DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA
Universidade de Vigo

Semiparametric estimation in the non-Markov three-state and illness-death progressive models

Ana Paula Costa da Conceição Amorim

PhD Dissertation
Doutor Europeus

2012

Semiparametric estimation in the non-Markov three-state and illness-death progressive models

Ana Paula Costa da Conceição Amorim

To João
and
Sofia, Miguel, Filipa,
Gabriela and Daniel

Realizado el acto público de Defensa y Mantenimiento de esta Tesis Doctoral, en la modalidad de Doctorado Europeo, el dia 3 de Febrero de 2012, en la Universidad de Vigo, ante el Tribunal formado por:

Presidente: Profa. Dra. Carmen María Cadarso Suárez

Vocal 1: Prof. Dr. Per Kragh Andersen

Vocal 2: Prof. Dr. César A. Sánchez Sellero

Vocal 3: Prof. Dr. Luís F. Meira Machado

Secretario: Profa. Dra. María del Carmen Iglesias Pérez

obtuvo la máxima calificación de SOBRESALIENTE CUM LAUDE, siendo el director de la misma el Prof. Dr. Jacobo de Uña Álvarez.

Acknowledgements

Firstly I want to express my extreme gratitude to my supervisor, Professor Jacobo de Uña-Álvarez, for his encouragement, patience, guidance and support. His expertise, vision and clarity of thought greatly improved this work.

My thanks to the Department of Statistics and Operations Research of the University of Vigo for receive me and provide me good working conditions.

My thanks are also directed to Professor Luís Filipe Meira Machado from University of Minho for receiving me in his department and helping me in solving technical questions.

I would also like to acknowledge the support of colleagues in the Department of Mathematics, University of Minho and Carla Moreira during these last years.

My appreciation to Dr. Fernando Campilho from the Portuguese Oncology Institute of Porto, for providing the Leukaemia data.

I would like to acknowledge the combined financial support provided by research Centre of Mathematics of the University of Minho through the FCT (Portuguese Fundação Ciência e Tecnologia) Pluriannual Funding Program Portuguese, and by research Grants MTM2005-01274 MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación.

Contents

1 Introduction **1**

1.1 Survival Analysis 2

1.1.1 Gap times data 3

1.1.2 Ilness-death model 4

1.2 Real data 4

1.3 Outline of the thesis 5

2 A semiparametric estimator for the gap times joint distribution **7**

2.1 Introduction 8

2.2 The semiparametric estimator. Consistency 9

2.3 Simulation study 13

2.4 Real data illustration 19

2.5 Asymptotic representation of the estimator 23

2.6 Proof to the consistency result 48

3 Presmoothing the transition probabilities in the illness-death model **53**

3.1 Introduction 54

3.2 The presmoothed estimator. Consistency 55

3.3 Simulation study 60

3.4 Real data illustration 64

4 R code and further examples **71**

4.1 Introduction 72

4.2 R code for the simulations in Section 2.3 72

4.3 A simple example in the illness-death model 89

4.4 Leukaemia data 102

5 Concluding remarks and future research **109**

5.1 Concluding remarks 110

5.2 Future research 111

6 Summary in Spanish **113**

7 Bibliography **127**

Chapter 1

Introduction

Contents

1.1 Survival Analysis **2**

1.1.1 Gap times data 3

1.1.2 Ilness-death model 4

1.2 Real data **4**

1.3 Outline of the thesis **5**

1.1 Survival Analysis

Survival Analysis is concerned with inter-event times. In a classical setup, the focus is on the time elapsed between two well-defined events: the starting event (or 'birth'), and the terminating event (or 'death'). This time is therefore called the 'lifetime' or the 'survival time'. Applications of Survival Analysis include medicine, biology, economics, astronomy, and engineering, among other fields. When analyzing survival data, one must face the important problem of censoring. A censored lifetime occurs when the observation of the terminating event is not possible. This may be due to time limitations in the study or because another relevant event occurs before the terminating event of interest. In this case, the recruited inter-event time is strictly less than the time of interest, and proper corrections are needed in order to perform consistent estimation of population parameters and curves.

In this scenario, the Kaplan-Meier product-limit estimator has become the standard method to estimate the survival probability in a nonparametric way. The statistical properties of the Kaplan-Meier estimator have been thoroughly investigated; see e.g. Klein and Moeschberger (1997). Besides, this estimator has been adapted to several problems such as the estimation of smooth curves (as the density function), conditional curves (e.g. the regression function and the conditional distribution function), multivariate distributions, and regression parameters.

However, one of the major drawbacks of the Kaplan-Meier estimator is that it exhibits a large variance when the proportion of censored lifetimes is large, particularly at the right tail of the distribution. In order to reduce the variance in estimation, several alternatives to the Kaplan-Meier curve have been proposed. These alternative estimators make use of some additional information on the censoring mechanism. The most famous example is the so-called Koziol-Green estimator, see Cheng and Lin (1987), which is based on the assumption that the hazard rate pertaining to the censoring variable is proportional to the hazard of ultimate interest. This assumption is equivalent to the conditional independence between the censoring indicator and the observable lifetime, which turns out to be unrealistic in practice. Still, by assuming that the conditional probability of censoring is a smooth, maybe non-constant, function of the observable lifetime, one can construct estimators with variance smaller than that of the Kaplan-Meier. This quite less restrictive assumption was used by several authors, see e.g. Dikta (1998) and Cao et al. (2005), to introduce what we in general term as 'presmoothed estimators'.

In this context, 'presmoothing' means to replace the no-censoring indicators by some smooth fit to the conditional probability of uncensoring given the observable lifetime. This has allowed to reduce the variance associated to the Kaplan-Meier-based estimators in different problems, including nonparametric curve estimation (Cao and Jácome (2004); Cao et al. (2005)) or regression analysis (de Uña-Álvarez and Rodríguez-Campos (2004); Yuan (2005); Iglesias-Pérez and de Uña-Álvarez (2008)). When the 'presmoothing' is performed on the basis of some parametric model, one

comes up with a semiparametric censorship model and, consequently, with some semiparametric substitute for the Kaplan-Meier estimator. This approach has been investigated in much detail by Dikta (1998, 2000, 2001), see also Dikta et al. (2005). One of the main results provided in that investigation is that the semiparametric estimator has smaller variance (when compared to Kaplan-Meier), being robust to miss-specifications of the parametric model otherwise. The goal of the present work is to use these ideas in the specific context of the three-state and the illness-death progressive multi-state models. This two multi-state models are briefly discussed in the following two sections.

