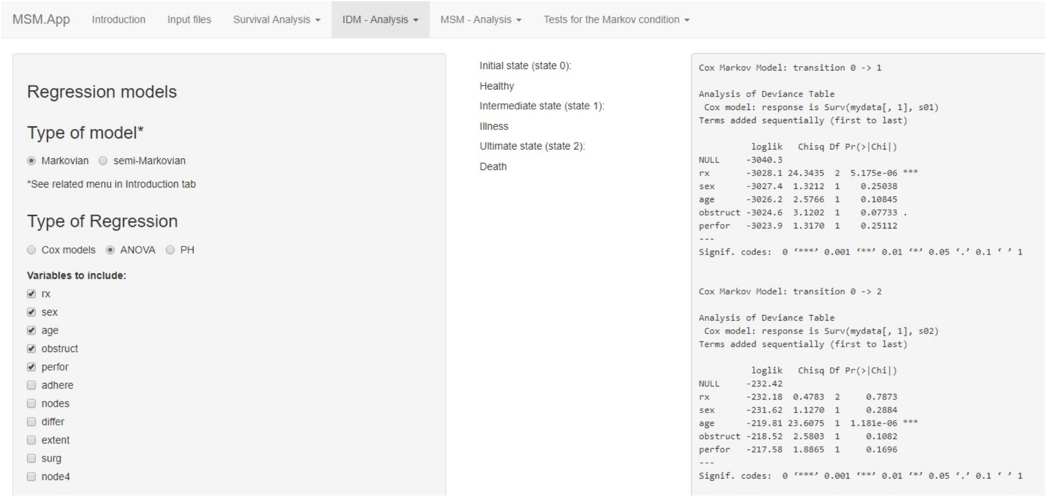
Initial state (state 0):
Healthy
Intermediate state (state 1):
Illness
Ultimate state (state 2):
Death

```
Cox Markov Model: transition 0 -> 1

                  coef exp(coef) lower 0.95 upper 0.95     Pr(>|z|)
rxLev+5FU -0.499494028 0.6068376  0.4800922  0.7670442 2.932274e-05
rxObs      0.009762298 1.0098101  0.8185304  1.2457893 9.274051e-01
sex       -0.103588026 0.9015967  0.7516051  1.0815208 2.645030e-01
age       -0.005527625 0.9944876  0.9868931  1.0021406 1.575820e-01
obstruct   0.188157387 1.2070235  0.9636701  1.5118303 1.014554e-01
perfor     0.299960615 1.3498056  0.8265243  2.2043820 2.306718e-01


Cox Markov Model: transition 0 -> 2

                  coef    exp(coef) lower 0.95 upper 0.95     Pr(>|z|)
rxLev+5FU   0.19928575 1.220531e+00  0.5462840   2.726961 6.270584e-01
rxObs       0.31675801 1.372670e+00  0.6015221   3.132427 4.517620e-01
sex         0.49427936 1.639316e+00  0.8466932   3.173946 1.425704e-01
age         0.08625773 1.090087e+00  1.0493666   1.132388 8.966451e-06
obstruct    0.72410896 2.062892e+00  0.9656328   4.406979 6.153014e-02
perfor    -16.10675048 1.011410e-07  0.0000000        Inf 9.959598e-01


Cox Markov Model: transition 1 -> 2

                  coef exp(coef) lower 0.95 upper 0.95   Pr(>|z|)
rxLev+5FU  0.282533676 1.3264864  1.0311061   1.706484 0.02792756
rxObs     -0.060556027 0.9412410  0.7516577   1.178641 0.59771279
sex        0.160110302 1.1736403  0.9645093   1.428117 0.10980608
age        0.009482975 1.0095281  1.0016386   1.017480 0.01783758
obstruct   0.217866731 1.2434213  0.9737161   1.587831 0.08073275
perfor    -0.405548466 0.6666111  0.3952698   1.124220 0.12829459
```

FIGURE 8.13: Results of the application of the Cox PH model with the following covariates: *rx*, *sex*, *age*, *obstruct* and *perfor*. Results for each of the three transition intensities of the *colonIDM*. A Markovian process is assumed.

8.3.4.3 Transition probabilities

The transition probabilities quantities are particularly of interest since they allow for long-term predictions of the multi-state process. These quantities can be nonparametrically estimated by the Aalen-Johansen estimator (Aalen and Johansen (1978) [61]), provided the



Regression models

Type of model*

● Markovian ○ semi-Markovian

*See related menu in Introduction tab

Type of Regression

○ Cox models ● ANOVA ○ PH

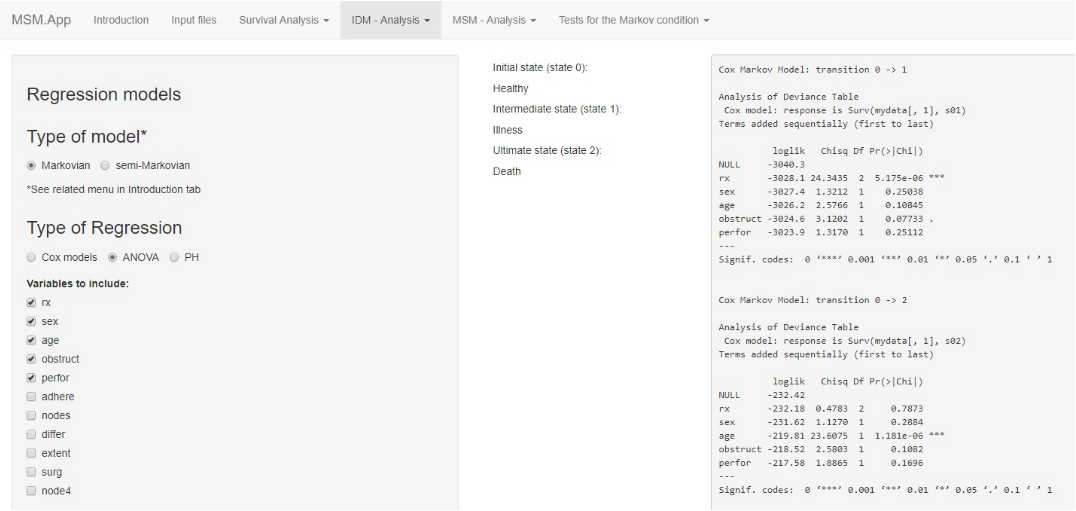
Variables to include:

☑ rx
☑ sex
☑ age
☑ obstruct
☑ perfor
□ adhere
□ nodes
□ differ
□ extent
□ surg
□ node4

Initial state (state 0):
Healthy
Intermediate state (state 1):
Illness
Ultimate state (state 2):
Death

```
Cox Markov Model: transition 0 -> 1

Analysis of Deviance Table
 Cox model: response is Surv(mydata[, 1], s01)
Terms added sequentially (first to last)

          loglik   Chisq Df Pr(>|Chi|)
NULL     -3040.3
rx       -3028.1 24.3435  2  5.175e-06 ***
sex      -3027.4  1.3212  1    0.25038
age      -3026.2  2.5766  1    0.10845
obstruct -3024.6  3.1202  1    0.07733 .
perfor   -3023.9  1.3170  1    0.25112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Cox Markov Model: transition 0 -> 2

Analysis of Deviance Table
 Cox model: response is Surv(mydata[, 1], s02)
Terms added sequentially (first to last)

          loglik   Chisq Df Pr(>|Chi|)
NULL     -232.42
rx       -232.18  0.4783  2     0.7873
sex      -231.62  1.1270  1     0.2884
age      -219.81 23.6075  1  1.181e-06 ***
obstruct -218.52  2.5803  1     0.1082
perfor   -217.58  1.8865  1     0.1696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 8.14: ANOVA results for transitions 0 ⟶ 1 and 0 ⟶ 2 with the following covariates: *rx*, *sex*, *age*, *obstruct* and *perfor* of the *colonIDM*. A Markovian process is assumed.



Regression models

Type of model*

● Markovian ○ semi-Markovian

*See related menu in Introduction tab

Type of Regression

○ Cox models ● ANOVA ○ PH

Variables to include:

☑ rx
☑ sex
☑ age
☑ obstruct
☑ perfor
□ adhere
□ nodes
□ differ
□ extent
□ surg
□ node4

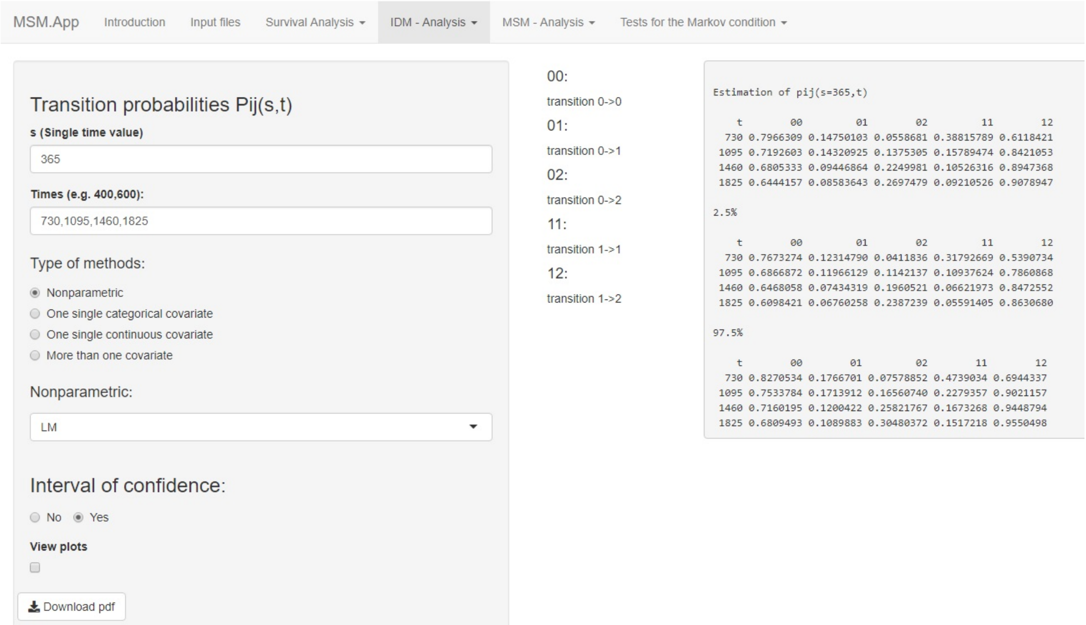
Initial state (state 0):
Healthy
Intermediate state (state 1):
Illness
Ultimate state (state 2):
Death

```
Cox Markov Model: transition 0 -> 1

Analysis of Deviance Table
 Cox model: response is Surv(mydata[, 1], s01)
Terms added sequentially (first to last)

          loglik   Chisq Df Pr(>|Chi|)
NULL     -3040.3
rx       -3028.1 24.3435  2  5.175e-06 ***
sex      -3027.4  1.3212  1    0.25038
age      -3026.2  2.5766  1    0.10845
obstruct -3024.6  3.1202  1    0.07733 .
perfor   -3023.9  1.3170  1    0.25112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Cox Markov Model: transition 0 -> 2

Analysis of Deviance Table
 Cox model: response is Surv(mydata[, 1], s02)
Terms added sequentially (first to last)

          loglik   Chisq Df Pr(>|Chi|)
NULL     -232.42
rx       -232.18  0.4783  2     0.7873
sex      -231.62  1.1270  1     0.2884
age      -219.81 23.6075  1  1.181e-06 ***
obstruct -218.52  2.5803  1     0.1082
perfor   -217.58  1.8865  1     0.1696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 8.15: The proportional hazards assumption was tested for the transitions 0 → 1 and 0 → 2 with the following covariates: *rx*, *sex*, *age*, *obstruct* and *perfor* of the *colonIDM*. A Markovian process is assumed.

system is Markovian. When the multi-state model is non-Markov, the Aalen-Johansen estimator may introduce some bias. In such cases, the use of methods that do not rely on this assumption is preferable. Among these methods are the estimators based on subsampling, also known as landmark methods. To be more specific, the landmark methods (`LM`) proposed by de Uña-Álvarez and Meira-Machado (2015) [69], as well as its presmoothed version (`PLM`) proposed by Meira-Machado (2016) [73], and the landmark Aalen-Johansen proposed by Putter and Spitoni (2018) [71]. All these non-parametric methods are available in `MSM.`app and can be accessed through this page. Suppose we are interested in obtaining the estimates of the transition probabilities from the initial single time $s = 365$

days to the next four years (730, 1095, 1460, and 1825 days) using the landmark approach. To this end, we start to fill in the first two text boxes on the left side of the page. This procedure will be the same for all types of estimation methods (given by the next four radio buttons). Then we select LM in "Nonparametric" and the results are automatically displayed on the right side of the page. If we are interested in seeing the confidence intervals, we just need to click "yes" on the radio button. As referred above, it is possible to see plots for each transition probability and export them using the last two HTML elements of the form (Figures 8.16 and 8.17).



Transition probabilities Pij(s,t)

s (Single time value)
365

Times (e.g. 400,600):
730,1095,1460,1825

Type of methods:
● Nonparametric
○ One single categorical covariate
○ One single continuous covariate
○ More than one covariate

Nonparametric:
LM

Interval of confidence:
○ No ● Yes

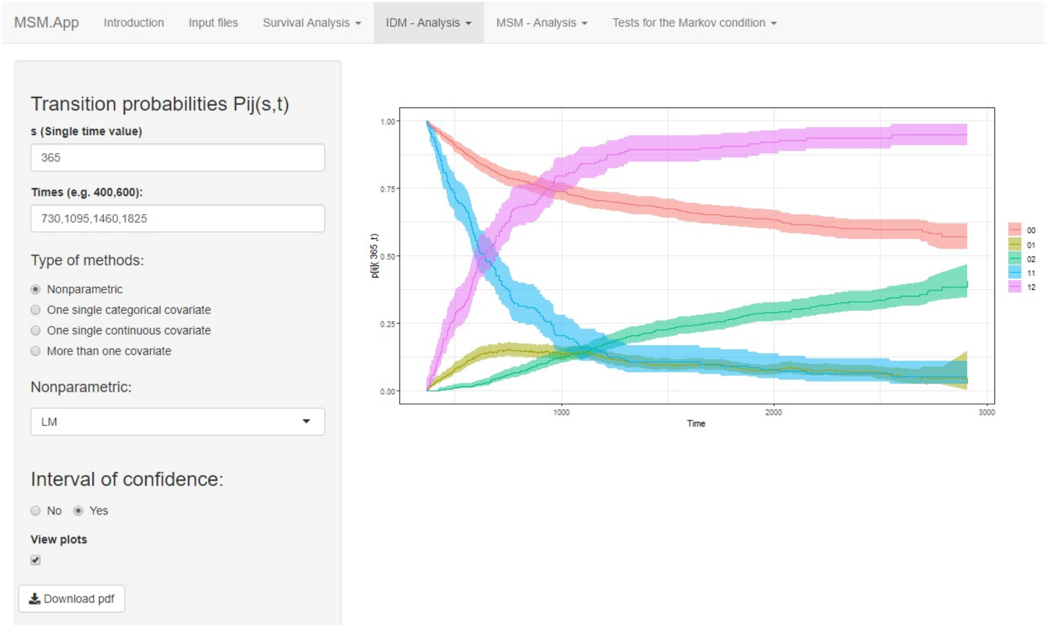
View plots
□

00:
transition 0->0
01:
transition 0->1
02:
transition 0->2
11:
transition 1->1
12:
transition 1->2

```
Estimation of pij(s=365,t)

    t       00         01        02         11        12
  730 0.7966309 0.14750103 0.0558681 0.38815789 0.6118421
 1095 0.7192603 0.14320925 0.1375305 0.15789474 0.8421053
 1460 0.6805333 0.09446864 0.2249981 0.10526316 0.8947368
 1825 0.6444157 0.08583643 0.2697479 0.09210526 0.9078947

2.5%

    t       00         01        02         11        12
  730 0.7673274 0.12314790 0.0411836 0.31792669 0.5390734
 1095 0.6866872 0.11966129 0.1142137 0.10937624 0.7860868
 1460 0.6468058 0.07434319 0.1960521 0.06621973 0.8472552
 1825 0.6098421 0.06760258 0.2387239 0.05591405 0.8630680

97.5%

    t       00        01         02        11        12
  730 0.8270534 0.1766701 0.07578852 0.4739034 0.6944337
 1095 0.7533784 0.1713912 0.16560740 0.2279357 0.9021157
 1460 0.7160195 0.1200422 0.25821767 0.1673268 0.9448794
 1825 0.6809493 0.1089883 0.30480372 0.1517218 0.9550498
```

FIGURE 8.16: Results of the estimates of the transition probabilities and the corresponding confidence intervals for $s = 365$ and times 730, 1090, 1460, and 1825 days for the *colonIDM* data set using the landmark estimators.

Categorical covariates can be included using all these four methods by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. To account for the effect of one continuous covariate, the nonparametric method proposed by Meira-Machado, de Uña-Álvarez and Datta (2015) [24] is implemented in this web tool by selecting "IPCW" from the radio button and the drop list "one single continuous covariate". This type of estimator is based on local smoothing, which is introduced using regression weights where the right censoring is handled by appropriate reweighting on observations using the Inverse Probability of Censoring Weighting (IPCW estimator). The next two steps to obtaining the transition probabilities are to choose the



Transition probabilities Pij(s,t)

s (Single time value)

365

Times (e.g. 400,600):

730,1095,1460,1825

Type of methods:

● Nonparametric
○ One single categorical covariate
○ One single continuous covariate
○ More than one covariate

Nonparametric:

LM

Interval of confidence:

○ No ● Yes

View plots

☑

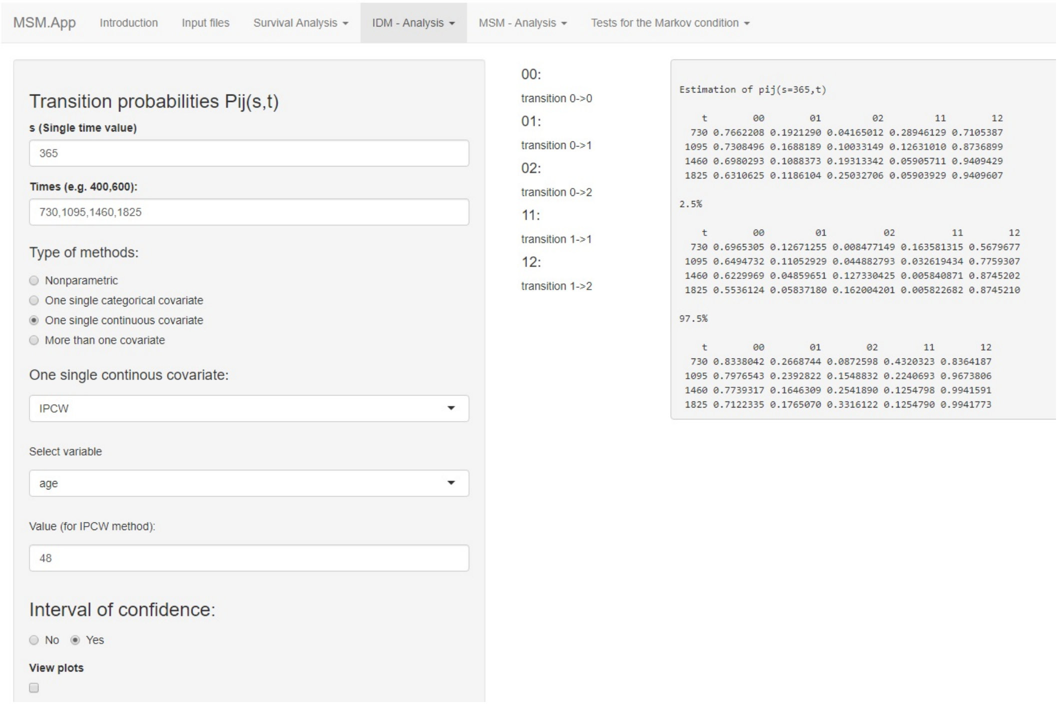
FIGURE 8.17: Transition probability estimates with confidence intervals for each transition for $s = 365$ using the landmark estimators.

continuous variable and fill in the corresponding value in the text box. In the example of Figure 8.18 we have selected the variable “age” taking the value of 48 years.