1.1.1 Gap times data

The statistical analysis of consecutive gap times is an issue of much importance in a number of fields, including engineering, economy, epidemiology, and survival analysis. Most of the times, one will be interested in describing not only the marginal distribution of the gap times but also the correlation structure among them. This happens, for example, when analyzing recurrent event data, which arise when each individual may go through a well-defined event several times along his history. Then, the inter-event times are referred to as the gap times, and they are of course determined by the times at which the recurrences take place (i.e. the recurrence times). See Cook and Lawless (2007) for an up-to-date revision of statistical methods for recurrent event data.

Alternatively, we may think about gap times as arising from a particular multi-state model. Multi-state models (Andersen et al. (1993); Meira-Machado et al. (2009)) are the most common models used for the description of longitudinal survival data. A multi-state model is a model for a stochastic process, which is characterized by a set of states and the possible transitions among them. The states represent different situations of the individual (healthy, diseased, etc) along a follow-up. Special multi-state models that have been widely used in biomedical applications are the three-state progressive model, the illness-death model, or the bivariate model (Hougaard (2000)). Recent reviews on multi-state models include Commenges (1999) , Hougaard (1999), Andersen and Keiding (2002), and Meira-Machado et al. (2009).

The three-state progressive model is formed by three states and two possible transitions: from state 1 to state 2, and from state 2 to state 3. Consequently, the observation of a process of such type provides information on two consecutive gap times (these are, the transition times among the three states). In practice, as in the classical Survival Analysis setup described above, complications arise in estimation due to the presence of censored information. See e.g. Lin et al. (1999) and references therein. This is the setup considered in Chapter 2.

1.1.2 Ilness-death model

The illness-death model is a generalization of the three-state progressive model in which a direct transition from state 1 to the final, absorbing state 3 is possible. This model is very important in applications. In this model one of the major goals is the estimation of the so-called transition probabilities (see Chapter 3 for a formal definition). Traditionally, this estimation is performed under a Markov assumption, which leads to the time-honored Aalen-Johansen estimator (Aalen and Johansen (1978)). However, in some applications the Markov condition is not fulfilled (e.g. Andersen et al. (2000)), and the Aalen-Johansen estimator may be inconsistent. To overcome this issue, Meira-Machado et al. (2006) introduced a substitute for the Aalen-Johansen estimator which does not depend on the Markov condition. Unfortunately, the variance of this alternative estimator may be very large in heavily censored scenarios. The possibility of improving Meira-Machado et al. (2006)'s estimator via presmoothing is explored in Chapter 3.

1.2 Real data

In this thesis some data sets will be used for illustration purposes. One of these data sets (the bladder cancer data) fits the three-state progressive model, while the colon cancer data is adapted to the illness-death model. Besides, we use several estimation methods to analyze new clinical data provided by the IPO (the Portuguese Institute of Oncology at Porto) on bone marrow transplants for acute leukaemia patients; this data set is analyzed also at the light of the illness-death model. These data sets are briefly presented now.

Bladder cancer data

The Veterans Administration Cooperative Urological Research Group was responsible by a cancer bladder study (Byar, 1980). In this study, patients had superficial bladder tumors that were remove transurethrally. Many patients had multiple recurrences of tumors, and new tumors were removed at each visit. Here we consider the 85 individuals in the placebo and thiotepa treatment groups; these data are listed in Wei et al. (1989). They are also available in the `survival` package of the `R` software (R-Development-Core-Team (2009)).

These data are used in Section 2.4 of Chapter 2 to illustrate the performance of the semiparametric estimator of the gap times joint distribution function. For this we only consider the first two recurrence times in the data set.

Colon cancer data

The colon cancer data is also available in `R`, package `survival`. These data come from a large clinical trial on Duke's stage III patients, affected by colon cancer, that underwent a curative surgery for colo-rectal cancer (Moertel et al. (1990)). In this study, from the total of 929 patients,

468 developed 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. Since recurrence can be expressed as an intermediate event, we use an illness-death model to represent these data.

In section 3.4 of Chapter 3 we use these data to illustrate the proposed semiparametric estimators of the transition probabilities.

Leukaemia data

The leukaemia data consist in all the individuals diagnosed from acute leukaemia (lymphocytic or myelocytic) between June 1989 and April 2009 at the IPO (the Portuguese Institute of Oncology at Porto). The number of individuals was 251. The standard treatment for acute leukemia is a bone marrow transplant. After the transplant, a relapse may occur. Relapse was defined on the basis of morphologic evidence of leukemia in bone marrow or other sites. In case of relapse, the patient will immediately undergo a second transplant, and so on. Here we only consider the first and the second transplant, and we investigate the times elapsed between the successive transplants and also the time to death (from any cause). This time variables are available (although maybe censored) because the data basis contains information on the date of the first bone marrow transplant, the date of the second transplant, and the date of last contact or death. As in the colon cancer data example, an illness-death model is suitable here.

This data are used in Chapter 4, where the several transition probabilities are estimated and graphically displayed.

1.3 Outline of the thesis

The thesis is organized as follows. In Chapter 2 we introduce a semiparametric estimator of the joint distribution function of a pair of possibly censored gap times. Consistency of a general functional based on this estimator is established (Section 2.2). A simulation study (Section 2.3) is performed to investigate the finite sample properties of the proposed estimator when compared to a purely nonparametric one. The simulation study includes the performance of a bootstrap estimator of the standard error. The real data illustration with the bladder cancer data example is given in Section 2.4. In Section 2.5, an asymptotic representation of the estimator as a sum of independent and identically distributed (i.i.d.) random variables is established and, as a consequence, the asymptotic normality of the estimator is obtained. The proof of the consistency result is deferred to Section 2.6.

In Chapter 3 a presmoothed estimator of the transition probabilities in the illness-death model is proposed. As in Chapter 2, the properties of the estimator are investigated both theoretically (consistency, Section 3.2) and through simulations (Section 3.3). Section 3.4 is devoted to the

illustration with the colon cancer data example.

In Chapter 4 we give some of the `R` code we have developed to implement the proposed methods. More specifically, in Section 4.2 the `R` code used for obtaining the simulation results of Section 2.3 is provided. In Section 4.3, a simple example (with a simulated data set) of the computation of the semiparametric estimators of the transition probabilities in the illness-death model is given. We also give the corresponding `R` code here. Finally, in Section 4.4 we estimate the transition probabilities for the leukaemia data, comparing several non-markov alternative estimators.