Finally, one standard method (particularly well-suited to the setting with multiple covariates) is to consider estimators based on a Cox’s model fitted marginally to each type of transition, with the corresponding baseline hazard function estimated by the Breslow’s method (Breslow (1972) [23]. In the `MSM.app` web application, the estimation of this type of transition probability conditional on several covariates can be done by clicking “more than one covariate” in the radio button ”type of methods”. Then all the possible covariates to include in the model appear on the form on the left side with the aspect of check boxes and a text box where the user can fill in with the corresponding values. Taking the example of Figure 8.19, three covariates were selected for the model (“rx”,“sex”, and “age”) and the respective values are: `Obs,1,48` (without spaces or quotation marks).

8.3.4.4 Cumulative Incidence Function (CIF)

Another quantity of interest in multi-state modeling is the cause-specific cumulative incidence function (CIF), as defined by Kalbfleisch and Prentice (1980) [50]. In the illness-death model, the cumulative incidence of the illness (intermediate state) is of particular interest. This quantity denotes the probability of the individual or item being or having been in the intermediate ‘diseased’ state at some particular time t. As for the transition



Transition probabilities Pij(s,t)

s (Single time value): 365

Times (e.g. 400,600): 730,1095,1460,1825

Type of methods:

○ Nonparametric
○ One single categorical covariate
● One single continuous covariate
○ More than one covariate

One single continous covariate: IPCW

Select variable: age

Value (for IPCW method): 48

Interval of confidence:

○ No ● Yes

View plots

□

```
00:
transition 0->0
01:
transition 0->1
02:
transition 0->2
11:
transition 1->1
12:
transition 1->2
```

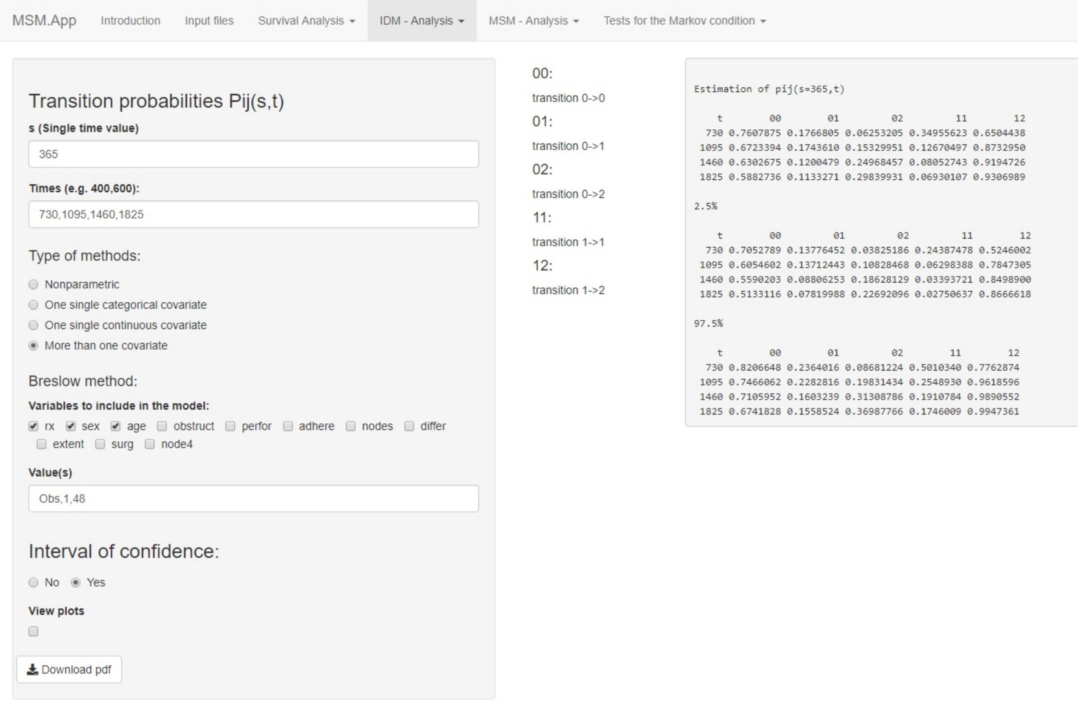
```
Estimation of pij(s=365,t)

    t        00        01         02         11        12
  730 0.7662208 0.1921290 0.04165012 0.28946129 0.7105387
 1095 0.7308496 0.1688189 0.10033149 0.12631010 0.8736899
 1460 0.6980293 0.1088373 0.19313342 0.05905711 0.9409429
 1825 0.6310625 0.1186104 0.25032706 0.05903929 0.9409607

2.5%

    t        00         01          02          11        12
  730 0.6965305 0.12671255 0.008477149 0.163581315 0.5679677
 1095 0.6494732 0.11052929 0.044882793 0.032619434 0.7759307
 1460 0.6229969 0.04859651 0.127330425 0.005840871 0.8745202
 1825 0.5536124 0.05837180 0.162004201 0.005822682 0.8745210

97.5%

    t        00        01        02        11        12
  730 0.8338042 0.2668744 0.0872598 0.4320323 0.8364187
 1095 0.7976543 0.2392822 0.1548832 0.2240693 0.9673806
 1460 0.7739317 0.1646309 0.2541890 0.1254798 0.9941591
 1825 0.7122335 0.1765070 0.3316122 0.1254790 0.9941773
```

FIGURE 8.18: Results of the estimates of the transition probabilities and the corresponding confidence intervals for $s = 365$ and times 730, 1090, 1460, and 1825 days for the *colonIDM* data set using the `IPCW` estimator.

probabilities, this quantity can be estimated conditional on a covariate, continuous or categorical. To get the outputs on the right side, first, the user should start by indicating the list of times (comma separated) in the text box labelled "times". To display the confidence interval for the estimates, one has to select "yes" in the corresponding radio button. To estimate this quantity conditional on covariates, one must choose which covariate to be considered in the drop list "Select variable". In the case of a continuous variable, we must also insert the value in the text box below. As an example, in Figure 8.20, it was selected "age" at 50 years. By default, the CIF estimates do not include any covariates since the drop list "Select variable" appears with "none".

8.3.5 MSM-Analysis pages

By clicking on the "MSM-analysis" we can extend some of the methods addressed in the previous section to more complex multi-state models involving more than three states and considering the possibility of reversible transitions. Three pages are available that allow one to obtain the number of movements of the individual among states, the outputs of the regression models for each transition, and the results for the transition probabilities using the classical Aalen-Johansen estimator that assumes the process to be Markovian and the



Transition probabilities Pij(s,t)

s (Single time value)

365

Times (e.g. 400,600):

730,1095,1460,1825

Type of methods:

○ Nonparametric
○ One single categorical covariate
○ One single continuous covariate
● More than one covariate

Breslow method:

Variables to include in the model:

☑ rx ☑ sex ☑ age □ obstruct □ perfor □ adhere □ nodes □ differ □ extent □ surg □ node4

Value(s)

Obs,1,48

Interval of confidence:

○ No ● Yes

View plots

□

00:
transition 0->0
01:
transition 0->1
02:
transition 0->2
11:
transition 1->1
12:
transition 1->2

```
Estimation of pij(s=365,t)

   t        00        01         02         11        12
 730 0.7607875 0.1766805 0.06253205 0.34955623 0.6504438
1095 0.6723394 0.1743610 0.15329951 0.12670497 0.8732950
1460 0.6302675 0.1200479 0.24968457 0.08052743 0.9194726
1825 0.5882736 0.1133271 0.29839931 0.06930107 0.9306989

2.5%

   t        00         01         02         11        12
 730 0.7052789 0.13776452 0.03825186 0.24387478 0.5246002
1095 0.6054602 0.13712443 0.10828468 0.06298388 0.7847305
1460 0.5590203 0.08806253 0.18628129 0.03393721 0.8498900
1825 0.5133116 0.07819988 0.22692096 0.02750637 0.8666618

97.5%

   t        00        01         02        11        12
 730 0.8206648 0.2364016 0.08681224 0.5010340 0.7762874
1095 0.7466062 0.2282816 0.19831434 0.2548930 0.9618596
1460 0.7105952 0.1603239 0.31308786 0.1910784 0.9890552
1825 0.6741828 0.1558524 0.36987766 0.1746009 0.9947361
```

FIGURE 8.19: Results of the estimates of the transition probabilities and the corresponding confidence intervals for $s = 365$ and times 730, 1090, 1460, and 1825 days for the *colonIDM* data set using the Breslow estimator.

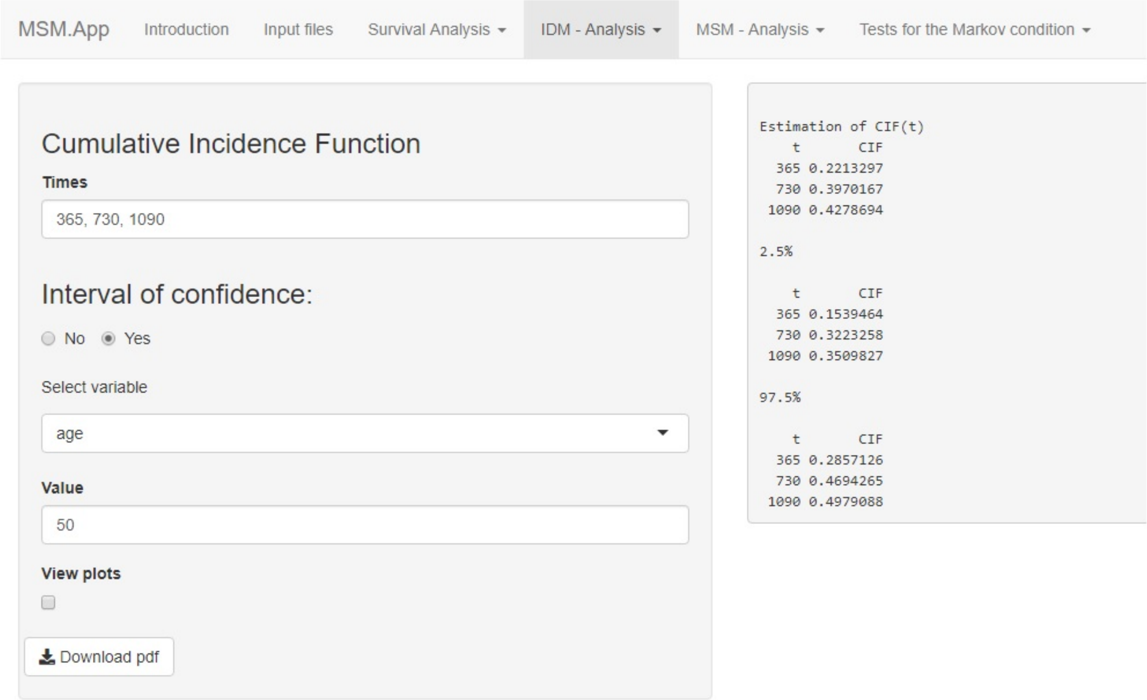
recent landmark Aalen-Johansen estimator that is free of this condition (Figure 8.21). All these methods were implemented by using several functions from the *mstate R* package (Putter *et al.* (2020) [143]. The following examples of applications were obtained from the European Group for Blood and Marrow Transplantation (Putter, Fiocco and Geskus (2007) [18].

8.3.5.1 Number of events

This page shows the movement of individuals among the states in the multi-state model. The outputs on the right side of the page change automatically when selecting the radio buttons "count" or "proportion" of transitions (Figure 8.22).

8.3.5.2 Regression models

In a similar way to that which occurred for IDM models, we are now interested in the estimation of covariate effects for each transition using Cox regression models. In this regard, we assume the decoupling of the process into various survival models, taking into account the delayed entry into each transition in accordance with Putter (2020) [153]. The summary outputs of the models are automatically shown on the right side of the page



Cumulative Incidence Function

Times: 365, 730, 1090

Interval of confidence: ○ No ● Yes

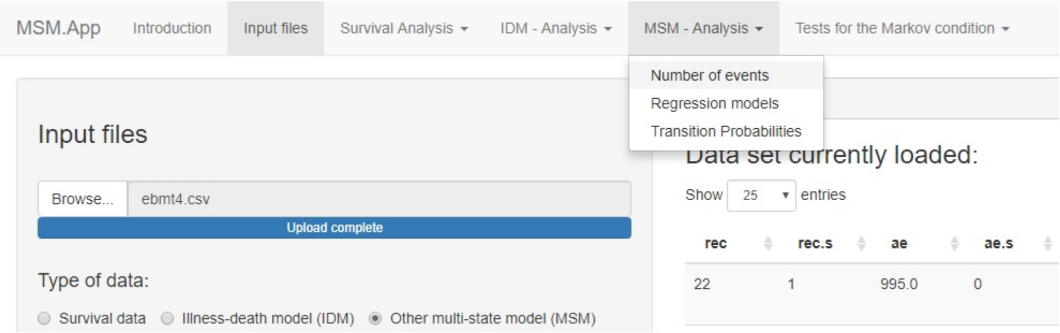
Select variable: age

Value: 50

View plots: □

```
Estimation of CIF(t)
    t       CIF
  365 0.2213297
  730 0.3970167
 1090 0.4278694

2.5%

    t       CIF
  365 0.1539464
  730 0.3223258
 1090 0.3509827

97.5%

    t       CIF
  365 0.2857126
  730 0.4694265
 1090 0.4979088
```

FIGURE 8.20: Cumulative recurrence incidence with 95% bootstrap confidence intervals. Data from a colon cancer study.



Input files

ebmt4.csv

Type of data: ○ Survival data ○ Illness-death model (IDM) ● Other multi-state model (MSM)

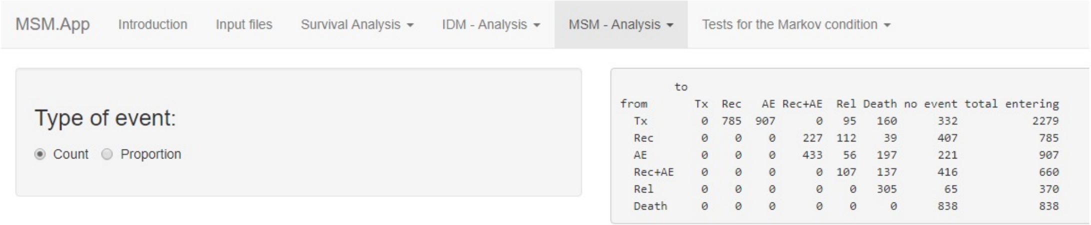
Data set currently loaded:

Show 25 entries

rec	rec.s	ae	ae.s
22	1	995.0	0

FIGURE 8.21: *TabPanels* shiny elements that were used for the pages to obtain the multi-state models (MSM) analysis.

as the user changes which covariates should be included in the models for each transition. This can be done through the check boxes where all possible covariates are presented and from drop lists with all combinations of transitions. As an example, in Figure 8.23, we have got the results of the Cox regression model for the transition 1 → 2 which includes the covariates “year”, “age” and “proph”.

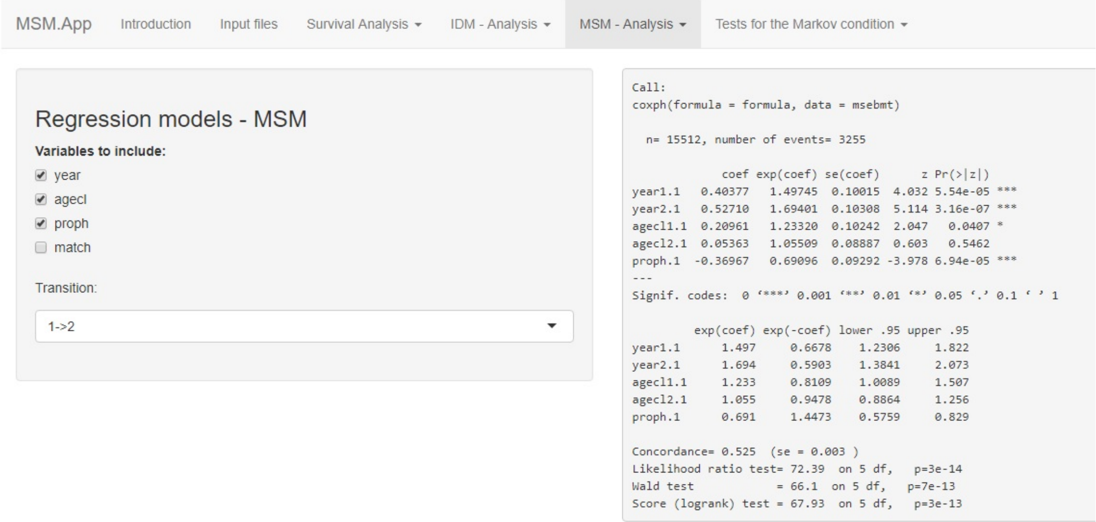


Type of event:

● Count ○ Proportion

```
         to
from      Tx  Rec   AE Rec+AE  Rel Death no event total entering
  Tx       0  785  907      0   95   160      332           2279
  Rec      0    0    0    227  112    39      407            785
  AE       0    0    0    433   56   197      221            907
  Rec+AE   0    0    0      0  107   137      416            660
  Rel      0    0    0      0    0   305       65            370
  Death    0    0    0      0    0     0      838            838
```

FIGURE 8.22: The number of transitions among the states of the *ebmt4* data set.