Chapter 5 contains the main conclusions of the several Chapters of the thesis (Section 5.1). We also give here some open problems which are interesting for our future research (Section 5.2).

The results in Chapter 2 (except for Section 2.5) are contained in the publication de Uña-Álvarez and Amorim (2011), while Chapter 3 is mostly reproduced in Amorim et al. (2011).

Chapter 2

A semiparametric estimator for the gap times joint distribution

Contents

2.1 Introduction 8
2.2 The semiparametric estimator. Consistency 9
2.3 Simulation study 13
2.4 Real data illustration 19
2.5 Asymptotic representation of the estimator 23
2.6 Proof to the consistency result 48

2.1 Introduction

As noted in Section 1.1, the statistical analysis of consecutive gap times is an issue of much importance in a number of fields, including engineering, economy, epidemiology, and survival analysis. Most of the times, one will be interested in describing not only the marginal distribution of the gap times but also the correlation structure among them. This happens, for example, when analyzing recurrent event data, which arise when each individual may go through a well-defined event several times along his history. Then, the inter-event times are referred to as the gap times, and they are of course determined by the times at which the recurrences take place (i.e. the recurrence times). See Cook and Lawless (2007) for an up-to-date revision of statistical methods for recurrent event data. In this Chapter, the interest is focused on a given couple of (successive) gap times. In our real data example in Section 2.4, these will be the time up to first recurrence and the time from first to second recurrence for bladder cancer patients. In order to formalize the discussion, we now introduce our notation.

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subject to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_1, T_2), and let $Y = T_1 + T_2$ be the total time. Due to censoring, rather than (T_1, T_2) we observe $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$, where $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$ and $\widetilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$, where $C_2 = (C - T_1)\, I\,(T_1 \leq C)$ is the censoring variable for the second gap time. Note that $\Delta_2 = 1$ implies $\Delta_1 = 1$. Hence, $\Delta_2 = \Delta_1\Delta_2 = I\,(Y \leq C)$ is the censoring indicator pertaining to the total time. We put $\widetilde{Y} = Y \wedge C$. Let $\left(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right)$, $1 \leq i \leq n$, be iid data with the same distribution as $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$. Since the censoring time is assumed to be independent of the process, the marginal distribution of the first gap time T_1 may be consistently estimated by the Kaplan-Meier estimator based on the $\left(\widetilde{T}_{1i}, \Delta_{1i}\right)$'s. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $\left(\widetilde{T}_{1i} + \widetilde{T}_{2i}, \Delta_{2i}\right)$'s. However, T_2 and C_2 will be in general dependent (because the expected correlation between the gap times), and hence the estimation of the marginal distribution of the second gap time is not such a simple issue. Also, it is not clear in principle how the bivariate distribution function $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$ can be efficiently estimated. This issue was investigated, among others, by Wang and Wells (1998), Lin et al. (1999), Wang and Chang (1999), Peña et al. (2001), van der Laan et al. (2002), Schaubel and Cai (2004), Van Keilegom (2004), or de Uña-Álvarez and Meira-Machado (2008).

In this Chapter we propose a semiparametric estimator for the bivariate distribution function of the gap times, $F_{12}(x, y)$. For this, we assume that the probability of censoring for T_2 given the (possibly censored) gap times belongs to a parametric family of binary regression curves. That is, letting $m(x, y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y)$, it is assumed that $m(x, y)$ follows some parametric

model. In Section 2.2 we will see that, in essence, this implies assuming a parametric (smooth) model for $m_1(x,y) = P(\Delta_2 = 1|\widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1)$. Note that, since $\widetilde{T}_1$, $\widetilde{Y}$, Δ_1, and Δ_2 are observed, this assumption is testable in practice, see e.g. Hosmer and Lemeshow (1989). On the basis of this parametric assumption, we are able to introduce a new estimator. Basically, the new method uses a presmoothed version of the Kaplan-Meier estimator (see e.g. Dikta (1998)) pertaining to the distribution of the total time (the Y) to weight the bivariate data. In the limit case of no presmoothing, the estimator we propose reduces to that in de Uña-Álvarez and Meira-Machado (2008), which was shown to have nice properties. However, the introduction of parametric presmoothing may greatly reduce the variance in the estimation, particularly at the right tail of the (bivariate) distribution or for heavy censoring on T_2. This will become clear below.

In Section 2.2 the consistency of the estimator is established. The finite sample performance of the estimator is investigated through simulations in Section 2.3. The simulation results are also used to evaluate the performance of a bootstrap standard deviation estimator. Real data illustration is provided in Section 2.4, while in Section 2.5 we derive and asymptotic representation of the estimator useful to establish a Central Limit Theorem. The proof to the consistency result is deferred to Section 2.6.

The idea of presmoothing the Kaplan-Meier estimator through a parametric model goes back to Dikta (1998), who termed this method as 'semiparametric censorship modeling'. See also Dikta (2000, 2001) and Dikta et al. (2005). Parametric presmoothing with covariates was considered by de Uña-Álvarez and Rodríguez-Campos (2004), Yuan (2005), or Iglesias-Pérez and de Uña-Álvarez (2008). All these references conclude that the presmoothed (semiparametric) estimators have improved variance when compared to purely nonparametric estimators. Here we show that presmoothing is also useful to improve efficiency in the multivariate setup of gap times.

2.2 The semiparametric estimator. Consistency

Let $\widetilde{Y}_i = \widetilde{T}_{1i} + \widetilde{T}_{2i}$ be the $i-$th recorded total time. Introduce the ordered $\widetilde{Y}-$statistics $\widetilde{Y}_{1:n} \leq \widetilde{Y}_{2:n} \leq \ldots \leq \widetilde{Y}_{n:n}$ and denote by $(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}, \Delta_{[1i:n]}, \Delta_{[2i:n]})$ the $i-$th concomitant. Let W_i be the Kaplan-Meier weight attached to $\widetilde{Y}_{i:n}$ when estimating the marginal distribution of Y from the $\left(\widetilde{Y}_i, \Delta_{2i}\right)$'s. That is,

$$W_i = \frac{\Delta_{[2i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{[2j:n]}}{n-j+1}\right].$$

Here, ties within the censored or within the uncensored times are ordered arbitrarily, and ties among the uncensored and censored times are treated as if the former precede the later. In the uncensored case we have $W_i = n^{-1}$ for each i. In de Uña-Álvarez and Meira-Machado (2008) the

following estimator was proposed:

$$\widehat{F}_{12}(x,y) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{[1i:n]} \leq x, \widetilde{T}_{[2i:n]} \leq y). \tag{2.1}$$

These authors showed that this estimator is consistent whenever $x+y$ is smaller than the upper bound of the support of the censoring time. In general, one only has (as usual)

$$\lim_{n\to\infty} \widehat{F}_{12}(x,y) = P\left(T_1 \leq x, T_2 \leq y, T_1 + T_2 \leq \tau_H\right) \equiv F_{12}^0(x,y),$$

where τ_H is the upper bound of the support of the distribution function H of $\widetilde{Y}$, assumed to be continuous throughout the Chapter. The estimator (2.1) was proved to be more efficient than previous estimators, while being more natural at the same time. Indeed, unlike other available estimators, it is an empirical distribution assigning nonnegative mass to each pair of gap times. Note that this estimator only assigns positive mass to those pairs of gap times with both components uncensored. Now we will modify this estimator in order to incorporate the semiparametric information.

Put $m(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y)$, that is, the probability of uncensoring for the total time Y given the observable information on both gap times. Note that this function is only defined for $x \leq y$; indeed, assuming $P(T_2 = 0) = 0$ (which of course holds under continuity), we have $m(x,x) = 0$, since the event $\left\{\widetilde{T}_1 = \widetilde{Y}\right\}$ corresponds exactly to $\Delta_1 = 0$ in this case, and since $\Delta_1 = 0$ implies $\Delta_2 = 0$. This shows the discontinuous nature of the function m, and consequently prevents us from using any smooth fit to this unknown curve. On the other hand, for $x < y$, we obtain $m(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1) \equiv m_1(x,y)$, since the event $\Delta_1 = 1$ is superfluous in the presence of $\widetilde{T}_1 < \widetilde{Y}$. Introduce the presmoothed Kaplan-Meier weights through

$$W_i(m) = \frac{m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n})}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n})}{n-j+1}\right],$$

that is, each censoring indicator $\Delta_{[2j:n]}, j = 1, ..., i$, in W_i is replaced by the conditional probability $m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n})$. We assume that $m(x,y) = m(x,y;\beta)$ where β is a vector of parameters and

$$m(x,y;\beta) = \begin{cases} 0 & \text{if} \quad x = y \\ m_1(x,y;\beta) & \text{if} \quad x < y, \end{cases}$$

and $m_1(.,.;\beta)$ stands for a (smooth) parametric binary regression model (e.g. logistic) for m_1. In practice, β is replaced by some consistent estimator β_n, which typically will be computed by maximizing the conditional likelihood of the Δ_2's given $\left(\widetilde{T}_1, \widetilde{T}_2\right)$, for those individuals with $\Delta_1 = 1$ (see e.g. Dikta (1998, 2000)). Thus, we introduce the parametrically presmoothed Kaplan-Meier weights as

$$W_i(\beta_n) = \frac{m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n}; \beta_n)}{n-j+1}\right],$$

where $m(x,y;\beta_n) = I(x < y)m_1(x,y;\beta_n)$ and where β_n is the maximizer of the conditional likelihood

$$L_1(\beta) = \prod_{\Delta_{1i}=1} m_1\left(\widetilde{T}_{1i},\widetilde{Y}_i;\beta\right)^{\Delta_{2i}}\left[1-m_1\left(\widetilde{T}_{1i},\widetilde{Y}_i;\beta\right)\right]^{1-\Delta_{2i}}$$

Note that this definition of $m(x,y;\beta_n)$ mimics the discontinuous behavior of the true m. On the basis of these weights, we introduce the new semiparametric estimator of $F_{12}(x,y)$ as

$$\widehat{F}_{12}^{sp}(x,y) = \sum_{i=1}^{n} W_i(\beta_n) I(\widetilde{T}_{[1i:n]} \le x, \widetilde{T}_{[2i:n]} \le y). \tag{2.2}$$

Unlike for (2.1), the estimator F_{12}^{sp} may attach positive mass to pairs of gap times with a censored T_2, while the weight attached to pairs with first gap time censored remains to be zero. As a consequence, the differences between (2.2) and (2.1) will be more evident when increasing the proportion of censoring on T_2 for the subpopulation $\Delta_1 = 1$.

More generally, we are concerned with the estimation of $S(\varphi) = E[\varphi(T_1,T_2)]$ for a given transformation φ. Specific transformations give the joint and the marginal distributions of the gap times, the moments of these variables, or the correlation coefficient. By noting $S(\varphi) = \int \varphi dF_{12}$, we introduce the following estimator of this expectation:

$$S_n(\varphi) = \int \varphi d\widehat{F}_{12}^{sp} = \sum_{i=1}^{n} W_i(\beta_n)\varphi(\widetilde{T}_{[1i:n]},\widetilde{T}_{[2i:n]}).$$

Note that this is just $\widehat{F}_{12}^{sp}(x,y)$ when we take $\varphi(u,v) = I(u \le x, v \le y)$. Next result establishes the strong consistency of $S_n(\varphi)$ under an integrability condition. We will also refer to the following assumption:

$$U: \sup_{x,y} |m_1(x,y;\beta_n) - m_1(x,y)| \to 0 \quad \text{as n} \to \infty \quad w.p.1,$$

which says that the function m_1 can be accurately approximated (in a uniform way) by some member of the parametric family $m_1(.,.;\beta)$, see Dikta (1998, 2000) for further discussion on this.

Theorem 2.2.1. *Assume $P(T_2 = 0) = 0$. Assume that H is continuous, that U hold, and that*

$$\int \frac{|\varphi(u,v)|\, F_{12}^0(du,dv)}{m_1(u,u+v)(1-H(u+v))^{\rho}} < \infty$$

is satisfied for some $\rho > 0$. Then, with probability 1

$$\int \varphi d\widehat{F}_{12}^{sp} \to \int \varphi dF_{12}^0.$$

The proof to Theorem 2.2.1 is similar to that of Theorem 2.1 in de Uña-Álvarez and Rodríguez-Campos (2004); here, the role of their covariate vector is played by the first gap time, while the total time Y is taken as the 'response'. Note that, since C is assumed to be independent of (T_1, T_2), the identifiability conditions H1 and H2 in de Uña-Álvarez and Rodríguez-Campos (2004) automatically hold. In our setup, these conditions read

H1. Y and C are independent

H2. $P(Y \leq C|T_1, Y) = P(Y \leq C|Y)$

which clearly follow from the independence between the censoring time and the gap times. However, since the first gap time is subject to right-censoring, the results in de Uña-Álvarez and Rodríguez-Campos (2004) do not directly apply here. Indeed, our presmoothing function vanishes on a line, and some care is needed in proofs to avoid zero denominators. We give an in-detail proof of Theorem 2.2.1 in Section 2.6.