Regression models - MSM

Variables to include:

☑ year
☑ agecl
☑ proph
□ match

Transition:

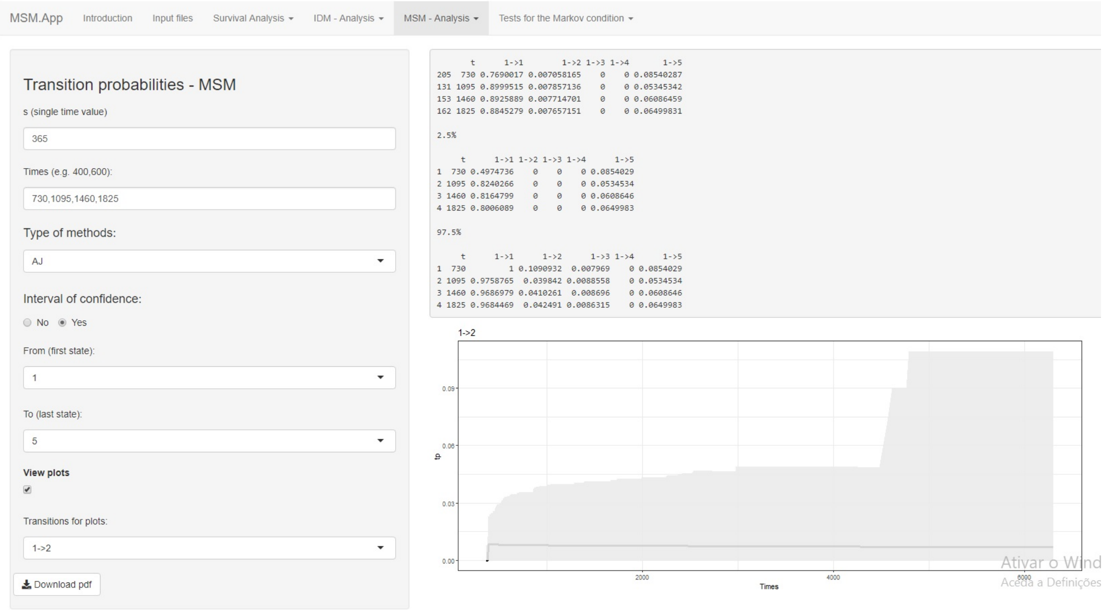
1->2 ▾

```
Call:
coxph(formula = formula, data = msebmt)

  n= 15512, number of events= 3255

             coef exp(coef) se(coef)      z Pr(>|z|)
year1.1   0.40377   1.49745  0.10015  4.032 5.54e-05 ***
year2.1   0.52710   1.69401  0.10308  5.114 3.16e-07 ***
agecl1.1  0.20961   1.23320  0.10242  2.047   0.0407 *
agecl2.1  0.05363   1.05509  0.08887  0.603   0.5462
proph.1  -0.36967   0.69096  0.09292 -3.978 6.94e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
year1.1      1.497     0.6678    1.2306     1.822
year2.1      1.694     0.5903    1.3841     2.073
agecl1.1     1.233     0.8109    1.0089     1.507
agecl2.1     1.055     0.9478    0.8864     1.256
proph.1      0.691     1.4473    0.5759     0.829

Concordance= 0.525  (se = 0.003 )
Likelihood ratio test= 72.39  on 5 df,   p=3e-14
Wald test            = 66.1  on 5 df,   p=7e-13
Score (logrank) test = 67.93  on 5 df,   p=3e-13
```

FIGURE 8.23: The output of the Cox regression model for the transition 1 → 2 that include the ’year”, “age” and “proph” covariates. *ebmt4* data set.

8.3.5.3 Transition probabilities

This page extends the estimation of the transition probabilities to multi-state systems that may involve more than three states with possible reversible transitions. The steps to obtain the results are quite similar to those used for the illness-death model. The user should start by introducing the initial (s) and the other times for obtaining the estimates in the text box elements. Two methods for inference are available: the markovian Aalen-Johansen estimator `AJ` and the non-markov `LMAJ` estimator that can be chosen through the drop list, “and the confidence intervals are displayed by selecting “yes”. As an example, Figure 8.24 shows the Aalen-Johansen estimates of the transition probabilities for the EBMT data set, considering all possible transition probabilities from the initial state 1 to the state 5 (which are chosen by selecting the corresponding drop list on the left side of the page), with $s = 365$ and the other times 730, 1095, 1460, and 1825. When clicking on the button “view plots” a new drop list with all possible transitions from the initial state appears, enabling you to see the plot for the transition probabilities. In this example, we have chosen the transition 1 → 2. Finally, it is also possible to export to a PDF format file the corresponding plot.



Transition probabilities - MSM

s (single time value): 365

Times (e.g. 400,600): 730,1095,1460,1825

Type of methods: AJ

Interval of confidence: ○ No ● Yes

From (first state): 1

To (last state): 5

View plots ☑

Transitions for plots: 1->2

```
        t      1->1        1->2 1->3 1->4       1->5
205  730 0.7690017 0.007058165    0    0 0.08540287
131 1095 0.8999515 0.007857136    0    0 0.05345342
153 1460 0.8925889 0.007714701    0    0 0.06086459
162 1825 0.8845279 0.007657151    0    0 0.06499831

2.5%

     t      1->1 1->2 1->3 1->4      1->5
1  730 0.4974736    0    0    0 0.0854029
2 1095 0.8240266    0    0    0 0.0534534
3 1460 0.8164799    0    0    0 0.0608646
4 1825 0.8006089    0    0    0 0.0649983

97.5%

     t      1->1      1->2      1->3 1->4      1->5
1  730         1 0.1090932  0.007969    0 0.0854029
2 1095 0.9758765  0.039842 0.0088558    0 0.0534534
3 1460 0.9686979 0.0410261  0.008696    0 0.0608646
4 1825 0.9684469  0.042491 0.0086315    0 0.0649983
```

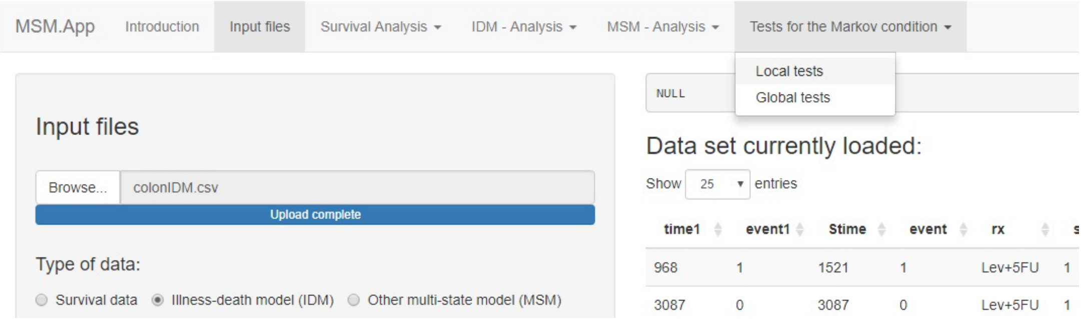
FIGURE 8.24: Estimates of all possible transition probabilities from the state 1 to 5, for s = 365 and times equal to 730, 1095, 1460, and 1825 using the AJ estimators. *ebmt4* data set.

8.3.6 Tests for the Markov condition pages

A critical aspect to take into account for the inference of regression models and transition probabilities is to check the Markov condition. The MSM.app web application provides two types of tests for checking this assumption using recent literature methods: (i) local tests, which are obtained by fixing a specific time value, s, and are especially useful for estimating transition probabilities; and (ii) global tests, which may be preferable for regression purposes.These tests can be accessed through the menu “Tests for the Markov condition” (Figure 8.25). The mathematics underlying the proposed methods has been discussed in the previous papers by Soutinho and Meira-Machado (2021) [108] and Titman and Putter (2020) [113].

8.3.6.1 Local tests

This page allows one to obtain the results for local tests that check the Markov assumption. First, the user must indicate which type of multi-state corresponds to the data set upload in the *input file page*: IDM or a general model with more than three states or that allows reversible transitions. Two types of methods are available for checking the local tests



Input files

colonIDM.csv

Type of data:

○ Survival data ● Illness-death model (IDM) ○ Other multi-state model (MSM)

NULL

Data set currently loaded:

Show 25 entries

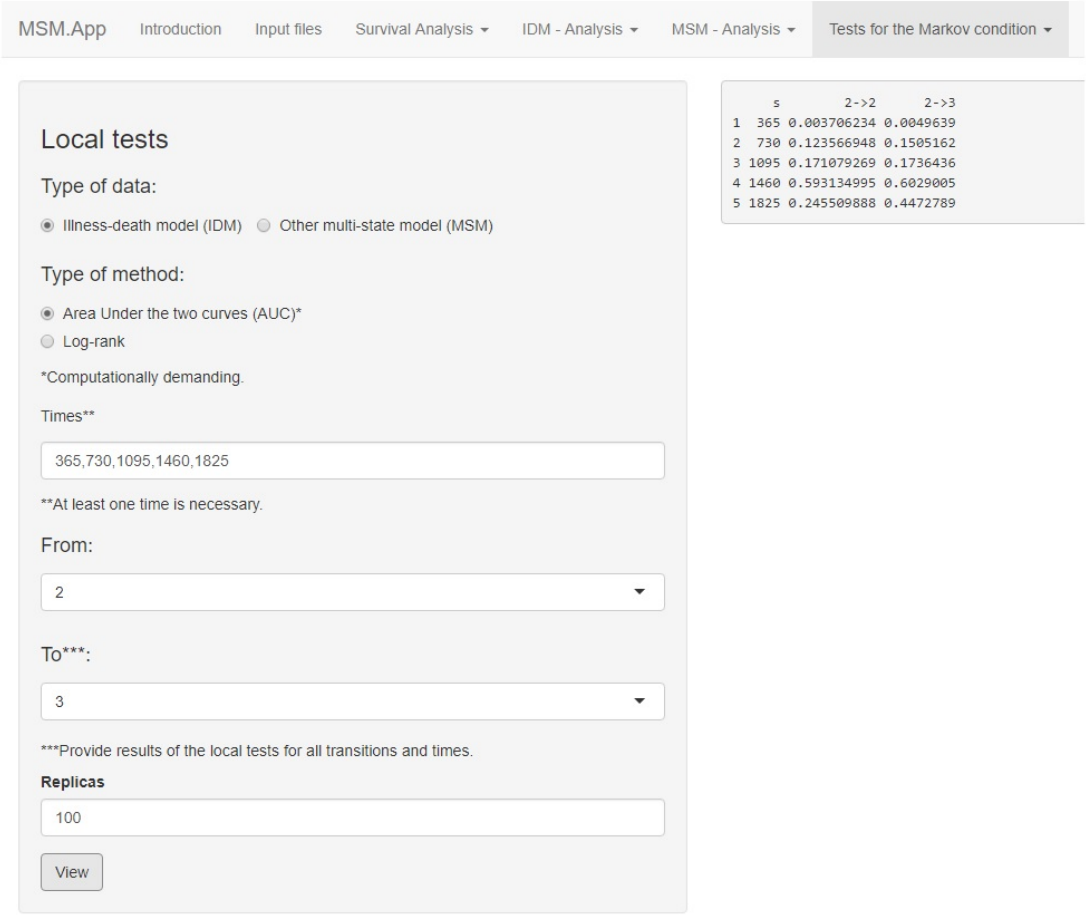
time1	event1	Stime	event	rx	s
968	1	1521	1	Lev+5FU	1
3087	0	3087	0	Lev+5FU	1

FIGURE 8.25: *TabPanels* shiny elements for the local and global tests pages for the tests for the Markov condition.

of the Markov condition: (i) the AUC method, which is based on measuring the discrepancy between the AJ estimator of the transition probabilities (which provides consistent estimates when the process is Markovian) and the landmark estimators (which are free of the Markov condition). In this case, the web tool uses the LM estimator for the progressive illness-death models and LMAJ in the case of more complex MSM models; (ii) the Log-rank method, which considers summaries from families of log-rank statistics where patients are grouped by the state occupied at different times. The next steps are to enter the specific values to be used to check the Markov assumption (via a text box element), indicate which transitions we are interested in obtaining the results from, and the number of replicas used in the tests in the text box "times". As an example, in Figure 8.26, we obtained the p-values for the local tests based on the s times 365, 730, 1095, 1460, and 1825 for the transition from state 2 to state 3. Results were obtained for an IDM model based on the colon cancer data using 100 replicas. The procedure to get results using the Log-rank test is quite similar, but in terms of output, the web tool provides results of the local test for the specific transition and times. It is suggested that this method be used with 500 replicas since the overall computation demand is reasonable. For more complex MSM models, the web form and the steps to obtain the outputs are the same with the only difference being the number of states in the drop list boxes for "from" and "to" are higher than the IDM models.

8.3.6.2 Global tests

On this page, we can obtain the results of the global tests to check the Markov assumption. Both for IMD and more complex MSM models, three global tests are available: (i) the first one is based on Cox models, from which it is possible to evaluate the effect of history on the process. In this case, this can be done by checking the significance of the covariate time



Local tests

Type of data:

● Illness-death model (IDM) ○ Other multi-state model (MSM)

Type of method:

● Area Under the two curves (AUC)*

○ Log-rank

*Computationally demanding.

Times**

365,730,1095,1460,1825

**At least one time is necessary.

From:

2

To***:

3

***Provide results of the local tests for all transitions and times.

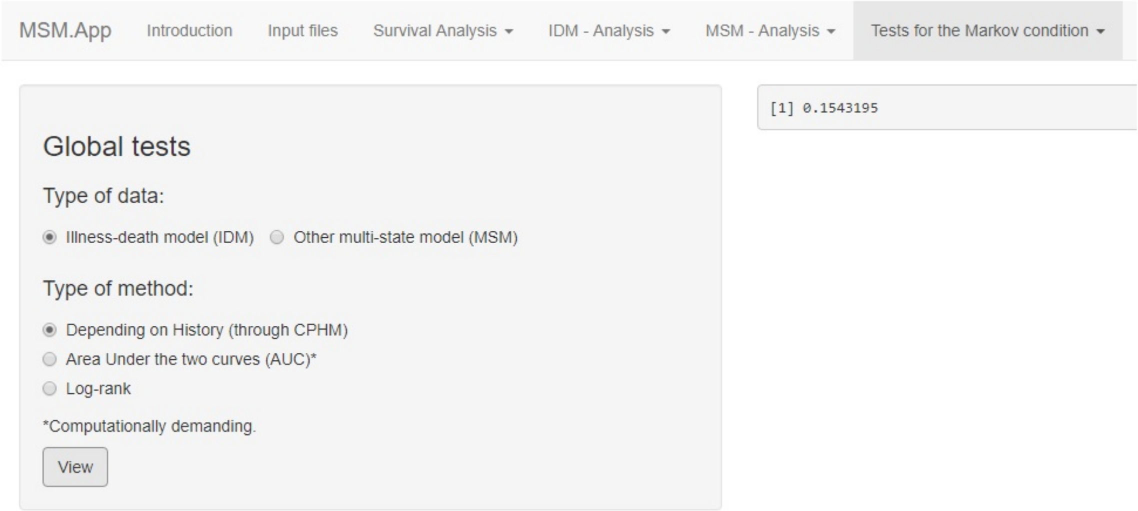
Replicas

100

```
     s       2->2      2->3
1  365 0.003706234 0.0049639
2  730 0.123566948 0.1505162
3 1095 0.171079269 0.1736436
4 1460 0.593134995 0.6029005
5 1825 0.245509888 0.4472789
```

FIGURE 8.26: Results of the local test for the illness-death model using the colon cancer data set, for $s = 365, 730, 1095, 1460$, and 1825 days, using the AUC test, from state 1 to state 3.

until entering the first state of a particular transition. As illustrated in Figure 8.27, we can conclude that there is no effect of the time spent in the initial state on the transition $2 \to 3$ (p-value = 0.1543195). The results were obtained using the Cox Proportional Hazards Model (CPHM) test for the IDM model based on the colon cancer data. In the analysis of more complex models, the user also needs to indicate which states correspond to the drop lists "from" and "to" that appear on the web tool form for this case; (ii) the recent global test proposed by Soutinho and Meira-Machado (2021) [108], based on the area under the curves (AUC), can be used. This test is based on the (AUC) local test results for specific percentiles. To do this, as shown in Figure 8.28, one only needs to indicate the type of test AUC, the corresponding states from ("1") and to ("5") and the number of replicas to compute the results, in this case, 100. The outputs on the right hand show the proportion of rejections of the test for all possible transitions between state 1 and state 5. For IDM models, the user only needs to indicate the number of replicas for the global test; (iii) it is also possible to use the global test based on the log-rank statistics (Titman and Putter, 2020) throughout the similar steps of the previous methods, after selecting Log-rank in the radio button `HTML` element. The outputs provide only the results of the tests for each transition indicated in the drop lists.



Global tests

Type of data:

● Illness-death model (IDM) ○ Other multi-state model (MSM)

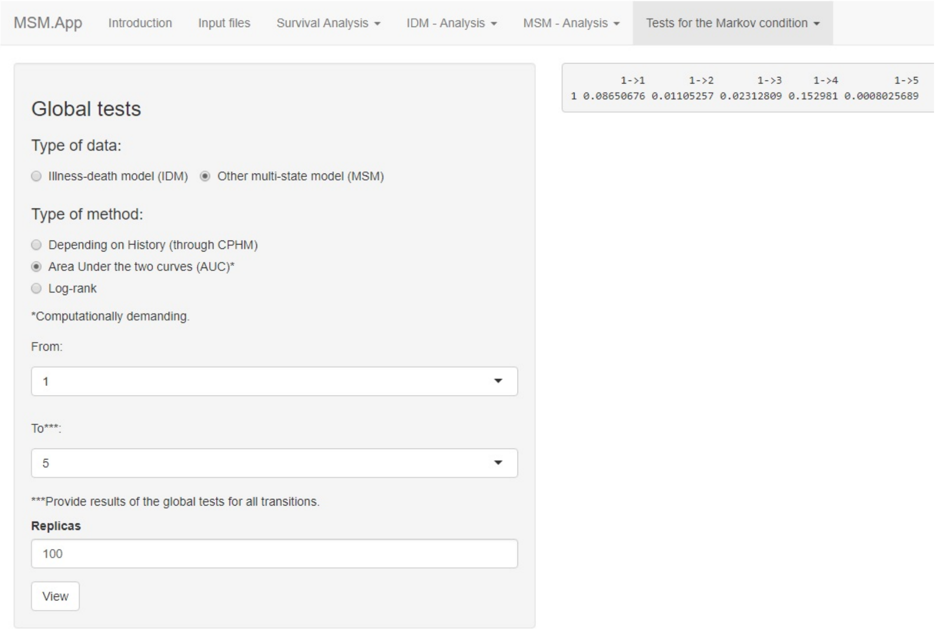
Type of method:

● Depending on History (through CPHM)
○ Area Under the two curves (AUC)*
○ Log-rank

*Computationally demanding.

```
[1] 0.1543195
```

FIGURE 8.27: Results for this global test given by the Cox PH model to our data indicated that the effect of the time spent in State 1 is not significant (p-value of 0.154), revealing no evidence against the Markov model for the colon data set.