Theorem 2.2.1 can be regarded as an adaptation of the Strong Law in Dikta (2000) to the context of censored gap times. Moreover, the result remains valid when using any presmoothing function $m_{1n}(x, y)$ satisfying assumption U, so it is not restricted to parametric presmoothing. We also indicate here that the integrability assumption in Theorem 2.2.1 is a consequence of estimating the binary regression $m_1(x, y)$ through $m_1(x, y; \beta_n)$; indeed, under the stronger assumption

$$U' : \sup_{x,y} \left| \frac{m_1(x, y; \beta_n)}{m_1(x, y)} - 1 \right| \to 0 \qquad w.p.1,$$

it is easily seen from the proofs in Section 2.6 that one can state Theorem 2.2.1 merely under

$$\int \frac{|\varphi(u, v)| \, F_{12}^0(du, dv)}{(1 - H(u + v))^\rho} < \infty,$$

which basically imposes the existence of the limit $\int \varphi dF_{12}^0$.

Now, an application of Theorem 2.2.1 to $\varphi(u, v) = I(u \leq x, v \leq y)$ leads to the pointwise convergence of $\widehat{F}_{12}^{sp}(x, y)$ to $F_{12}^0(x, y)$. Then, a standard uniformity argument gives the uniform consistency of the semiparametric estimator. This is stated as a Corollary.

Corollary 2.2.1. *Under the conditions of Theorem 2.2.1, with probability 1*

$$\sup_{x,y} \left| \widehat{F}_{12}^{sp}(x, y) - F_{12}^0(x, y) \right| \to 0.$$

From (2.2) we can obtain an estimator for the marginal distribution of the second gap time, $F_2(y) = P(T_2 \leq y)$, namely

$$\widehat{F}_2^{sp}(y) = \widehat{F}_{12}^{sp}(\infty, y) = \sum_{i=1}^{n} W_i(\beta_n) I(\widetilde{T}_{[2i:n]} \leq y). \tag{2.3}$$

Note that $\widehat{F}_2^{sp}(y)$ is not Dikta (1998)'s presmoothed Kaplan-Meier estimator based on the $\left(\widetilde{T}_{2i}, \Delta_{2i}\right)$'s. This is because the weights $W_i(\beta_n)$ refer to the $\widetilde{Y}$ (rather than the $\widetilde{T}_2$) ordered statistics. Indeed, since T_2 and C_2 are expected to be dependent, the ordinary Kaplan-Meier estimator of F_2 will be in general inconsistent. As for (2.2), in general we have (assuming continuity for H)

$$\lim_{n\to\infty} \widehat{F}_2^{sp}(y) = P(T_2 \leq y, T_1 + T_2 \leq \tau_H) \equiv F_2^0(y),$$

and again the restriction $T_1 + T_2 \leq \tau_H$ plays a role. Hence, it is interesting to discuss the conditions under which both estimators $\widehat{F}_{12}^{sp}(x, y)$ and $\widehat{F}_2^{sp}(y)$ converge to their respective targets.

Let F and G denote the distribution functions of Y and C, respectively. Let τ_F be the upper bound of the support of F, and similarly define τ_G. Assume again that H is continuous (see de Uña-Álvarez and Meira-Machado (2008), for a more general discussion). In essence, two different situations are possible. (A) If $\tau_F \leq \tau_G$, then we get that $\widehat{F}_{12}^{sp}(x, y)$ is consistent for any (x, y). (B) If $\tau_G < \tau_F$, then $\tau_H < \tau_F$ and consistency is only ensured for $x + y \leq \tau_H$. This is not surprising, since in this case relevant information on F is missing on the whole interval $(\tau_G, \tau_F]$. The bivariate estimators proposed in Wang and Wells (1998), Lin et al. (1999) and de Uña-Álvarez and Meira-Machado (2008) suffer from the same problem, which is related to a support restriction and cannot be solved by using any kind of presmoothing. Similar comments hold for (2.3). However, note that in this latter case, to get consistency of $\widehat{F}_2^{sp}(y)$ in situation (B) one should require $P(T_1 \leq \tau_H - y) = 1$, a condition that will typically fail for y at the right tail of F_2. Specifically, if τ_1 stands for the upper bound of the support of T_1, we have $\widehat{F}_2^{sp}(y) \to F_2(y)$ w.p.1 for $y \leq \tau_H - \tau_1$. In many applications τ_1 will be close (or even equal) to τ_H, and hence the marginal distribution of the second gap time cannot be estimated in this way. In the real medical data illustration of Section 2.4 we rather estimate the distribution of T_2 for the subpopulation undergoing the first recurrence before some (relatively small) time x, that is, $T_1 \leq x$. Clearly, this guarantees consistency at least on the interval $[0, \tau_H - x]$.

2.3 Simulation study

In this Section we investigate the performance of the proposed estimator $\widehat{F}_{12}^{sp}(x, y)$ through simulations. The simulated scenario is the same as that described in Lin et al. (1999) and de Uña-Álvarez

and Meira-Machado (2008). To be precise, the gap times (T_1, T_2) were generated according to the bivariate distribution

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

where the marginal distribution functions F_1 and F_2 are exponential with rate parameter 1. This corresponds to the so-called Farlie-Gumbel-Morgenstern copula, where the single parameter θ controls for the amount of dependency between the gap times. The parameter θ was set to 0 for simulating independent gap times, and also to 1, corresponding to 0.25 correlation between T_1 and T_2. Specifically, we performed the following steps:

(1) $V_1 \sim U(0,1), V_2 \sim U(0,1)$ are independently generated;

(2) $U_1 = V_1, A = \theta(2U_1 - 1) - 1, B = (1-\theta(2U_1-1))^2 + 4\theta V_2(2U_1-1)$

(3) $U_2 = 2V_2/\left(\sqrt{B} - A\right)$

(4) $T_1 = \ln(1/(1-U_1)), T_2 = \ln(1/(1-U_2))$

An independent uniform censoring time C was generated, according to models $U[0,4]$ and $U[0,3]$. The first model resulted in 24% of censoring on the first gap time, and in 47% of censoring on the second gap time. The second model increased these censoring levels to 32% and about 57%, respectively. Sample sizes 50, 100, 250 and 500 were considered. In each simulation, 1,000 samples were generated.