Global tests

Type of data:

○ Illness-death model (IDM) ● Other multi-state model (MSM)

Type of method:

○ Depending on History (through CPHM)
● Area Under the two curves (AUC)*
○ Log-rank

*Computationally demanding.

From: 1

To***: 5

***Provide results of the global tests for all transitions.

Replicas 100

```
        1->1       1->2       1->3     1->4         1->5
1 0.08650676 0.01105257 0.02312809 0.152981 0.0008025689
```

FIGURE 8.28: Outputs of the global test for the illness-death model based on the *ebmt* data set using the `AUC` test, from state 1 to state 5. Results for the `AUC` local test are also shown.

8.4 Discussion

The shiny package has grown in popularity as a means of displaying interactive outputs and graphs via web applications. Once developed, the web tools can be shared via the cloud and run on any browser, allowing any user, regardless of computing skills, to dynamically analyze data sets. In terms of biomedical applications, following this increasing interest in this type of web tool, there have been some recent contributions in the literature that enable carring out multi-state analysis. In this chapter we have presented the MSDshiny, the MSM-shiny and the MSMplus applications from which it is possible to perform some specific aspects of the multi-state model analysis, such as planning and powering clinical trials, simulation studies, obtaining results for the inference of transition probabilities or identifying factors that influence the transition intensities of the models (Peterson (2019) [144], Lacy (2021) [145] and Skourlis *et al.* (2021) [146]).

However, although the importance of these models is well recognized, after analyzing existing web applications, we can conclude their use by non-statisticians has been limited. An important reason for this is the lack of friendly software that covers the main goals involving survival analysis and multistate models on the same platform.

For this reason, we have developed a shiny application called MSM.app that allows users to explore various types of multi-state models and perform regression inference as well as obtain several predictive measures of interest, such as the occupation probabilities, the transition probabilities, and the cumulative incidence functions. Recent methods for checking the Markov assumption are also implemented. Throughout the chapter, we have highlighted from three data sets all the main functionalities of the web application and the steps to carry out the analysis. As part of future research work, we plan to constantly update MSM.app to improve its limitations and to cope with recent developments.

This software is available at the Shiny Apps repository https://gsoutinho.shinyapps.io/appmsm/.

Chapter 9

Conclusions and future research

In this thesis, we presented several methodological contributions concerning statistical issues encountered in multi-state models. In this context, we addressed special attention to the estimation of transition probabilities, namely, through the extension of the recent landmark approach to include presmoothing methods of estimation or by including repeated measures and event history data using the joint modeling. Due the importance of checking the Markov condition in the inference in multi-state models, we have also proposed new tests by measuring the discrepancies between estimators that do not rely of this assumption and the Aalen-Johansen estimator (that provides consistent estimates in Markovian processes). All methods proposed in this thesis were analyzed from simulation studies and applied to biomedical data sets. Next, we go through the main conclusions drawn from the results obtained and raise some open questions that motivate future research.

In Chapter 2, we provided practical algorithms for simulating data from a wide class of multivariate copulas. They are suitable for this purpose since they can be used to introduce dependence between time and covariates, or between times of different transitions in more complex survival systems. The dependence measures involving copulas given by Kendall's tau τ or Spearman's ρ were also presented. The algorithms of copulas described are based on three of the most used techniques for generating multivariate data from copulas: the conditional distribution method (such as, Clayton, Frank, FGM or AMH); based on the bivariate distribution of the copula or sampling algorithms based on numerical inversion of Laplace transforms. Finally, four types of survival data and random number generation involving time-to-event data, recurrent events data, competing risks data and progressive illness-death multi-state models were introduced.

In Chapter 3, we proposed new non-parametric estimators (`WCH`) for the estimation of transition probabilities for cases that the process is not necessarily Markovian using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. Several simulation studies were conducted for illness-death models. Results confirmed the good accuracy of the proposed estimator comparing to the other. For the transition probabilities (p_{22}) this new estimator is equivalent to the landmark estimator (`LM`). Application to a real data set of colorectal cancer also revealed for cases where it was visible a failure of the Markov assumption that the `LM` and `WCH` estimators are preferred over the Aalen-Johansen estimator.

The landmark approach is a recent contribution in the literature that allows the improvement the performance of the transition probability estimation, by reducing bias, in case of failure of the Markov condition in multi-state models. Nevertheless, since this methodology is based on (differences between) Kaplan-Meier estimators obtained from a subset, when the sample sizes are small this may lead to increase the variance of the estimates, as we shown from simulation studies. In Chapter 4, we introduced presmoothed estimators to improve of the accuracy of the estimates by replacing the indicator variable of the LM estimators. To model the binary regression function $p(t)$ some parametric families were considered such as logit, probit and generalized additive logistic models. It was also considered a non-parametric modeling of $p(t)$ based on the Nadaraya-Watson kernel estimator.

In this thesis, we were also interested to include the effect of the progression of the disease among states, for each individual, on the estimation of transition probabilities. To account the trajectory of repeated markers in multi-state models, in Chapter 5, we proposed a new estimators (`JMLM`) based on a joint modeling between the longitudinal sub model (that includes random effect errors associated to the individuals) and survival data (that are represented by the transition intensities for each transition of the multi-state process) under the landmark approach. These new estimators for the inference of the transition probabilities were compared to the standard Breslow's method (that comprises for each individual only a single observation for covariate), and the nonparametric LM estimator (that provides the same estimates for all individuals). To illustrate the ability of the proposed method to deal with the evolution of the repeated measures, we have used two data sets obtained from simulation. From the result, we can conclude that the estimation of transition probabilities conditionally on covariates observed with repeated measures seems to be more efficient than competing estimators that do not take in consideration all information across the states of the process.

In Chapter 6, two new tests for testing the Markov condition in multi-state models were proposed. The 'local' test is based on the areas under the two curves (i.e., the curves of the estimated transition probabilities) that can be used for a general multi-state model (considering the `LMAJ` estimators). To approximate the distributions of the test statistic, bootstrap method, with a large number of resamples, was used. The test rejects the null hypothesis of Markovianity when the absolute value of the standard test statistic is above of the critical value (1.96, in case of 95% of confidence). We also proposed a 'global' test given by a grid of points obtained from the percentiles 5, 10, 20, 30 and 40 of the sojourn time in State 1. For general multi-state models, is recommended the use of the same percentiles of the subject specific arrival time at the corresponding state. Results of simulation and the application to three real data sets reported that the proposed 'global' test has better accuracy to identify failure of Markovianity being much more powerful than the standard parametric method based on the proportional hazard specification which relies on a priori model specification that may fail in practice. The use of 'local' tests is also recommended whenever the interest is focused on the estimation of the transition probabilities. From a grid of fixed values, for the different scenarios, the proposed 'local' test confirmed the ability to detect a failure of the Markov condition. This is more evident in the non-Markov scenario. Simulation results also revealed that the proposed 'local' test and the log-rank test have similar power to identify the failure of the Markovianity.

An important aspect of the statistical research is the development of user-friendly software to facilitate the use of new methodologies. With this regard, in Chapter 7, we detailed described the implementation of the `survidm` `R` package whose contents are currently under review. However, a description of the main functionalities of other packages covering the methods proposed in this thesis are available as supplementary material [A]. Due to the increasing importance of dynamic visualization of results and graphs in biomedical applications during the period of this thesis, we have developed a web application, called `MSM.app`, to perform an interactive multi-state survival data analysis by using a user-friendly interface. In Chapter 8, we introduced the main functionalities of this web tool, which are available at the Shiny Apps repository at `https://gsoutinho.shinyapps.io/appmsm/`. At present, the `MSM.app` allows users to perform a traditional survival analysis, and the main goals described in this thesis involve progressive illness-death models and the extension to more complex multi-state models. It is also possible to use some of the most recent methods for checking the Markov assumption, such as Soutinho and Meira-Machado (2021) [108] and Titman and Putter (2020) [113]. As a result of this work, we have already submitted a paper for publication [B]. As next steps, since this type of web application is always an unfinished project, we are interested in updating the MSM.app application with the new findings of the future research work.

Most of the topics addressed in this thesis may be and should be extended in several research directions. During the course of this thesis, some research work has already been done apart from that which has been present in this thesis. As an example, in terms of transition probabilities, we have developed new methods for improving the accuracy of the transition probabilities under the landmark approach. In particular, we have already submitted a paper to a journal where we propose new estimators that make use of the flexibility of the generalized gamma distribution to model survival functions [B]. Results drawn from simulation studies and the application to a cancer data set confirm a decrease in the variance for both the new proposed estimators when compared to the non-parametric LM estimator for small subsets. However, in some cases, the use of LM estimators may be preferable to avoid misspecification of the GGLM in short lag times $(t - s)$.

As future work, we also plan to apply our proposed methods to the estimation of transition probabilities, namely, through the extension of to cope with left truncation, (ii) to interval censoring, and (iii) to the illness-death model with recovery as well as to more complex models involving several transient states. Still regarding the estimation of transition probabilities conditional covariates given by repeated measures (Chapter 5), we also intend to investigate our methods to more complex longitudinal models comprising several covariates and more complicated random errors. To this end, the author of this thesis has already started a academic collaboration with the Hospital Nossa Sra. da Oliveira, in Guimarães, in order to access a biomedical information and data sets for future publications. He has also established partnerships with doctors of hospitals that resulted in other publications beyond those in the scope of this thesis.

We are also interested in developing new approaches to estimate the marginal and joint distribution functions for recurrent event data. To this end, we have already submitted a paper to a journal in which we introduce new nonparametric estimators and their extensions to several gap times are also given. Nonparametric inference based on current or past covariate measures is also taken into account [B]. Due to computational demand, we did not apply our methods for checking the Markov condition to complex multistate models by simulation studies. In future work, we aim to confirm that the accuracy of the 'local' and 'global' tests to identify failures of the Markov assumption in this type of multi-state model is in accordance with what we have observed in this thesis through the application to real data sets with more than three states and reversible transitions.

ROC (Receiver Operating Characteristic)-based approaches were first developed for classification studies with categorical outcomes. They have been extensively used in biomedical studies because of their flexibility and robustness. ROC curves are most useful when predictions are continuous and the problem is to compute the sensitivity for various thresholds of a marker or combination of markers (Li and Ma (2011) [154]). The AUC (Area under the ROC curve) provides a quantitative measure of the ROC curve and summaries the discrimination accuracy of a test (Chambless and Diao (2006) [155]). Many disease outcomes are time depend and ROC curves that vary as a function of time may be more appropriate (Heagerty, Lumley and Pepe (2000) [156]). To incorporate the time-varying nature of the clinical onset time of the disease, various definitions of time-dependent sensitivity, specificity and ROC curves have been proposed (Zheng, Cai and Feng (2006) [157]). As future research work, we also aim to develop new nonparametric estimators of the cumulative-dynamic time-dependent ROC curve that allows accounting for the possible modifying effect of covariate measures on the discriminatory power of the biomarker. Several methods may be considered, for example the use of single-index models. To account for the covariate effect, one standard method (particularly well-suited to the setting with multiple covariates) could consider estimators based on a Cox's model. However, besides of imposing the so-called proportional hazards assumption, these methods also rely on a parametric specification of the covariates' effects on the intensity functions. In this topic, we should aim to study the implementation of these methods to the case of flexible additive Cox models (e.g. using a P-spline fit).

Appendix A

Supplementary material

A.1 survCopula: a R package for multivariate Dependence Modeling with Copulas for survival data

This software and source code are all available at the GitHub repository at https://github.com/gsoutinho/survCopula. Below, we provide a specific description and details on the usage of each functions of the package.

The function dgCopula performs the simulation of survival data for the mortality model. The description of arguments of the functions are presented in Table A.1.

For illustration purposes, suppose we are interested to simulate survival data for the mortality model. One possibility would be using a bivariate copula with marginal functions uniformly distributed on (0; 5), where the survival (denoted by T) could be for instances the survival time (in years) of lung cancer since diagnosis, and tumor size (in cm) is a covariate value measured for each individual. Individuals alive at the end of the follow-up have right censored observations (i.e., Delta = 0). Such data can be obtained using the dgCopula function through the following input commands:

```
> library(survCopula)
> setseed(2345)
> sim.data<-dgCopula(typeCopula ='clayton', theta=1,
            typeX='Unif', num1_X=0, num2_X=5,
            typeY='Unif',  num1_Y=0, num2_Y=5,
            typeCens='Unif', num1_Cens=0, num2_Cens=7,
            nsim=250, typeSurvData='time-to-event')
> head(sim.data)
```

Argument	Description
`typeCopula`	Type of copula. Possible options are `clayton`, `frank`, `FGM`, `AMH`, `gumbel-hougaard` and `joe`. Defaults to `clayton`.
`theta`	A numeric value for the space parameter.
`typeX`	Type of marginal distribution. Possible options are `Exp`, `Norm`, `Unif` and `Gamma`. Defaults to `Exp`.
`num1_X`	A numeric value for the first parameter of the first marginal distribution.
`num2_X`	A numeric value for the second parameter of the first marginal distribution. Only required for two parameter distributions.
`typeY`	Type of marginal distribution. Possible options are `Exp`, `Norm`, `Unif` and `Gamma`. Defaults to `Exp`.
`num1_Y`	A numeric value for the first parameter of the second marginal distribution.
`num2_Y`	A numeric value for the second parameter of the second marginal distribution. Only required for two parameter distributions.
`typeCens`	Type of censuring distribution. Possible options are `None`, `Unif`, `Exp` and `Wei`. Defaults to `None`.
`num1_Cens`	A numeric value for the first parameter of the censoring distribution.
`num2_Cens`	A numeric value for the second parameter of the censoring distribution. Only required for two parameter distributions.
`typeSurvData`	Type of survival data. Possible options are `time-to-event`, `recurrent`, `competing-risks` and `illness-death`. Defaults to `illness-death`.
`state2.prob`	Probability of a individual move to the intermediate state in illness-death model. Only required if typeSurvData ='illness-death'. Default to 0.7.
`nsim`	Number of observations to be generated.

TABLE A.1: Summary of the arguments of the function `dgCopula`.

```
          T         Z  Delta
  1 3.3786641 5.3939928     1
  2 4.5925602 6.3436964     1
  3 1.9646380 1.9646380     0
  4 0.5421364 5.1900408     1
  5 0.4418575 0.5881083     1
  6 2.1502214 2.1502214     0
```

Following the same procedure, a simulated survival data in an illness-death model could be given by this imput commmads:

```
> sim.data2<-dgCopula(typeCopula ='frank', theta=10,
                     typeX='Exp', num1_X=0.5,
                     typeY='Exp', num1_Y=1.5,
                    typeCens='Unif', num1_Cens=0, num2_Cens=4,
                    nsim=250, typeSurvData='illness-death',
                    state2.prob=0.6)
> head(sim.data2)
      T1      Delta1    T       Delta     Z
1 1.6494293      1 1.6494293     1 1.7036172
2 0.1866107      1 0.3812326     1 2.4967757
3 0.2455007      0 0.2455007     0 0.2455007
4 1.3421718      1 1.3421718     1 2.0920786
5 0.7569676      0 0.7569676     0 0.7569676
6 2.9718201      0 2.9718201     0 2.9718201
```

The function `copula` performs a random number generation for Bivariate Copula Functions. Only returns a single pair of random values from a bivariate copula with marginal distributions X and Y. The arguments of this function are presented in Table A.2.

Argument	Description
`v1`	A numeric value belong to the interval [0,1], corresponding to the cumulative density of the first marginal distribution.
`v2`	A numeric value belong to the interval [0,1], corresponding to the cumulative density of the first marginal distribution.
`theta`	A numeric value for the space parameter.
`type`	Type of copula. Possible options are `clayton`, `frank`, `FGM`, `AMH`, `gumbel-hougaard` and `joe`. Defaults to `clayton`
`typeX`	Type of marginal distribution. Possible options are `Exp`, `Norm`, `Unif` and `Gamma`. Defaults to `Exp`.
`num1_X`	A numeric value for the first parameter of the first marginal distribution. Defaults to `Exp`
`num2_X`	A numeric value for the second parameter of the first marginal distribution. Only required for two parameter distributions.
`typeY`	Type of marginal distribution. Possible options are `Exp`, `Norm`, `Unif` and `Gamma`. Defaults to `Exp`.
`num1_Y`	A numeric value for the first parameter of the second marginal distribution.
`num2_Y`	A numeric value for the second parameter of the second marginal distribution. Only required for two parameter distributions.

TABLE A.2: Summary of the arguments of the function `copula`.

As result, the function provides 2-dimensional vector for the random variables as we can see through the following input commands for the copulas `clayton` and `AMH`:

```
> clay<-copula(0.6, 0.4, theta=2, type='clayton', typeX='Exp', num1_X=0.56,
        typeY='Exp', num1_Y=0.90)
```

```
> clay
[1] 1.6362334 0.8804935

>AMH<-copula(0.4, 0.4, theta=0.59, type='AMH', typeX='Norm', num1_X=0.56,
      num2_X=0.3, typeY='Gamma', num1_Y=0.90, num2_Y=0.30)
> AMH
[1] 0.4839959 0.1231871
```

The function `rcopula` allows to obtain random number generation for Bivariate Copula Functions. Returns a number equal to the indicated size sample of pairs of random values from a bivariate copula with marginal distributions X and Y. The arguments of the function are presented in Table A.3:

Argument	Description
typeCopula	Type of copula. Possible options are clayton, frank, FGM, AMH, gumbel-hougaard and joe. Defaults to **clayton**
theta	A numeric value for the space parameter.
typeX	Type of marginal distribution. Possible options are Exp, Norm, Unif and Gamma. Defaults to Exp.
num1_X	A numeric value for the first parameter of the first marginal distribution.
num2_X	A numeric value for the second parameter of the first marginal distribution. Only required for two parameter distributions.
typeY	Type of marginal distribution. Possible options are Exp, Norm, Unif and Gamma. Defaults to Exp.
num1_Y	A numeric value for the first parameter of the second marginal distribution.
num2_Y	A numeric value for the second parameter of the second marginal distribution. Only required for two parameter distributions.
nsim	Number of observations to be generated.

TABLE A.3: Summary of the arguments of the function rcopula.

For illustration, the 2-dimensional random vector with the results of the simulation considering two size samples 250 can be obtained through the input commands:

```
> res1<-rcopula(typeCopula = 'clayton', theta = 2, typeX='Exp', num1_X=0.9,
               typeY='Exp', num1_Y=0.3, nsim=1000)

> head(res1)

       X           Y
1 1.9802811  4.3365143
2 3.0396864 14.1342913
3 0.2920836  6.7020863
4 0.3780575  0.9226155
5 0.5374659  5.8153233
6 2.6236131 13.6893266

> res2<-rcopula(typeCopula = 'AMH', theta = 2,
        typeX='Norm', num1_X=0.9, num2_X=0.3,
        typeY='Gamma', num1_Y=3, num2_Y=2, nsim=1000)

> head(res2)

```

```
          X        Y
1 1.4816224 5.379682
2 0.5164510 4.827510
3 0.7882364 2.033678
4 0.5850117 1.641286
5 0.7162522 2.530820
6 0.8377601 1.333447
```

The function `invF` is used to obtain the value of the inverse cumulative distribution and is included inside the previous functions. It is composed of the following four parameters:

Argument	Description
`u`	A numeric value belong to the interval [0,1], corresponding to the cumulative density.
`type`	Type of marginal distribution. Possible options are `Exp`, `Norm`, `Unif` and `Gamma`. Defaults to `Exp`.
`num1`	A numeric value for the first parameter of the marginal distribution.
`num2`	A numeric value for the second parameter of the marginal distribution. Only required for two parameter distributions.

TABLE A.4: Summary of the arguments of the function `invF`.

As result, we obtain numeric values corresponding to the distribution of the marginal function as presented in following two examples:

```
> invF(0.2)
[1] 0.2231436
> invF(0.2,type = 'Norm', num1 = 0.2,num2 = 0.1)
[1] 0.1158379
```

A.2 presmTP: a R package for obtaining unsmoothed and presmoothed estimates of the transition probabilities in the illness-death model

presmTP package is available at the CRAN repository at https://cran.r-project.org/web/packages/presmTP. This package comprises three functions presmTP, summary.pstp and plot.pstp which arguments are described in Tables A.5, A.6 and A.7. Functions summary.pstp and plot.pstp return, respectively, a data.frame or a list containing the estimates of the probabilities and draws the estimated probabilities obtained by presmTP function.

Argument	Description
data	A numeric value to be squared
s	The first time for obtaining estimates for the transition probabilities.
method	The method used to compute the transition probabilities. Possible options are uns, np logit, logit.gam, probit and cauchit. Defaults to uns.
estimand	An optional character string identifying the function to estimate: S for survival function and H for cumulative hazard function. Defaults to S.
bw.selec	An optional (partially matched) character string specifying the method of bandwidth selection. fixed if no bandwidth selection is done, in which case the bandwidth(s) given by the fixed.bw argument is (are) used, plug-in for plug-in bandwidth selection and bootstrap for bootstrap bandwidth selection. Defaults to fixed.
fixed.bw	An optional numeric vector with the fixed bandwidth(s) used when the value of the bw.selec argument is fixed. It must be of length 1 for estimating survival and cumulative hazard functions, and of length 2 for density and hazard functions (in this case, the first element is the presmoothing bandwidth).
bound	An optional numeric vector with the fixed bandwidth(s) used when the value of the bw.selec argument is fixed. It must be of length 1 for estimating survival and cumulative hazard functions, and of length 2 for density and hazard functions (in this case, the first element is the presmoothing bandwidth).

TABLE A.5: Summary of the arguments of the function presmTP.

Argument	Description
object	A fitted pstp object as produced by presmTP.
state_ini	Initial state of the transition. Defaults to state_ini=0.
times	Vector of times; the returned data frame will contain 1 row for each time.

TABLE A.6: Summary of the arguments of the function summary.pstp.

Argument	Description
`x`	A fitted pstp object as produced by `presmTP`.
`state_ini`	Initial state of the transition. Defaults to `state_ini`=0.

TABLE A.7: Summary of the arguments of the function `plot.pstp`.

Below, we provide some examples on the usage of the functions of the **presmTP** package from a real data set.

```
> data("colonIDM")
> #Unsmoothed
> res1<- presmTP(data = colonIDM, s = 365,method = "uns" )
> res1$est0$t[1:10]
 [1] 365 366 369 370 372 374 378 379 380 382
> res1$est0$p01[1:10]
 [1] 0.000000000 0.001430615 0.002861230 0.004291845 0.005722461 0.007153076
    0.008583691 0.010014306 0.012875536 0.014306152
> res1$est1$t[1:10]
 [1] 365 366 372 376 381 382 384 389 390 400
> summary(res1, state_ini=1, time=365*1:5)

Estimation of pij(s=365,t)

        t        p11       p12
1     365 1.00000000 0.0000000
366   730 0.38815789 0.6118421
731  1095 0.15789474 0.8421053
1096 1460 0.10526316 0.8947368
1461 1825 0.09210526 0.9078947
> plot(res1)
> res1$call
presmTP(data = colonIDM, s = 365, method = "uns")
> class(res1)
[1] "Unsmooth" "pstp"
> #Nonparametric
> res2<- presmTP(data = colonIDM, s = 365,method = "np" )
> res3<- presmTP(data = colonIDM, s = 365,method = "np", estimand="S")
> res4<- presmTP(data = colonIDM, s = 365,method = "np", estimand="H")
> res5<- presmTP(data = colonIDM, s = 365,method = "np",
+                  bw.selec="fixed", fixed.bw=30)
> #Presmoothed - Logit
> res6<- presmTP(data = colonIDM, s = 365,method = "logit" )
> summary(res6, state_ini=1, time=365*1:5)

Estimation of pij(s=365,t)

         t        p11       p12
1      365 1.00000000 0.0000000
366    730 0.38817214 0.6118279
731   1095 0.15796118 0.8420388
1096  1460 0.10541699 0.8945830
1461  1825 0.09282557 0.9071744
> #Presmoothed - Logit GAM
> res7<- presmTP(data = colonIDM, s = 365,method = "logit.gam" )
```

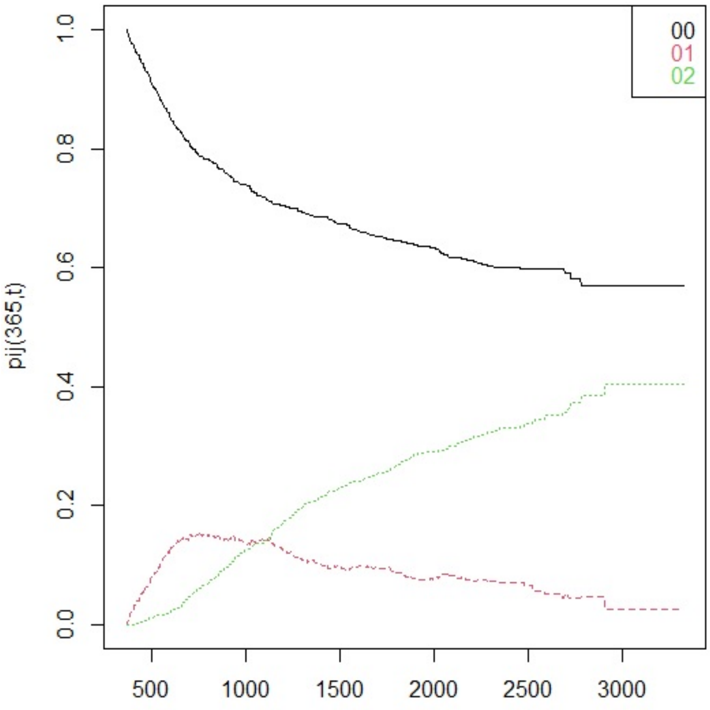


FIGURE A.1: Transition probabilities estimates using unsmoothed estimator for the transitions 0→1, 0→1 and 0→2. Colon cancer data.

Function	Description
`global.test`	Performs a global test for checking the Markov condition using the Area under the two Curves (AUC) method.
`local.test`	Performs a local test of the Markov assumption based on the Area under the two Curves AUC for selected times.
`LR.test`	Log-rank based test for the validity of the Markov assumption.
`plotMSM`	Plot for an object of class `markovMSM`.
`eventsMSM`	Counts the number of observed transitions in the multi-state model.
`prepMSM`	Prepares the data set for multi-state modeling in a long format from a data set in wide format.
`transMatMSM`	Define transition matrix for a multi-state model.
`print.markovMSM`	Print for an object of class
textttmarkovMSM.	
`summary.markovMSM`	Summary for an object of class `markovMSM`.

TABLE A.8: Summary of functions in the `markovMSM` package.

A.3 markovMSM: a R package for testing Markovianity

This section offers the guidelines to use the `markovMSM` package (Soutinho and Meira-Machado (2021) [119]), a software application for R statistical program (R Core Team (2019) [86]), with the purpose to perform the local and global tests for testing the Markov assumption presented in chapter 6. To this end, a description of the functions of the package is illustrated using three real data sets. The first one involve data from a clinical trial on colon cancer modeled using the progressive illness-death model (Moertel *et al.* (1990) [115]). Extensions to progressive processes beyond the three-state illness-death model are discussed using data from the European Group for Blood and Marrow Transplantation (EBMT) (Putter, Fiocco and Geskus (2007) [18]). Finally, we use data from a study with liver cirrhosis patients subjected to a prednisone treatment (Andersen *et al.* (1993) [13]). The package comprises 7 main functions which are briefly summarized in Table A.8.

A.3.1 Data manipulation

Following, we reanalyse data from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer (Moertel *et al.* (1990) [74]). Of the 929 patients, 468 developed a recurrence and among these 414 died; 38 patients died without recurrence. The remaining 423 patients were alive and disease-free up to the end of the follow-up. Besides the two event times (time to recurrence and time to death) and the corresponding indicator statuses, a vector of covariates including age, sex, number of lymph nodes and extent of local spread are also available. Below is an excerpt of the data frame in a wide format with one row per individual. Individuals were chosen in order to represent all possible combinations of movements among the three states.

```
> library(markovMSM)
> data("colonMSM")
> db_wide <- colonMSM
> head(db_wide[c(1:2,16,21),1:11])

   time1 event1 Stime event      rx sex age obstruct perfor adhere nodes
1    968      1  1521     1 Lev+5FU   1  43        0      0      0     5
2   3087      0  3087     0 Lev+5FU   1  63        0      0      0     1
16  1323      1  3214     0     Obs   1  68        0      0      0     1
21  2789      0  2789     1     Obs   1  64        1      0      0     1
```

The four initial variables describe the movement of the patients among the three states of the illness-death model: time1 denote the time measured in days from surgery to recurrence, whereas Stime is the total time or the time to death or censoring; event1 and event denote the corresponding status/censoring indicator (1 for an event and 0 for censoring). Patient 1 had a recurrence after 968 days (i.e., observed a transition from the initial state to the intermediate state) and then he/she died after 1521 days in study. Patient 2 remain alive and without recurrence at the end of follow-up (event1 = 0 and event = 0). The two event times are equal in these cases. Patient represented in the third line had a recurrence after 1323 days but remain alive at the end of the follow-up (i.e. in State 2). Finally, the patient represented in the last line died after 2789 days in study without experiencing a recurrence.

As the original data set is in the wide format, the next step to implement the proposed methods will be to convert the data into a long format which is given by one line for each transition for which a subject is at risk. This can be done using functions transMatMSM and prepMSM. Function defines the transition matrices revealing which transitions are possible, whereas prepMSM provides a new dataset in a long format data for which each row will correspond to a transition for which a patient is at risk. For the progressive illness-death model these two functions are used as follow:

```
> positions<-list(c(2, 3), c(3), c())
> namesStates =  c("Alive", "Rec",  "Death")
> tmat <-transMatMSM(positions, namesStates)
         timesNames = c(NA, "time1","Stime")
         status=c(NA, "event1","event")
> trans = tmat
> db_long<- prepMSM(data=db_wide, trans, timesNames, status)

> db_long[1:10,]

   id from to trans Tstart Tstop time status
1   1    1  2     1      0   968  968      1
2   1    1  3     2      0   968  968      0
3   1    2  3     3    968  1521  553      1
4   2    1  2     1      0  3087 3087      0
5   2    1  3     2      0  3087 3087      0
6   3    1  2     1      0   542  542      1
7   3    1  3     2      0   542  542      0
8   3    2  3     3    542   963  421      1
9   4    1  2     1      0   245  245      1
10  4    1  3     2      0   245  245      0
```

Finally, in terms of manipulation of data, a useful function is `eventsMSM` since it allows to summarise the number of transitions among states and their percentages:

```
> eventsMSM(db_long)

$Frequencies
       to
from    Alive Rec Death no event total entering
  Alive     0 468    38      423             929
  Rec       0   0   414       54             468
  Death     0   0     0      452             452

$Proportions
       to
from        Alive       Rec     Death  no event
  Alive 0.0000000 0.5037675 0.0409042 0.4553283
  Rec   0.0000000 0.0000000 0.8846154 0.1153846
  Death 0.0000000 0.0000000 0.0000000 1.0000000
```

A.3.2 Methods for testing the Markov condition in the illness-death model

Traditionally, the Markov assumption is checked by including covariates depending on the history. In the particular case of the colon cancer data set, we are interested to assess if the transition rate from the recurrence state into death is unaffected by the time spent in the previous state. This can be done using function PHM.test. A a brief description of the arguments of this function is shown in Table A.9. Results for this global test to our data indicated that the effect of the time spent in State 1 is not significant (p value of 0.154) revealing no evidence against the Markov model for the colon data. The corresponding input codes are the following

```
> res <- PHM.test(data=db_long, from=2, to=3)
[1] 0.1543195
> res
$p.value
[1] 0.1543195
$from
[1] 2
$to
[1] 3
```

Argument	Description
db_long	A data frame in the long format containing the subject id; from corresponding to the starting state; the receiving state, to; the transition number, trans; the starting time of the transition given by Tstart; the stopping time of the transition, Tstop, and status (for the status variable, with 1 indicating an event (transition), 0 a censoring).
from	The starting state of the transition to check the Markov condition.
to	The last state of the considered transition to check the Markov condition.

TABLE A.9: Summary of the arguments of the function PHM.test.

In the markovMSM package the local test proposed in Section 2.3 is performed using function AUC.test, through argument type='local'. A summary of the arguments of this function is presented in Table A.10.

The input commands to perform the AUC local test, for a fixed time $s = 180$ and transitions 1→2 and 1→3 are the following

```
> set.seed(1234)
> res2<-AUC.test(db_long, db_wide, times=180, from=1, to=3, type=’local’,
        replicas=100, tmat = tmat)
> res2$localTest
    s    1->1        1->2        1->3
1 180 0.2902191 0.002982042 0.002992007
```

As result, function `AUC.test` returns the probability values, for all attainable transitions from the initial state. To obtain the same local test for transition 2→3, we only need to put 2 in the parameter `from` as follow

```
> set.seed(1234)
> res3<-AUC.test(db_long, db_wide, times=180, from=2, to=3, type=’local’,
        replicas=100, tmat = tmat)
> res3$localTest
    s     2->2         2->3
1 180  0.02547708   0.04816232
```

Results reveal a possible failure of the Markov assumption with low probability values for transitions 1→2 and 2→3 for $s = 180$ (less than 5%). These findings are in agree with results depicted in Figure A.2 that reports the estimated transition probabilities. In fact, we can observe departures between the two Markov-free estimators (`LM` and `LMAJ`) and the Aalen-Johansen estimator (`AJ`) revealing a possible failure of the Markov assumption. The input command to obtain the plots shown in Figure A.2 are the following

```
> plot(res2, to=2, axis.scale=c(0,0.25), difP=FALSE)
> plot(res3, to=3, axis.scale=c(0,1), difP=FALSE)
```

Putting the parameter `difP=TRUE`, we can also obtain the discrepancy between the Aalen-Johansen estimator (Markovian) and the landmark non-Markovian estimator (`LMAJ`), for $p_{12}(s,t)$ and $p_{23}(s,t)$, for $s = 180$, measured through $D_{hj} = \widehat{p}_{hj}^{\texttt{AJ}}(s,t) - \widehat{p}_{hj}^{\texttt{LMAJ}}(s,t)$, $h = 1,2$, $j = h+1$. The 95% pointwise confidence limits were obtained using simple bootstrap (Figure A.3). The corresponding input commands in this case are the following

```
> plot(res2, to=2, axis.scale=c(-0.03,0.03), difP=TRUE)
> plot(res3, to=3, axis.scale=c(-0.30,.10), difP=TRUE)
```

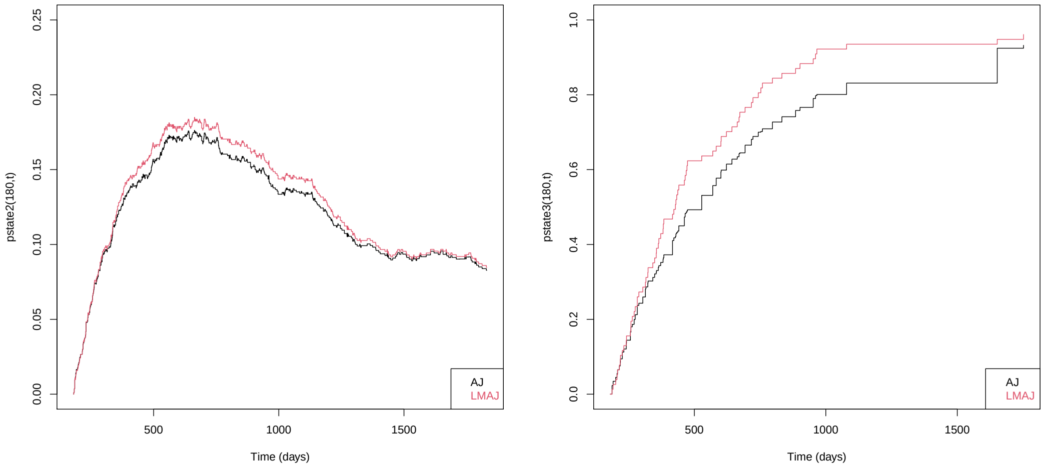


FIGURE A.2: Estimates of the transition probabilities for the Aalen-Johansen (AJ) and Markov-free estimators (landmark and landmark Aalen-Johansen) for $s = 180$. Colon cancer data.