We considered as (x,y) pairs four different points, corresponding to the four different combinations of the percentiles 20% and 80% of the marginal distributions of the gap times. In this manner, we were able to explore the relative behavior of the estimator at the different corners of the joint distribution. As a measure of efficiency, we took the Mean Squared Error (MSE) of $\widehat{F}_{12}^{sp}(x,y)$ along the 1,000 trials. In the simulations, the MSE's were mainly determined by the variances, while the bias terms (squared) were of a smaller order of magnitude (results not shown). Hence, the estimator with the smallest MSE is that enjoying of minimum variance. In Tables 2.1 and 2.2 we report the MSE's attained by the proposed estimator when based on several presmoothing functions. The row labeled with m corresponds to presmoothing with the true function $m(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y)$. This is unrealistic in practice, because this function will be typically unknown, but the figures are relevant because they represent the optimal situation in which the presmoothing function is 'perfectly estimated' (so the attained MSE's are expected to be lower bounds for the error of any realistic estimator). In the simulated models the function m is given by (for $x < y$)

$$m(x,y) = \frac{1}{1+\eta(x,y)}, \qquad \text{where} \qquad \eta(x,y) = \frac{\lambda_G(y)}{\lambda_{2|1}(y-x|x)},$$

and where $\lambda_G(.)$ and $\lambda_{2|1}(.|x)$ stand for the hazard rate functions of C and T_2 given $T_1 = x$,

respectively. Note that $\lambda_G(y) = 1/(\tau_G - y)$ when $C \sim U[0, \tau_G]$ and that $\lambda_{2|1}(.|x)$ is given by

$$\lambda_{2|1}(y-x|x) = \frac{2 + 4\exp(-y) - 2\exp(-x) - 2\exp(-y+x)}{2 + 2\exp(-y) - 2\exp(-x) - \exp(-y+x)} \quad if\ \theta = 1,$$

being 1 when $\theta = 0$.

Table 2.1: $10^3 \times MSE$ of $\widehat{F}_{12}^{sp}(x,y)$ for several presmoothing functions (see text) along 1,000 simulated samples, case $\theta = 0$. From top to bottom: $(x,y) = \left(F_1^{-1}(0.2), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.8), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.2), F_2^{-1}(0.8)\right)$, and $\left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right)$.

	$C \sim U[0,4]$				$C \sim U[0,3]$			
n	50	100	250	500	50	100	250	500
$m(.;\beta)$	0.7024	0.3247	0.1244	0.0708	0.7347	0.3293	0.1309	0.0663
$m(.;\gamma)$	0.7250	0.3411	0.1352	0.0786	0.7582	0.3444	0.1380	0.0725
m	0.6749	0.3095	0.1246	0.0690	0.6495	0.2900	0.1186	0.0591
KM	0.8298	0.3987	0.1604	0.0865	0.8408	0.4094	0.1579	0.0839
$m(.;\beta)$	2.9085	1.4435	0.5471	0.2989	3.0520	1.4900	0.5476	0.2821
$m(.;\gamma)$	2.9595	1.4500	0.5526	0.3080	3.0670	1.4964	0.5507	0.2842
m	2.6497	1.2990	0.5148	0.2759	2.5405	1.2782	0.4842	0.2549
KM	3.4877	1.7482	0.6752	0.3537	3.7107	1.9175	0.7235	0.3641
$m(.;\beta)$	2.9347	1.3820	0.5378	0.2664	3.2162	1.4967	0.5657	0.2922
$m(.;\gamma)$	2.9575	1.3994	0.5486	0.2737	3.2462	1.5109	0.5742	0.2970
m	2.7510	1.2487	0.5123	0.2499	2.7115	1.2622	0.5006	0.2511
KM	3.5112	1.6836	0.6582	0.3406	3.9618	1.8774	0.7539	0.3862
$m(.;\beta)$	6.8489	3.4705	1.4054	0.6695	10.211	4.7643	2.0116	0.9723
$m(.;\gamma)$	6.9993	3.5538	1.4405	0.7007	10.447	4.9738	2.1642	1.0673
m	5.4665	2.8684	1.1615	0.5388	6.5614	3.0111	1.2640	0.5878
KM	8.3579	4.3358	1.7308	0.8117	13.083	7.1644	2.8870	1.3184

Secondly, the row labeled with $m(.;\beta)$ corresponds to a presmoothing based on a certain parametric family which contains the true m. Specifically, we consider a logistic model with a

preliminary transformation of the variables $\widetilde{T}_1 = x$ and $\widetilde{Y} = y$, as follows. When $\theta = 0$, for $x < y$ we took

$$m(x, y; \beta) = \frac{1}{1 + \exp(\beta_0 + \beta_1 \psi(x) + \beta_2 \psi(y))}$$

where $\psi(s) = \ln \lambda_G(s)$. Hence, the true m corresponds to $\beta_0 = \beta_1 = 0$, $\beta_2 = 1$ in this case. When $\theta = 1$, we just took $(x < y)$

$$m(x, y; \beta) = \frac{1}{1 + \exp(\beta_0 + \beta_1 \ln(\eta(x, y)))},$$

so again the true presmoothing function is included in the parametric family, specifically it corresponds to $\beta_0 = 0$ and $\beta_1 = 1$. In order to investigate the robustness of the proposed estimator with respect to miss-specifications of the binary regression family, we considered also presmoothing via a standard logistic model, without any preliminar transformation of the gap times. This is labeled with $m(.; \gamma)$ in Tables 2.1 and 2.2. Note that the true m does not belong to this parametric family, which is explicitly given by