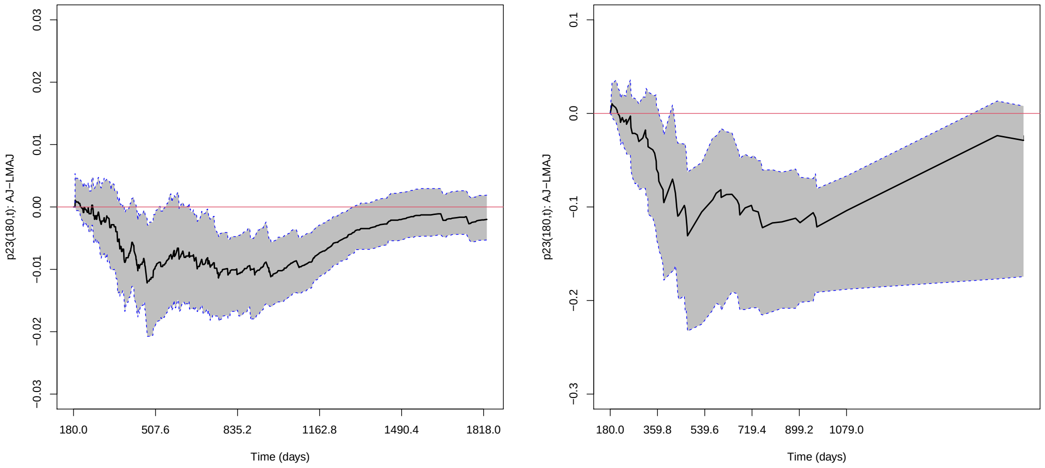


FIGURE A.3: Local graphical test for the Markov condition, for $s = 180$. Test based on the discrepancy between the Aalen-Johansen estimator (Markovian) and the Markov-free estimator (LM). Colon cancer data.

Plots shown in Figure A.3 can be seen as graphical local tests for the Markov assumption. As expected, in both cases they reveal differences between the two methods for $s = 180$ since they show a clear deviation with respect to the straight line $y = 0$. From these one gets some (graphical) evidence on the lack of Markovianity of the underlying process beyond of half an year after surgery. Thereby for this specific time the application of the Aalen-Johansen method may not be recommended here, due to possible biases.

In Section 6.2, a global test was also introduced that combines the probability values of the local test over different times (given by the percentiles of the sojourn time in State 1). In the `markovMSM` package this can be obtained using function `AUC.test`. The arguments of this function are described in Table A.10. Some examples of how to perform the proposed AUC global test are shown in the following input commands, in which we consider the default percentiles in the argument `quantiles`.

```
> set.seed(1234)
> res4<-AUC.test(db_long, db_wide, from=1, to=3, replicas = 100, tmat=tmat)
> round(res4$globalTest,3)
   1->1  1->2  1->3
1 0.067 0.012 0.012

> set.seed(1234)
> res5<-AUC.test(db_long, db_wide, from=2, to=3, type='global', replicas = 100,
        tmat=tmat)
> round(res5$globalTest,3)
    2->3
1  0.006
```

Results reported by the first command lines provide the probability values for the global test based on the AUC for the three transitions leaving State 1 (i.e., 1 → 1, 1 → 2 and 1 → 3). As expected, an higher probability value was obtained for transition 1→1 while the two remaining transitions reveal evidences against the Markov condition. Results reported in the second set of input commands are in agree with previous findings, reporting a probability value of 0.016 for 2→3. Among the objects saved by this function, `AUC.test` displays the probability values for each percentile times (default to 5, 10, 20, 30 and 40) through the following codes

```
> round(res4$localTest,3)
       s   1->1   1->2   1->3
1  102.4  0.978  0.081  0.081
2  173.0  0.118  0.025  0.025
3  290.6  0.015  0.000  0.000
4  469.2  0.679  0.056  0.056
5  726.8  0.635  0.176  0.176

> round(res5$localTest,3)
```

```
      s    2->2   2->3
1 102.4   0.015  0.015
2 173.0   0.011  0.011
3 290.6   0.000  0.000
4 469.2   0.050  0.050
5 726.8   0.156  0.156
```

These outputs show, for instance, that the probability value of the AUC local test for the second percentile time ($s = 173$), is 0.011. Plots with the graphical local tests for the respective quantiles can be easily obtained using the following input commands:

```
> plot(res4, quantileOrder=3, axis.scale=c(-0.04, 0.02))

> plot(res5, quantileOrder=3, axis.scale=c(-0.10, 0.20))
```

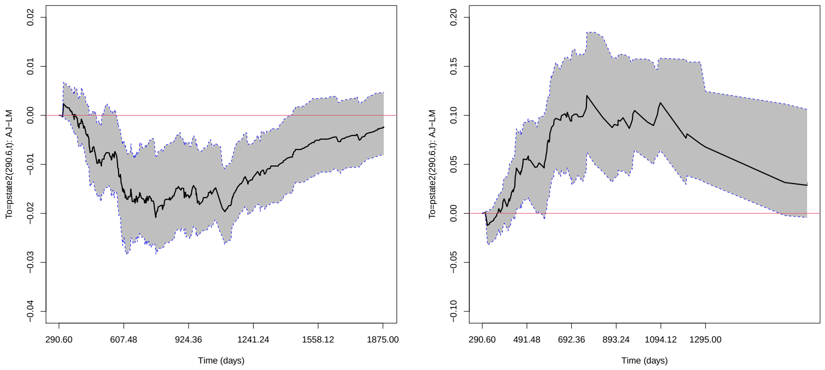


FIGURE A.4: Local graphical test for the Markov condition, for s equal to 173 (2nd percentil sojourn time in State 0. Colon cancer data.

The plots shown in Figure A.4 display the differences between the AJ and LMAJ estimates for the second quantile (quantileOrder) of the sojourn time in the initial state being in accordance with the p-values obtained in the local tests.

The markovMSM package can also be used to compute the results of the global and local tests proposed by [113] which is based on log-rank statistics. A summary of the arguments of the LR.test function is shown in Table A.11. The following input commands illustrate the usage of LR.test function to implement these tests:

```
> set.seed(1234)
> res6 <- LR.test(db_long=db_long, times=180, from = 2, to = 3, replicas = 1000)
> res6$globalTestLR
 [1] 0.047
```

Results of the local and the global test based on the log-rank statistics also confirm the failure of Markovianity for lower values of s.

A.3.3 Extending the Tests for Markov assumption to more complex multi-state models

Following, we use two data sets to illustrate the extension of the previous functions to more complex multi-state models. As a first example, we consider the data of 2279 patients transplanted at the European Society for Blood and Marrow Transplantation (EBMT) and, as second example data from liver cirrhosis patients subjected to a prednisone treatment. Further datails on the description of the data can be found in Putter, Fiocco and Geskus (2007) [18] and Andersen *et al.* (1993) [13], respectively. The steps on the usage of the functions are quite similar to those introduced for the illness-death model. To extend the proposed local and global tests to more complex models we make use of the LMAJ estimator that produces consistent estimates of the transition probabilities in case of non-Markovianity of the process. We start to consider the data set comprising 2279 patients who suffered a blood cancer and who were treated at the EBMT between 1985 and 1998 after a transplant. The movement of the patients among the six states can be modelled through the multi-state model with the following six states: 'Alive and in remission, no recovery or adverse event' (state 1); 'Alive in remission, recovered from the treatment' (state 2); 'Alive in remission, occurrence of the adverse event' (state 3); 'Alive, both recovered and adverse event' (state 4); 'Alive, in relapse' (treatment failure) (state 5) and 'Dead (treatment failure)' (state 6). In total there are 12 transitions, three intermediate events given by recovery (Rec), adverse event (AE) and a combination of the two (AE and Rec), and two absorbing states: Relapse and Death (Figure A.5).

Since the original data `ebmt4` is in the wide format, before implementing a global test we need to convert it into the long format using functions `transMatMSM`, `prepMSM` before using function `AUC.test` with the argument `type='global'`:

```
> data("ebmt4")
> db_wide <- ebmt4
> positions=list(c(2, 3, 5, 6), c(4, 5, 6), c(4, 5, 6),
```

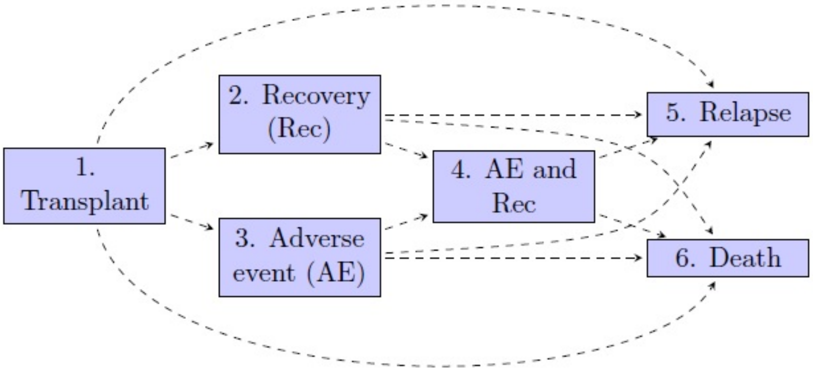


FIGURE A.5: A six-states model for leukemia patients after bone marrow transplantation.

```
                 c(5, 6), c(), c())
> namesStates =  c("Tx", "Rec", "AE", "Rec+AE", "Rel",  "Death")
> tmat <-transMatMSM(positions, namesStates)
> timesNames = c(NA, "rec", "ae","recae", "rel", "srv")
> status=c(NA, "rec.s", "ae.s", "recae.s","rel.s", "srv.s")
> trans = tmat
> db_long<- prepMSM(data=db_wide, trans, timesNames, status)
> db_long[1:10,]

Data:
   id from to trans Tstart Tstop time status
1   1    1  2     1      0    22   22      1
2   1    1  3     2      0    22   22      0
3   1    1  5     3      0    22   22      0
4   1    1  6     4      0    22   22      0
5   1    2  4     5     22   995  973      0
6   1    2  5     6     22   995  973      0
7   1    2  6     7     22   995  973      0
8   2    1  2     1      0    12   12      0
9   2    1  3     2      0    12   12      1
10  2    1  5     3      0    12   12      0

> set.seed(1234)
> res7<-AUC.test(db_long, db_wide, from=1, to=5, type='global',
        quantiles=c(.05, .10, .20, .30, 0.40),
        tmat = tmat, replicas = 100,
        positions=positions, namesStates=state.names,
```

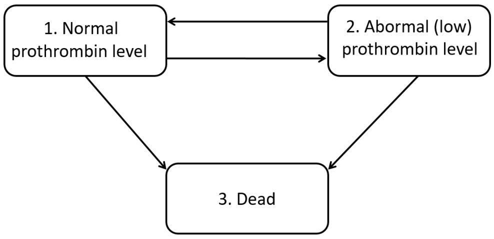

1. Normal prothrombin level
2. Abormal (low) prothrombin level
3. Dead

FIGURE A.6: The reversible illness-death model for patients with liver cirrhosis.

```
        timesNames=timesNames, status=status)

> round(res7$globalTest, 4)
    1->1    1->2    1->3   1->4    1->5
1  0.1423  0.0099 0.0234  0.1016  0.0017

> round(res7$localTests,4)
     s   1->1    1->2    1->3    1->4    1->5
1  9.9  0.2805  0.5994  0.0187  0.0891  0.0040
2 12.0  0.2844  0.5248  0.0282  0.4794  0.0000
3 15.0  0.0001  0.2018  0.0404  0.0177  0.0034
4 18.0  0.4610  0.0051  0.1811  0.1855  0.0015
5 21.0  0.0067  0.0147  0.4796  0.2471  0.0287
```

The interpretation of the output for the global test and for local tests of the five percentile times are similar to those shown for the illness-death model. Thus, object `res7` is used to get a list with components giving the probability values for the Markov test for all transitions leaving State 1 (i.e., for 1→ j with $j \in \{1, 2, 3, 4, 5\}$). For instance, the probability value 0.0099 correspond to the result of the AUC global test for transition 1→2, while 0.0147 is the probability value for the local test, for $s = 21$ for the same transition.

The proposed methods can be also used in reversible multi-state models such as those applied to the data set of liver cirrhosis patients who were included in a randomized clinical trial at several hospitals in Copenhagen between 1962 and 1974. The study aimed to evaluate whether a treatment based on prednisone prolongs survival for patients with cirrhosis (Andersen *et al.* (1993) [13]. State 1 corresponds to 'normal prothrombin level', State 2 to 'low (or abnormal) prothrombin level', and the State 3 to 'dead'. The movement of the patients among these three states can be modeled using the reversible illness-death model shown in Figure A.6.

Note that the original data set *prothr* is already in the long format. Thus to obtain the probability values for the transitions started in State 2, the input command is the following

```
> set.seed(1234)
> res8 <- AUC.test(db_long = prothr, db_wide = NULL, from=2, to=3,
      type='global', replicas=100, limit=0.90,
      quantiles=c(.05, .10, .20, .30, 0.40))
> round(res8$globalTest,5)
     2->1    2->2    2->3
1 0.00063 0.00067 0.30514

> round(res8$localTests,4)
      s   2->1    2->2    2->3
1  73.5  0.0021  0.0006  0.2092
2 117.0  0.0009  0.0007  0.6885
3 223.0  0.0003  0.0021  0.2566
4 392.0  0.0085  0.0401  0.3537
5 681.0  0.2047  0.2825  0.6736
```

The interpretation of the output for the global test is the same shown previously and therefore, for instance, the probability value for transitions 2→1 and 2→3, are 0.00063 and 0.30514, respectively; the result for the local test for $s = 117$ for 2→3 is 0.6885.

Below we report, for the same data set `prothr`, the results for global and local tests proposed by [113], which are based on log-rank statistics, for transition 4 (between the states 2 and 3) and transition 3 (between the states 2 and 1), with times corresponding to the percentiles 5, 10, 20, 30 and 40 (the default percentile values also used by the AUC global test). The corresponding input commands are the following

```
> set.seed(1234)
> times <- c(73.5, 117, 223, 392, 681)
> res9 <- LR.test(db_long=prothr, times=times, from = 2, to = 3, replicas = 1000)

> res9$localTestLR
[1] 0.907 0.330 0.758 0.516 0.193

> res9$globalTestLR
[1] 0.576
```

```
> set.seed(1234)
> res10 <- LR.test(db_long=prothr, times=times, from = 2, to = 1, replicas =
    1000)

> res10$localTestLR
[1] 0.012 0.007 0.107 0.044 0.500

> res10$globalTestLR
[1] 0.012
```

The interpretation of results is similar to those obtained through the `AUC.test` function. Thus, for instance for $s = 681$, the probability values, for transitions 2→3 and 2→1, are 0.193 and 0.50, respectively. The probability values for the global test, are 0.576 and 0.012, respectively. A summary of the arguments of the `LR.test` function is presented in Table A.11.

Argument	Description
`db_long`	A data frame in the long format containing the subject `id`; `from` corresponding to the starting state; the receiving state, `to`; the transition number, `trans`; the starting time of the transition given by `Tstart`; the stopping time of the transition, `Tstop`, and `status` for the status variable, with 1 indicating an event (transition), 0 a censoring.
`db_wide`	Data frame in wide format in which to interpret time, status, id or keep, if appropriate.
`from`	The starting state of the transition probabilities.
`to`	The last receiving state considered for the estimation of the transition probabilities. All the probabilities among the first and the last states are also computed.
`type`	Type of test for checking the Markov condition: `local` or `global`. By default `type='global'`.
`times`	For the local test, times represents the starting times of the transition probabilities. In case of a global test, the argument is given by times between the minimum time and the third quartile times used in the formula of this test. Default to `NULL`.
`quantiles`	Quantiles used in the formula of the global test for the AUC methods.
`tmat`	The transition matrix for multi-state model.
`replicas`	Number of replicas for the Monte Carlo simulation to standardization of the T-statistic given by the difference of the areas of `AJ` and `LMAJ` transition probabilities estimates.
`limit`	Percentile of the event time used as the upper bound for the computation of the AUC-based test.
`positions`	List of possible transitions; x[[i]] consists of a vector of state numbers reachable from state i.
`namesStates`	A character vector containing the names of either the competing risks or the states in the multi-state model specified by the competing risks or illness-death model. names should have the same length as the list x (for `transMat`), or either K or $K+1$ (for `trans.comprisk`), or 3 (for `trans.illdeath`).
`timesNames`	Either 1) a matrix or data frame of dimension $n \times S$ (n being the number of individuals and S the number of states in the multi-state model), containing the times at which the states are visited or last follow-up time, or 2) a character vector of length S containing the column names indicating these times. In the latter cases, some elements of time may be `NA`
`status`	Either 1) a matrix or data frame of dimension $n \times S$, containing, for each of the states, event indicators taking the value 1 if the state is visited or 0 if it is not (censored), or 2) a character vector of length S containing the column names indicating these status variables. In the latter cases, some elements of status may be `NA`

TABLE A.10: Summary of the arguments of function `AUC.test`.

Argument	Description
`db_long`	Multi-state data in `msdata` format. Should also contain (dummy coding of) the relevant covariates; no factors allowed.
`times`	Grid of time points at which to compute the statistic.
`from`	The starting state of the transition to check the Markov condition.
`to`	The last state of the considered transition to check the Markov condition.
`replicas`	Number of wild bootstrap replications.
`formula`	Right-hand side of the formula. If `NULL` will fit with no covariates (formula="1" will also work), offset terms can also be specified.
`fn`	A list of summary functions to be applied to the individual zbar traces (or a list of lists).
`fn2`	A list of summary functions to be applied to the overall chi-squared trace.
`dist`	Distribution of wild bootstrap random weights, either `poisson` for centred Poisson (default), or `normal` for standard normal.
`min_time`	The minimum time for calculating optimal weights.
`other_weights`	Other (than optimal) weights can be specified here.

TABLE A.11: Summary of the arguments of the function `LR.test`.

Appendix B

Other submitted papers

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

markovMSM: An R package for checking the Markov condition in multi-state survival data

Gustavo Soutinho
ISPUP - University of Porto

Luís Meira-Machado
Centre of Mathematics, University of Minho

Abstract

Multi-state models can be successfully used to describe processes in which an individual move through a finite number of states in continuous time. These models allow a detailed view of the evolution or recovery of the process, and can be used to study the effect of a vector of explanatory variables on the transition intensities or to obtain prediction probabilities of future events, after a given event history. In both cases, before using these models, we have to evaluate whether the Markov assumption is tenable. This paper introduces the **markovMSM** package, a software application for R, which considers tests of the Markov assumption that are applicable to general multi-state models. Three approaches using existing methodology are considered: a simple method based on including covariates depending on the history; methods based on measuring the discrepancy of the non-Markov estimators of the transition probabilities to the Markovian Aalen-Johansen estimators; and, finally, methods that were developed by considering summaries from families of log-rank statistics where patients are grouped by the state occupied of the process at a particular time point. The main functionalities of the **markovMSM** package are illustrated using real data examples.

Keywords: Markov assumption, Multi-state models, Transition probabilities.