$$m(x, y; \gamma) = \frac{1}{1 + \exp(\gamma_0 + \gamma_1 x + \gamma_2 y)}.$$

Finally, we also report in Tables 2.1 and 2.2 the errors pertaining to the estimator in de Uña-Álvarez and Meira-Machado (2008), which corresponds to the situation with no presmoothing at all. This is labeled in the Tables as KM. Some expected features are clearly seen in the Tables. For example, we see that the MSE goes down with an increasing sample size, while it increases at the right corners of the joint distribution, where the censoring effects are stronger. Besides, results for $C \sim U[0, 3]$ are in general worse than those for $C \sim U[0, 4]$, although this is not true for all the situations; a possible explanation is that the presmoothing induces a kind of informative censoring model, a discussion that goes back at least to Cheng and Lin (1987). On the other hand, the MSE tends to be a bit larger when introducing some correlation between the gap times (case $\theta = 1$), although some exceptions are found at the right corner of the joint distribution. More interestingly, from Tables 2.1 and 2.2 we see that the minimum MSE is attained by the estimator which makes use of the true m. Compared to the estimator without any presmoothing (KM), it is seen that the relative efficiency of this one is about 67%-75% when taking the average along the four considered (x, y) points for each simulated scenario. However, a more careful inspection of the results reveals that, in special cases, this relative efficiency is as small as 42%. As expected, these cases correspond to the right corner of the joint distribution $((x, y) = \left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right))$ and the heavily censored case. As discussed above, in practice one has to estimate the function m. In Tables 2.1 and 2.2, the best performance among the realistic versions of $\widehat{F}_{12}^{sp}(x, y)$ corresponds to the estimator based on the right parametric family of binary regression curves. The relative efficiency of KM with respect to this estimator is about 82%-85% on average, but again in some extreme situations (right corner, heavy censoring) it goes down to only 67%. Finally, we see that

Table 2.2: $10^3 \times MSE$ of $\widehat{F}_{12}^{sp}(x,y)$ for several presmoothing functions (see text) along 1,000 simulated samples, case $\theta = 1$. From top to bottom: $(x,y) = \left(F_1^{-1}(0.2), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.8), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.2), F_2^{-1}(0.8)\right)$, and $\left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right)$.

	$C \sim U[0,4]$				$C \sim U[0,3]$			
n	50	100	250	500	50	100	250	500
$m(.;\beta)$	1.2979	0.5735	0.2168	0.1173	1.0919	0.5928	0.2437	0.1153
$m(.;\gamma)$	1.2958	0.5708	0.2162	0.1174	1.1091	0.6047	0.2486	0.1180
m	1.2600	0.5572	0.2158	0.1141	1.0253	0.5841	0.2335	0.1120
KM	1.4068	0.6267	0.2408	0.1345	1.2210	0.6647	0.2776	0.1313
$m(.;\beta)$	3.0332	1.4798	0.5670	0.3137	3.0125	1.4339	0.6353	0.3242
$m(.;\gamma)$	3.2090	1.5668	0.6083	0.3311	3.2587	1.5223	0.6781	0.3655
m	2.9112	1.3844	0.5405	0.2969	2.7051	1.3348	0.5770	0.2857
KM	3.6242	1.8101	0.6789	0.3747	3.8507	1.8790	0.8021	0.4182
$m(.;\beta)$	3.0088	1.4905	0.6743	0.3225	3.3173	1.5772	0.6647	0.3621
$m(.;\gamma)$	3.0129	1.4956	0.6723	0.3233	3.3363	1.5768	0.6683	0.3621
m	2.8146	1.4273	0.6459	0.3079	3.0422	1.5135	0.6214	0.3390
KM	3.3812	1.6898	0.7565	0.3565	3.8003	1.8664	0.7748	0.4177
$m(.;\beta)$	6.6111	3.3523	1.4540	0.7006	9.2472	4.3998	1.8009	0.9804
$m(.;\gamma)$	6.8618	3.4152	1.4742	0.7402	10.233	4.8078	2.0860	1.2115
m	5.1991	2.7842	1.1823	0.5716	5.6046	2.6988	1.1484	0.6081
KM	8.0523	3.9276	1.6765	0.7967	13.055	6.8854	2.7521	1.6888

the presmoothed estimator based on the wrong parametric model $m(.;\gamma)$ is still (much) better than KM; the practical message is that it is worthwhile doing some parametric presmoothing even when we are not completely sure about the parametric family. This recommendation is reinforced by the testability of the parametric presmoothing function in practice (e.g. Hosmer and Lemeshow (1989)), since it only involves observable variables. An interesting point to discuss is that of the

relative benefits of presmoothing when increasing the sample size. The values in Tables 2.1 and 2.2

suggest that there exist a first order improvement related to presmoothing. That is, if the MSE of the KM estimator in de Uña-Álvarez and Meira-Machado (2008) is $MSE(KM) \sim c_{KM}/n$, and if the MSE pertaining to the semiparametric estimator is $MSE(SP) \sim c_{SP}/n$, then we would have $c_{SP}/c_{KM} < 1$. This is an interesting feature, since it is known that presmoothing ideas only lead to second-order improvements of the error in a number of applications (see e.g. Cao et al. (2005)). In practice, one will want to compute the standard error of the provided

Table 2.3: Mean and standard deviation of $se_m^B\left(\widehat{F}_{12}^{sp}(x,y)\right)/se_{MC}\left(\widehat{F}_{12}^{sp}(x,y)\right)$ with $B = 100$ along 500 trials of sample size $n = 100$ taken from the four different models (see text). From top to bottom: $(x,y) = \left(F_1^{-1}(0.2), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.8), F_2^{-1}(0.2)\right)$, $\left(F_1^{-1}(0.2), F_2^{-1}(0.8)\right)$, and $\left(F_1^{-1}(0.8), F_2^{-1}(0.8)\right)$.