1. Introduction

Multi-state models have been widely used to analyze complex longitudinal survival data involving several events of interest (Andersen, Borgan, Gill, and Keiding 1993; Hougaard 2000; Putter, Fiocco, and Geskus 2007; Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, and Andersen 2009; Meira-Machado and Sestelo 2019). These models can be considered as a generalization of the survival process where survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. For instances, in some biomedical applications, besides the 'healthy'

ARTICLE TEMPLATE

Parametric landmark estimation of the transition probabilities in survival data with multiple events

Gustavo Soutinho[a] and Luís Meira-Machado[b]

[a]EPIUnit - University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal; [b]Centre of Mathematics, University of Minho, Portugal

ARTICLE HISTORY
Compiled July 13, 2021

ABSTRACT
Multi-state models are a useful tool for analyzing survival data with multiple events. The transition probabilities play an important role in these models since they allow for long-term predictions of the process in a simple and summarized manner. Recent papers have used the idea of subsampling to estimate these quantities, providing estimators with superior performance in case of strong violation of the Markov condition. Subsampling, also referred to as landmarking, leads to small sample sizes and usually to heavily censored data leading to estimators with higher variability. Here, we use the flexibility of the generalized gamma distribution, combined with the same idea of subsampling to obtain estimators free of the Markov condition with less variability. Simulation studies show the good small sample properties of the proposed estimators. The proposed methods are illustrated using real data.

KEYWORDS
Multi-state models; Transition probabilities; Generalized gamma distribution; Landmark approach

1. Introduction

Multi-state models are models for a stochastic process, which at any time occupies one of a set of discrete states [1–4]. These models provide a relevant modeling framework to deal with complex survival data in which individuals may experience more than one event. A multi-state model can be represented schematically by diagrams with boxes representing the states and arrows the possible transitions. The complexity of a multi-state model greatly depends on the number of states defined and also on the transitions allowed between these states.

Among the examples, the simplest case is the mortality model for survival data which involves only two states and one transition. The competing risks model [5,6] can be seen as an extension of the mortality model considering that a subject may reach the ultimate state due to any of several causes. The irreversible illness-death or disability model is a special case of multi-state model, commonly used in the literature to introduce theoretical background of multi-state models, in which the individuals may pass from the initial state (State 1) to the intermediate state (State 2) and then to the absorbing state (State 3) (Figure 1). Individuals are at risk of death in each

CONTACT Gustavo Soutinho. Email: gdsoutinho@gmail.com

Estimation of multivariate distributions for recurrent event data

Gustavo Soutinho[1*] and Luís Meira-Machado[2†]

[1*]EPIUnit, University of Porto, Rua das Taipas 135, 4050-600, Porto, Portugal.
[2]Centre of Mathematics and Department of Mathematics, University of Minho, Campus de Azurém, 4800-058, Guimarães, Portugal.

*Corresponding author(s). E-mail(s): gdsoutinho@gmail.com;
Contributing authors: lmachado@math.uminho.pt;
†These authors contributed equally to this work.

Abstract

In many longitudinal studies information is collected on the times of different kinds of events. Some of these studies involve repeated events, where a subject or sample unit may experience a well-defined event several times along his history. Such events are called recurrent events. In this paper we introduce nonparametric methods for estimating the marginal and joint distribution functions for recurrent event data. New estimators are introduced and their extensions to several gap times are also given. Nonparametric inference conditional on current or past covariate measures is also considered. We study by simulation the behavior of the proposed estimators in finite samples considering two or three gap times. Our proposed methods are applied to the study of (multiple) recurrence times in patients with bladder tumors. Software in the form of an R package has been developed implementing all methods.

Keywords: Censoring, Gap times, Kaplan-Meier, Multiple events, Recurrent events

MSM.app: a Web-Based Tool for the Analysis of Multi-state Survival Data

Gustavo Soutinho[a,], Luís Meira-Machado[b]

[a]EPIUnit - University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal
[b]Centre of Mathematics, University of Minho, Portugal

Abstract

The development of applications for obtaining interpretable results in a simple and summarized manner in multi-state models is a research field with great potential, namely in terms of using open source tools that can be easily implemented in biomedical applications. This paper introduces `MSM.app`, an interactive web application using `shiny` package for the `R` language, which enables any user, regardless of their previous knowledge of informatics, to perform a dynamic analysis involving the most important topics in multi-state models. The `MSM.app` can be used to relate the individual characteristics with the intensity rates through a covariate vector, but can also be used to report additional interpretable results in a simple and summarized manner. It can be used to obtain the results from some newly developed methods, such as the estimation of transition probabilities or the recent methods for checking the Markov assumption. The classical survival analysis can be seen as a particular multi-state model and therefore is also included regarding the estimation of survival curves, the comparison of several curves, or the inference in regression models. The `MSM.app` application comprises a set of dynamic web forms, tables, and graphics whose usability is illustrated using real data examples.

Keywords: R language, Shiny package, Survival analysis, Multi-state models.

Email address: `gdsoutinho@gmail.com` (Gustavo Soutinho)

Bibliography

[1] P. Hougaard, *Analysis of Multivariate Survival Data*, ser. Statistics for Biology and Health. New York: Springer-Verlag, 2000. [Cited on pages 1, 5, and 21.]

[2] J. Klein and M. Moeschberger, *Survival analysis - techniques for censored and truncated data*. New York: Springer-Verlag, 1997. [Cited on pages 1 and 2.]

[3] M. Tableman and J. Kim, *Survival analysis using S*. Chapman & Hall Ltd, 2003. [Cited on page 1.]

[4] D. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*. New York: Springer-Verlag, 2012. [Cited on page 1.]

[5] D. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. John Wiley & Sons, Inc., 2008. [Cited on pages 1 and 3.]

[6] D. Cox, "Regression models and life tables," *Journal of the Royal Statistical Society Series B*, vol. 34, pp. 187–220, 1972. [Cited on pages 3, 6, 7, 49, 74, 75, 100, 133, 148, and 150.]

[7] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, pp. 269–276, 1975. [Cited on page 3.]

[8] C. A. Struthers and J. D. Kalbfleisch, "Misspecified proportional hazard models," *Biometrika*, vol. 73, no. 2, pp. 363–369, 1986. [Cited on page 3.]

[9] G. L. Anderson and T. R. Fleming, "Model misspecification in proportional hazards regression," *Biometria*, vol. 82, pp. 527–541, 1995. [Cited on page 3.]

[10] T. Martinussen and T. H. Scheike, *Dynamic regression models for survival data*. New York: Springer-Verlag, 2006. [Cited on page 4.]

[11] T. Hastie and R. Tibshirani, *Generalized additive models*. Chapman & Hall Ltd, 1990. [Cited on page 4.]

[12] P. Eilers and B. Marx, "Flexible smoothing with b-splines and penalties," *Statistical Science*, vol. 11, pp. 89–121, 1996. [Cited on pages 4 and 124.]

[13] P. Andersen, Ø. Borgan, R. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*. New York: Springer-Verlag, 1993. [Cited on pages 4, 5, 21, 29, 99, 103, 180, 188, and 190.]

[14] P. Hougaard, "Multi-state models: a review," *Lifetime Data Analysis*, vol. 5, pp. 239–264, 1999. [Cited on pages 4 and 5.]

[15] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. Andersen, "Multi-state models for the analysis of time to event data," *Statistical Methods in Medical Research*, vol. 18, pp. 195–222, 2009. [Cited on pages 4, 5, 21, 49, 100, and 141.]

[16] L. Meira-Machado and M. Sestelo, "Estimation in the progressive illness-death model: A nonexhaustive review," *Biometrical Journal*, vol. 61, no. 2, pp. 245–263, 2019. [Cited on pages 4, 5, 7, 21, 100, 118, and 141.]

[17] P. Andersen and N. Keiding, "Multi-state models for event history analysis," *Statistical Methods in Medical Research*, vol. 11, pp. 91–115, 2002. [Cited on pages 4, 68, and 90.]

[18] H. Putter, M. Fiocco, and G. R.B., "Tutorial in biostatatistics: Competing risks and multi-state models," *Statistics in Medicine*, vol. 26, no. 11, pp. 2389–2430, 2007. [Cited on pages 4, 5, 21, 49, 100, 141, 143, 156, 180, and 188.]

[19] R. Gentleman, F. Lawless, J. Lindsey, and P. Yan, "Multi-state markov models for analysing incomplete disease history data with illustrations for hiv diseas," *Statistics in Medicin*, vol. 13, pp. 805–821, 1994. [Cited on page 5.]

[20] P. Andersen, S. Esbjerj, and T. Sorensen, "Multistate models for bleeding episodes and mortality in liver cirrhosi," *Statistics in Medicine*, vol. 19, pp. 587–599, 2000. [Cited on pages 5, 68, and 90.]

[21] R. Pérez-Ocón, J. Ruiz-Castro, and M. Gámiz-Pérez, "Non-homogeneous markov models in the analysis of survival after breast cancer," *Journal of the Royal Statistical Society: Series C*, vol. 50, pp. 111–124, 2001. [Cited on page 5.]

[22] R. Kay, "A markov model for analyzing cancer markers and disease states in survival studies," *Biometrics*, vol. 42, pp. 457–481, 1986. [Cited on pages 6, 90, and 91.]

[23] N. Breslow, "Discussion of paper by dr cox," *Journal of Royal Statistical Society, Series B*, vol. 34, pp. 216–217, 1972. [Cited on pages 7, 75, 133, and 154.]

[24] L. Meira-Machado, J. de Uña-Álvarez, and S. Datta, "Nonparametric estimation of conditional transition probabilities in a non-markov illness-death model," *Computational Statistics*, vol. 30, pp. 377–397, 2015. [Cited on pages 7 and 153.]

[25] M. Rodríguez-Álvarez, L. Meira-Machado, and E. Abu-Assi, "Nonparametric estimation of time-dependent roc curves conditional on a continuous covariate," *Statistics in medicine*, vol. 35:7, pp. 1090–1102, 2016. [Cited on page 7.]

[26] G. Soutinho and L. Meira-Machado, "Some of the most common copulas for simulating complex survival data," *International Journal of Mathematics and Computers in Simulation*, vol. 14, pp. 28–37, 2020. [Cited on page 11.]

[27] A. Burton, D. Altman, P. Royston, and R. Holder, "The design of simulation studies in medical statistics," *Statistics in Medicine*, vol. 25, pp. 4279–4292, 2006. [Cited on page 12.]

[28] A. Sklar, "Fonctions de répartition à n dimension et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959. [Cited on pages 12 and 13.]

[29] D. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, vol. 65, pp. 141–152, 1978. [Cited on page 14.]

[30] M. Frank, "On the simultaneous associativity of f(x,y) and x+y-f(x,y)," *Aequationes Mathematicae*, vol. 19, pp. 124–226, 1979. [Cited on page 14.]

[31] E. Gumbel, "Bivariate exponential distributions," *Journal of the American Statistical Association*, vol. 55, pp. 698–707, 1960. [Cited on pages 14 and 15.]

[32] M. Ali, N. Mikhail, and M. Haq, "A class of bivariate distributions including the bivariate logistic," *Journal of Mulvariate Analysis*, vol. 8, no. 3, pp. 405–412, 1978. [Cited on page 14.]

[33] H. Joe, *Multivariate Models and Dependence Concepts*. Chapman & Hall Ltd, 1997. [Cited on pages 14 and 20.]

[34] D. Farlie, "The performance of some correlation coefficients for a general bivariate distribution," *Biometrika*, vol. 47, pp. 307–323, 1960. [Cited on page 15.]

[35] D. Morgenstern, "Einfache beispiele zweidimensionaler verteilungen," *Mitteilungsblatt fürMathematische Statistik*, vol. 8, pp. 234–235, 1956. [Cited on page 15.]

[36] E. Bouyè, V. Durrleman, A. Bikeghbali, G. Riboulet, and T. Roncall, *Copulas for finance - A reading guide and some applications*. Working paper, Group de Rechercher Opérationnelle, Crédit Lyonnais, 2000. [Cited on page 18.]

[37] C. Genest and L.-P. Rivest, "Statistical inference procedures for bivariate archimedean copulas," *J Amer Statist Assoc*, vol. 36, no. 88, pp. 1034–1043, 1993. [Cited on page 18.]

[38] F. Wu, E. A. Valdez, and M. Sherris, "Simulating exchangeable multivariate archimedean copulas and its applications," *ommunications in Statistics - Simulation and Computation*, vol. 36, no. 5, pp. 1019–1034, 2006. [Cited on page 19.]

[39] A. Marshall and I. Olkin, "Families of multivariate distributions," *Journal of the American Statistical Association*, vol. 83, pp. 834–841, 1988. [Cited on page 20.]

[40] M. Hofert, "Sampling archimedean copulas," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5163–5174, 2008. [Cited on page 20.]

[41] J. Nolan, *Stable distributions - models for heavy tailed data*. Boston, MA: Birkhauser, 2007. [Cited on page 20.]

[42] G. Escarela and J. Carrière, "Fitting competing risks with an assumed copula," *Statistical Methods in Medical Research*, vol. 12, no. 4, pp. 333–349, 2003. [Cited on page 21.]

[43] V. Kaishev, D. Dimitrova, and S. Haberman, "Modelling the joint distribution of competing risks survival times using copula functions," *Insurance: Mathematics and Economics*, vol. 41, no. 3, pp. 339–361, 2007. [Cited on page 21.]

[44] G. Soutinho, L. Meira-Machado, and P. Oliveira, "A comparison of presmoothing methods in the estimation of transition probabilities," *Communications in Statistics - Simulation and Computation*, 2020. [Cited on pages 21, 36, and 55.]

[45] J. J. Harden and J. Kropko, "Simulating duration data for the cox model," *Political Science Research and Methods*, vol. 7, no. 4, pp. 921–928, 2018. [Cited on page 21.]

[46] J. Kropko and J. Harden, "Beyond the hazard ratio: Generating expected durations from the cox proportional hazards model," *British Journal of Political Science*, vol. 50, no. 1, pp. 303–320, 2018. [Cited on page 21.]

[47] R. Cook and J. Lawless, *The Statistical Analysis of Recurrent Events*. New York: Springer-Verlag, 2007. [Cited on pages 21 and 23.]

[48] A. Malehi, E. Hajizadeh, K. Ahmadi, and P. Mansouri, "Joint modelling of longitudinal biomarker and gap time between recurrent events: copula-based dependence," *Journal of Applied Statistics*, vol. 42, no. 9, pp. 1931–1945, 2015. [Cited on page 21.]

[49] F. Rotolo, C. Legrand, and I. Van Keilegom, "A simulation procedure based on copulas to generate clustered multi-state survival data," *Computer Methods and Programs in Biomedicine*, vol. 109, no. 3, pp. 305–312, 2013. [Cited on page 21.]

[50] J. D. Kalbfleisch and P. R. L., *The statistical analysis of failure time data*. New York: John Wiley & Sons, 1980. [Cited on pages 21, 117, and 154.]

[51] J. de Uña Álvarez and L. Meira-Machado, "A simple estimator of the bivariate distribution function for censored gap times," *Statistics & Probability Letters*, vol. 78, pp. 2440–2445, 2008. [Cited on page 21.]

[52] A. Moreira and L. Meira-Machado, "survivalbiv: Estimation of the bivariate distribution function for sequentially ordered events under univariate censoring," *Journal of Statistical Software*, vol. 46, no. 13, pp. 1–16, 2012. [Cited on page 21.]

[53] L. Meira-Machado and M. Sestelo, "condsurv: An r package for the estimation of the conditional survival function for ordered multivariate failure time data," *The R Journal*, vol. 8, no. 2, pp. 460–473, 2016. [Cited on page 21.]

[54] J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher, "Simulating competing risks data in survival analysis," *Statistics in Medicine*, vol. 28, pp. 956–971, 2009. [Cited on page 24.]

[55] G. Soutinho and L. Meira-Machado, “Estimation of the transition probabilities in multi-state survival data: New developments and practical recommendations,” *WSEAS Transactions on Mathematics*, vol. 19, pp. 353–366, 2020. [Cited on page 28.]

[56] L. Ferrer, V. Rondeau, J. Dignam, T. Pickles, H. Jacqmin-Gadda, and C. Proust-Lima, “Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer,” *Statistics in Medicine*, vol. 35, no. 22, pp. 3933–3948, 2016. [Cited on pages 75, 79, 80, and 86.]

[57] A. Allignol and A. Latouche, *CRAN Task View: Survival Analysis, Version 2019-09-01*, 2019. [Online]. Available: http://CRAN.Rproject.org/view=Survival [Cited on pages 25 and 116.]

[58] L. Meira-Machado and S. Faria, “A simulation study comparing modeling approaches in an illness-death model,” *Communications in Statistics - Simulation and Computation*, vol. 43, no. 5, pp. 929–946, 2014. [Cited on page 25.]

[59] M. Crowther and P. Lambert, “Simulating biologically plausible complex survival data,” *Statistics in Medicine*, vol. 32, no. 23, pp. 4118–4134, 2013. [Cited on page 25.]

[60] D. Morina and A. Navarro, “The r package survsim for the simulation of simple and complex survival data,” *Journal of statistical software*, vol. 59, no. 2, pp. 1–20, 2014. [Cited on page 25.]

[61] O. Aalen and S. Johansen, “An empirical transition matrix for nonhomogeneous markov chains based on censored observations statistics in medicine,” *Scandinavian Journal of Statistics*, vol. 5, pp. 141–150, 1978. [Cited on pages 29, 31, 33, 37, 90, and 151.]

[62] J. Beyersmann, M. Schumacher, and A. Allignol, *Competing Risks and Multistate Models with R*. New York: Springer, New York, 2011. [Cited on page 31.]

[63] E. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *Journal of the American Statistical Association*, vol. 53, pp. 457–481, 1958. [Cited on pages 31 and 33.]

[64] A. Moreira, J. de Uña-Álvarez, and L. Meira-Machado, “Presmoothing the aalen-johansen estimator in the illness-death model,” *Electronical Journal of Statistics*, vol. 7, pp. 1491–1516, 2013. [Cited on pages 31 and 32.]

[65] G. Dikta, “On semiparametric random censorship models,” *Journal of Statistical Planning and Inference*, vol. 66, pp. 253–279, 1998. [Cited on pages 31, 56, and 130.]

[66] R. Cao, I. Lopez-de Ullibarri, P. Janssen, and N. Veraverbeke, “Presmoothed kaplan-meier and nelson-aalen estimators,” *Journal of Nonparametric Statistics*, vol. 17, pp. 31–56, 2005. [Cited on pages 31, 58, 61, and 130.]

[67] S. Datta and G. Satten, “Validity of the aalen-johansen estimators of stage occupation probabilities and nelson aalen integrated transition hazards for non-markov models,” *Statistics & Probability Letters*, vol. 55, pp. 403–411, 2001. [Cited on page 32.]

[68] L. Meira-Machado, J. de Uña-Álvarez, and C. Cadarso-Suárez, “Nonparametric estimation of transition probabilities in a non-markov illness-death model,” *Lifetime Data Analysis*, vol. 12, pp. 325–344, 2006. [Cited on pages 32, 33, 34, 37, 40, 52, 75, and 90.]

[69] J. de Uña-Álvarez and L. Meira-Machado, “Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study,” *Biometrics*, vol. 71, no. 2, pp. 364–375, 2015. [Cited on pages 32, 34, 35, 36, 37, 38, 53, 54, 55, 59, 68, 69, 75, 90, 91, 116, and 152.]

[70] H. C. van Houwelingen, “Dynamic prediction by landmarking in event history analysis,” *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 70–85, marzo 2007. [Cited on page 32.]

[71] H. Putter and C. Spitoni, “Non-parametric estimation of transition probabilities in non-markov multi-state models: The landmark aalen-johansen estimator,” *Statistical Methods in Medical Research*, vol. 27, no. 7, pp. 2081–2092, 2018. [Cited on pages 32, 36, 91, and 152.]

[72] O. Borgan, *Encyclopedia of biostatistics: Aalen-Johansen estimator*. John Wiley & Sons, 2005. [Cited on page 33.]

[73] L. Meira-Machado, “Smoothed landmark estimators of the transition probabilities,” *SORT-Statistics and Operations Research Transactions*, vol. 40, no. 2, pp. 375–398, Jul-Dec 2016. [Cited on pages 36, 56, 100, and 152.]

[74] C. G. Moertel, T. R. Fleming, J. S. Macdonald, D. G. Haller, J. A. Laurie, P. J. Goodman, J. S. Ungerleider, W. A. Emerson, D. C. Tormey, J. H. Glick, M. H. Veeder, and

J. A. Mailliard, "Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma," *New England Journal of Medicine*, vol. 322, pp. 352–358, 1990. [Cited on pages 49, 118, 143, and 180.]

[75] L. Meira-Machado, M. Sestelo, and A. Gonlçalves, "Nonparametric estimation of the survival function for ordered multivariate failure time data: A comparative study," *Biometrical Journal*, vol. 58, no. 2, pp. 623–634, 2016. [Cited on page 56.]

[76] R. Cao and M. Jácome, "Presmoothed kernel density estimator for censored data," *Journal of Nonparametric Statistics*, vol. 16, pp. 289–309, 2004. [Cited on page 56.]

[77] J. de Uña-Álvarez and C. Rodríguez-Campos, "Strong consistency of presmoothed kaplan-meier integrals when covariables are present," *Statistics*, vol. 38, pp. 483–496, 2004. [Cited on page 56.]

[78] M. Jácome and M. Iglesias, "Presmoothed estimation of the density function with truncated and censored data," *Statistics*, vol. 44, pp. 217–234, 2010. [Cited on page 56.]

[79] J. de Uña Álvarez and A. Amorim, "A semiparametric estimator of the bivariate distribution function for censored gap times," *Biometrika*, vol. 53, pp. 113–127, 2011. [Cited on page 56.]

[80] A. Amorim, J. de Uña Álvarez, and L. Meira-Machado, "Presmoothing the transition probabilities in the illness-death model," *Statistics & Probability Letters*, vol. 81, no. 2, pp. 797–806, 2011. [Cited on pages 56 and 59.]

[81] M. Wand and M. Jones, *Kernel Smoothing*. Chapman and Hall, 1997. [Cited on pages 56 and 131.]

[82] D. W. Hosmer, S. Lemeshow, and S. R. X., *Applied Logistic Regression, Third Edition*. John Wiley & Sons, 2013. [Cited on pages 58, 61, and 69.]

[83] G. Dikta, M. Kvesic, and C. Schmidt, "Bootstrap approximations in model checks for binary data," *Journal of the American Statistical Association*, vol. 101, pp. 521–530, 2006. [Cited on page 58.]

[84] S. Wood, "mgcv: Mixed gam computation vehicle with automatic smoothness estimation. r package version 1.8-29 (2019)." [Online]. Available: https://cran.r-project.org/web/packages/mgcv [Cited on pages 61 and 71.]

[85] P. S. Group, "Prophylaxis of first hemorrhage from esophageal varices by sclerotherapy, propranolol or both in cirrhotic patients: a randomized multicenter trial," *Hepatology*, vol. 14, pp. 1016–1024, 1991. [Cited on page 67.]

[86] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.r-project.org/ [Cited on pages 71, 118, 140, and 180.]

[87] G. Soutinho, L. Meira-Machado, and P. Oliveira, "presmtp: Methods for transition probabilities." *R package version 1.1.0.*, pp. URL https://cran.r–project.org/web/packages/presmTP, 2019. [Cited on page 71.]

[88] I. López-de Ullibarri and M. Jácome, "survpresmooth: An r package for presmoothed estimation in survival analysis," *Journal of Statistical Software*, vol. 54, no. 11, pp. 1–26, 2013. [Cited on page 71.]

[89] G. Soutinho, L. Meira-Machado, and P. Oliveira, "Estimation of the transition probabilities conditional on repeated measures in multi-state models," *Conference: 35th International Workshop on Statistical Modelling*, 2020. [Cited on page 73.]

[90] Q. Xu and Y. Bai, "Semiparametric statistical inferences for longitudinal data with nonparametric covariance modelling," *Statistics*, vol. 51, no. 6, pp. 1280–1303, 2017. [Cited on page 74.]

[91] L. Edwards, "Modern statistical techniques for the analysis of longitudinal data in biomedical research," *Pediatric Pulmonology*, vol. 30, no. 4, pp. 330–344, 2000. [Cited on page 74.]

[92] G. Molenberghs and G. Verbeke, "A review on linear mixed models for longitudinal data, possibly subject to dropout," *Statistical Modelling*, vol. 1, no. 4, pp. 235–269, 2001. [Cited on page 74.]

[93] G. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint models of longitudinal and time-to-event data with more than one event time outcome: A review," *The International Journal of Biostatistics*, vol. 14, no. 1, pp. 1–19, 2018. [Cited on page 74.]

[94] D. Rizopoulos, “Jm: An r package for the joint modelling of longitudinal and time-to-event data,” *Journal of Statistical Software*, vol. 35, no. 9, pp. 1–33, 2010. [Cited on pages 74, 78, and 80.]

[95] J. Ibrahim, H. Chu, and L. Chen, “Basic concepts and methods for joint models of longitudinal and survival data,” *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796–801, 2010. [Cited on page 74.]

[96] P. Y. Self S., “Modeling a marker of disease progression and onset of disease,” *In: Jewell N.P., Dietz K., Farewell V.T. (eds) AIDS Epidemiology. Birkhäuser*, 1992. [Cited on page 74.]

[97] A. Tsiatis, V. Degruttola, and M. Wulfsohn, “Modelling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids,” *Journal of the American Statistical Association*, vol. 90, pp. 27–37, 1995. [Cited on page 74.]

[98] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, 2012. [Cited on pages 74, 75, and 80.]

[99] P. Williamson, R. Kolamunnage-Dona, P. Philipson, and A. Marson, “Joint modelling of longitudinal and competing risks data,” *Statistics in Medicine*, vol. 27, pp. 6426–6438, 2008. [Cited on page 74.]

[100] L. Liu and X. Huang, “Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome,” *Journal of the Royal Statistical Society, Series C.*, vol. 58, no. 1, pp. 65–81, 2009. [Cited on page 74.]

[101] E.-R. Andrinopoulou, D. Rizopoulos, J. Takkenberg, and E. Lesaffre, “Joint modeling of two longitudinal outcomes and competing risk data,” *Statistics in Medicine*, vol. 33, no. 18, pp. 3167–3178, 2014. [Cited on page 74.]

[102] E. Andrinopoulou, D. Rizopoulos, J. Takkenberg, and E. Lesaffre, “Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data,” *Statistical Methods in Medical Research*, vol. 26, no. 4, pp. 1787–1801, 2017. [Cited on page 74.]

[103] A. Król, A. Mauguen, Y. Mazroui, A. Laurent, S. Michiels, and V. Rondeau, "Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event," *Journal of Statistical Software*, vol. 81, no. 3, pp. 1–52, 2017. [Cited on page 75.]

[104] G. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes," *BMC Medical Research Methodology*, vol. 18, no. 50, pp. 1–14, 2018. [Cited on page 75.]

[105] D. Dabrowska and W. Lee, "Nonparametric estimation of transition probabilities in a two-stage duration model," *Journal of Nonparametric Statistics*, vol. 7, pp. 75–103, 1996. [Cited on page 75.]

[106] L. Azarang, T. Scheike, and J. de Uña-Álvarez, "Direct modeling of regression effects for transition probabilities in the progressive illness–death model," *Statistics in Medicine*, vol. 36, pp. 1964–1976, 2017. [Cited on page 75.]

[107] R. Hoff, H. Putter, I. Mehlum, and J. Gran, "Landmark estimation of transition probabilities in non-markov multistate models with covariates," *Lifetime Data Anal*, vol. 25, no. 4, pp. 660–680, 2019. [Cited on page 76.]

[108] G. Soutinho and L. Meira-Machado, "Methods for checking the markov condition in multi-state survival data," *Comput Stat*, 2021. [Cited on pages 28, 89, 141, 159, 161, and 168.]

[109] A. Allignol, J. Beyersmann, T. Gerds, and A. Latouche, "Nonparametric estimation of transition probabilities in a non-markov illness-death model," *Lifetime data analysis*, vol. 20, no. 4, pp. 495–513, 2014. [Cited on page 90.]

[110] A. Titman, "Transition probability estimates for non-markov multi-state models," *Biometrics*, vol. 71, no. 4, pp. 1034–1041, 2015. [Cited on page 90.]

[111] M. Rodriguez-Girondo and J. de Uña-Álvarez, "A nonparametric test for markovianity in the illness-death model," *Statistics in Medicine*, vol. 31, no. 30, pp. 4416–4427, 2012. [Cited on page 90.]

[112] M. Rodriguez-Girondo and J. Uña-Álvarez, "Methods for testing the markov condition in the illness-death model: a comparative study," *Statistics in Medicine*, vol. 35, no. 20, pp. 3549–3562, 2016. [Cited on pages 90, 94, and 99.]

[113] A. Titman and H. Putter, "General tests of the markov property in multi-state models," *Bio-statistics*, 2020. [Cited on pages 90, 92, 95, 96, 159, 168, 187, and 191.]

[114] S. Chiou, J. Qian, E. Mormino, and R. Betensky, "Permutation tests for general dependent truncation," *Computational Statistics & Data Analysis*, vol. 128, pp. 308–324, 2018. [Cited on page 90.]

[115] C. Moertel, T. Fleming, J. Macdonald, D. Haller, J. Laurie, C. Tangen, J. Ungerleider, W. Emerson, D. Tormey, J. Glick, M. Veeder, and J. Mailliard, "Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii. colon carcinoma: A final report," *The Annals of Internal Medicine*, vol. 122, no. 5, pp. 321–326, 1995. [Cited on pages 99 and 180.]

[116] B. Genser and K. Wernecke, "Joint modelling of repeated transitions in follow-up data – a case study on breast cancer data," *Biometrical Journal*, vol. 47, no. 3, pp. 388–401, 2005. [Cited on page 100.]

[117] R. Pérez-Ocón, J. Ruiz-Castro, and M. Gámiz-Pérez, "Non-homogeneous markov models in the analysis of survival after breast cancer," *Journal of the Royal Statistical Society, Series C*, vol. 50, no. 1, pp. 111–124, 2001. [Cited on page 100.]

[118] W. Sauerbrei and P. Royston, "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials," *Journal of the Royal Statistical Society, Series A.*, vol. 161, no. 1, pp. 71–94, 1999. [Cited on page 102.]

[119] G. Soutinho and L. Meira-Machado, "The markovmsm package: Methods for checking the markov condition in multi-state survival data," 2021. [Online]. Available: https://cran.r-project.org/web/packages/markovMSM [Cited on pages 114 and 180.]

[120] G. Soutinho, M. Sestelo, and L. Meira-Machado, "survidm: An r package for inference and prediction in an illness-death model," *R Journal*, 2021. [Cited on pages 116, 141, and 150.]

[121] L. Meira-Machado and J. Roca-Pardiñas, "p3state.msm: Analyzing survival data from an illness-death model," *Journal of Statistical Software*, vol. 38:3, 2011. [Cited on page 116.]

[122] A. Araújo, j. Roca-Pardiñas, and L. Meira-Machado, "Tpmsm: Estimation of the transition probabilities in 3-state models," *Journal of Statistical Software*, vol. 62, pp. 1–29, 2014. [Cited on page 116.]

[123] A. Allignol, M. Schumacher, and J. Beyersmann, "Empirical transition matrix of multi-state models: The etm package," *Journal of Statistical Software*, vol. 38:4, pp. 1–15, 2011. [Cited on page 116.]

[124] L. de Wreede, M. Fiocco, and H. Putter, "mstate: An r package for the analysis of competing risks and multi-state models," *Journal of Statistical Software*, vol. 38:7, pp. 1–30, 2011. [Cited on page 116.]

[125] V. Balboa and J. de Uña-Álvarez, "Estimation of transition probabilities for the illness-death model: Package tp.idm," *Journal of Statistical Software*, vol. 83:10, pp. 1–19, 2018. [Cited on page 116.]

[126] C. Jackson, "Multi-state models for panel data: The msm package for r," *Journal of Statistical Software*, vol. 38:8, pp. 1–28, 2011. [Cited on page 117.]

[127] O. O. Aalen, O. Borgan, and H. Fekjaer, "Covariate adjustment of event histories estimated from markov chains: The additive approach," *Biometrics*, vol. 57, no. 4, pp. 993–1001, 2001. [Cited on page 117.]

[128] L. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in Medicine*, vol. 11, pp. 1871–1879, 1992. [Cited on page 117.]

[129] R. B. Geskus, "Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring," *Biometrics*, vol. 67, no. 1, pp. 39–49, 2011. [Cited on pages 117 and 134.]

[130] G. A. Satten and S. Datta, "Marginal estimation for multi-stage models: waiting time distributions and competing risks analyses," *Statistics in Medicine*, vol. 21, no. 1, pp. 3–19, 2002. [Cited on pages 118 and 136.]

[131] T. M. Therneau, *A Package for Survival Analysis in R*, 2021, r package version 3.2-11. [Online]. Available: https://CRAN.R-project.org/package=survival [Cited on pages 124, 141, and 146.]

[132] A. Kirk, *Data visualization: a successful design process*. Birmingham: Packt Publishing, 2012. [Cited on page 140.]

[133] N. Yau, *Data Points: Visualization that means something*. John Wiley & Sons, 2013. [Cited on page 140.]

[134] P. Govan, "eanalytics: Dynamic web-based analytics for the energy industry," *Journal of Open Research Software*, vol. 4, no. e45, 2016. [Cited on pages 140 and 142.]

[135] L. A. and G. orffy B., "Web-based survival analysis tool tailored for medical research (kmplot): Development and implementation," *J Med Internet Res*, vol. 23, no. 7, 2021. [Cited on page 140.]

[136] R. Muenchen, "The popularity of data science software." [Online]. Available: http://r4stats.com/articles/popularity/ [Cited on page 140.]

[137] J. Varma and V. Virmani, "Shiny alternative for finance in the classroom," *Institute of Management. W.P.*, 2017. [Cited on pages 140 and 142.]

[138] K. Walker, "Tools for interactive visualization of global demographic concepts in r," *Spatial Demography*, vol. 4, no. 3, pp. 207–220, 2016. [Cited on pages 140 and 142.]

[139] H. A. Wojciechowski, J. and R. Upton, "Interactive pharmacometric applications using r and the shiny package," *CPT: pharmacometrics & systems pharmacology*, vol. 4, no. 3, pp. 146–159, 2015. [Cited on pages 140, 142, and 143.]

[140] K. B. Powers, R. and W. Martinez, "Developing tools for analysis of text data," *JSM 2016 - Section on Statistical Computing*, 2016. [Cited on page 141.]

[141] M. Dunning, S. Vowler, E. Lalonde, H. Ross-Adams, P. Boutros, I. Mills, A. Lynch, and D. Lamb, "Mining human prostate cancer datasets: The "camcapp" shiny app," *EBioMedicine*, vol. 17, pp. 5–6, 2017. [Cited on page 141.]

[142] P. Murrell and S. Potter, "The gridsvg package," *The R Journal*, vol. 6, no. 1, pp. 133–143, 2014. [Cited on page 141.]

[143] H. Putter, L. de Wreede, M. Fiocco, R. Geskus, E. Bonneville, and D. Manevski, "mstate: Data preparation, estimation and prediction in multi-state models," *The R Journal*, 2020. [Cited on pages 141 and 156.]

[144] R. Peterson, "Msdshiny: The multistate simulation designer shiny application," 2019. [Online]. Available: https://github.com/petersonR/MSDshiny/ [Cited on pages 141 and 163.]

[145] S. Lacy, "msm-shiny," 2021. [Online]. Available: https://stulacy.shinyapps.io/msm-shiny/ [Cited on pages 141 and 163.]

[146] N. Skourlis, M. Crowther, T. Andersson, and P. Lambert, "Msmplus," 2021. [Online]. Available: https://nskbiostatistics.shinyapps.io/MSMplus/ [Cited on pages 142 and 163.]

[147] W. Chang, "Web application framework for r," *CRAN*, 2017. [Cited on page 142.]

[148] S. Kaushik, "Creating interactive data visualization using shiny app in r (with examples)," *Analytics Vidhya*, 2016. [Cited on page 142.]

[149] A. Seal and D. Wild, "Netpredictor: R and shiny package to perform drug-target bipartite network analysis and prediction of missing links," *Cold Spring Harbor Laboratory*, 2016. [Cited on pages 142 and 143.]

[150] RStudio, "Shinyapps.io user guide," 2021. [Online]. Available: http://docs.rstudio.com/shinyapps.io/ [Cited on page 143.]

[151] C. Beeley, *Web Application Development with R Using Shiny*. Birmingham: Packt Publishing, 2013. [Cited on page 143.]

[152] R. Miller, *Survival Analysis*. John Wiley & Sons, 1997. [Cited on page 143.]

[153] H. Putter, "Tutorial in biostatistics: Competing risks and multi-state models analyses using the mstate package," *CRAN*, 2020. [Cited on page 156.]

[154] J. Li and S. Ma, "Time-dependent roc analysis under diverse censoring patterns," *Statistics in Medicine*, vol. 30, pp. 1266–1277, 2011. [Cited on page 169.]

[155] L. Chambless and G. Diao, "Estimation of time-dependent area under the roc curve for long-term risk prediction," *Statistics in Medicine*, vol. 25, pp. 3474–3486, 2006. [Cited on page 169.]

[156] P. J. Heagerty, T. Lumley, and M. S. Pepe, “Time-dependent roc curves for censored survival data and a diagnostic marker,” *Biometrics*, vol. 6, pp. 337–344, 2000. [Cited on page 169.]

[157] Y. Zheng, T. Cai, and Z. Feng, “Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers,” *Biometrics*, vol. 62, pp. 279–287, 2006. [Cited on page 169.]