	Model 1	Model 2	Model 3	Model 4
$Mean$	1.0049	0.9602	1.0059	0.9794
$S.D.$	0.2581	0.2623	0.2081	0.1973
$Mean$	0.9836	1.0225	0.9494	1.0451
$S.D.$	0.1379	0.1622	0.1210	0.1305
$Mean$	1.0397	1.0062	0.9980	0.9801
$S.D.$	0.1464	0.1658	0.1220	0.1295
$Mean$	1.0260	1.0741	0.9681	1.0039
$S.D.$	0.0977	0.1404	0.0934	0.1293

estimator. This can be done by using resampling methods such as the bootstrap (Efron (1981)). Akritas (1986) showed that the simple bootstrap performs consistently under random censoring and that it can be recommended for the computation of confidence limits. Hence, we propose and investigate an estimator of the standard error of $S_n(\varphi)$ based on the simple bootstrap. To formalize

things, let $\left(\widetilde{T}_{1i}^b, \widetilde{T}_{2i}^b, \Delta_{1i}^b, \Delta_{2i}^b\right)$, $i = 1, ..., n$, be the b -th bootstrap resample ($b = 1, ..., B$), which is formed by resampling (with replacement) each $\left(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right)$ with probability $1/n$. Let $S_n^b(\varphi)$ be the estimator $S_n(\varphi)$ when computed from the b-th bootstrap resample, and introduce the average value $S_n^\bullet(\varphi) = B^{-1}\sum_{b=1}^B S_n^b(\varphi)$. Then, the bootstrap estimator of the standard error

$se\left(S_n\left(\varphi\right)\right)=\sqrt{Var\left(S_n\left(\varphi\right)\right)}$ is defined as the standard deviation of the $S_n^b\left(\varphi\right)$'s, that is:

$$se^B\left(S_n\left(\varphi\right)\right)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(S_n^b\left(\varphi\right)-S_n^\bullet\left(\varphi\right)\right)^2}.$$

In order to explore the accuracy of the bootstrap standard error $se^B\left(S_n\left(\varphi\right)\right)$, we performed a new Monte Carlo experiment with 500 trials and sample size $n=100$ for the four models above (crossing the censoring distributions $C\sim U[0,4]$ and $C\sim U[0,3]$ with the two levels of dependency $\theta=0,1$). For this study, $B=100$ bootstrap resamples were taken. As in Tables 2.1 and 2.2, we considered $\varphi\left(u,v\right)=I(u\leq x,v\leq y)$ (so $S_n\left(\varphi\right)$ reduces to $\widehat{F}_{12}^{sp}(x,y)$) and four different points (x,y) combining the 20 and 80% percentiles of the marginal distributions of the gap times. In each case, the target $se\left(S_n\left(\varphi\right)\right)$ was approximated by the standard deviation of $S_n\left(\varphi\right)$ along the 500 Monte Carlo trials, $se_{MC}\left(S_n\left(\varphi\right)\right)$ say. Table 2.3 reports the mean and standard deviations of the quotients

$$\frac{se_m^B\left(S_n\left(\varphi\right)\right)}{se_{MC}\left(S_n\left(\varphi\right)\right)},\qquad m=1,...,500,$$

for the four distinct models, namely: $C\sim U[0,4]$ and $\theta=0$ (Model 1), $C\sim U[0,3]$ and $\theta=0$ (Model 2), $C\sim U[0,4]$ and $\theta=1$ (Model 3), and $C\sim U[0,3]$ and $\theta=1$ (Model 4). In this Table 2.3 it is seen that the bootstrap standard error is almost perfectly unbiased in all the considered situations.

2.4 Real data illustration

In this Section we consider data from a cancer bladder study (Byar (1980)) conducted by the Veterans Administration Cooperative Urological Research Group. In this study, patients had superficial bladder tumors that were remove transurethrally. Many patients had multiple recurrences of tumors during the study, and new tumors were removed at each visit. Here we analyze for illustration purposes the n =85 individuals in the placebo and thiotepa treatment groups; these data are listed in Wei et al. (1989). Only the first two recurrence times T_1 and Y (or the corresponding gap times T_1 and $T_2=Y-T_1$) are considered. Among the 85 patients, 47 relapsed at least once (45% of censoring on T_1) and, among these, 29 had another recurrence (38% of extra censoring). The presence of a reasonable amount of censored Y's among the uncensored T_1's suggests that presmoothing could lead to an important reduction of variance in estimation. We will quantify this below. In Figure 2.1 we represent the 85 observed values for the recurrence times $\left(\widetilde{T}_1,\widetilde{Y}\right)$ (months). Cases with times censored are located on the line $y=x$. On the other hand, 18 points among those out of this line (labelled with a cross) correspond to observations with second gap time censored. From this Figure it is not clear in principle which type of correlation (if any) exists between both gap times T_1 and T_2. Figure 2.2 depicts the survival curves corresponding to T_1

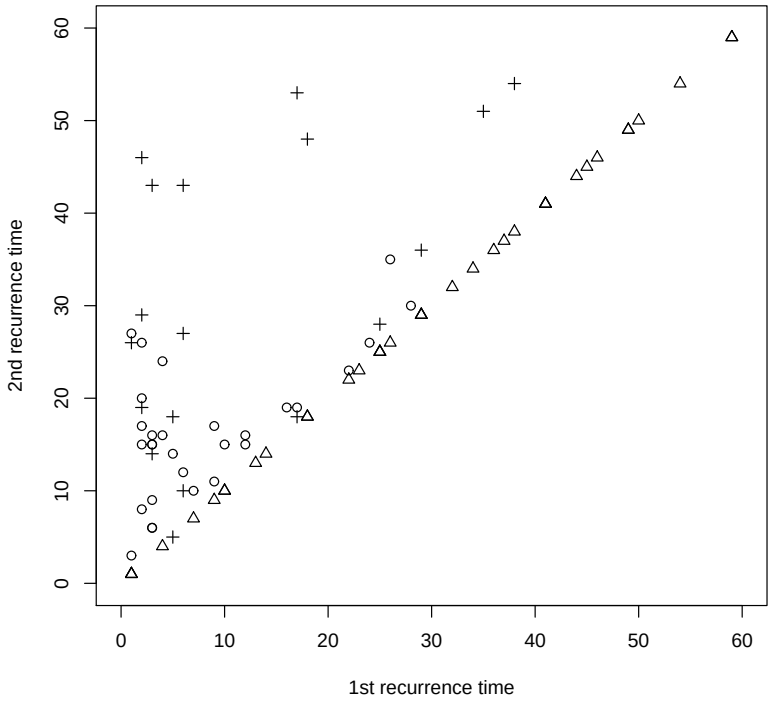

Figure 2.1: Time to first recurrence vs. time to second recurrence for the 85 cases of bladder cancer. Triangles indicate censoring in both times, while crosses indicate censoring on the second gap time.

(solid line) and Y (dashed line). It is clearly seen that the first recurrence is almost restricted to the first 3 years after randomization, while a large proportion of patients (about 60%) do not relapse in 5 years. In order to compute the semiparametric estimator (2.2), we have fitted a logistic model to the binary regression $m_1(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1)$. The results indicate that $\widetilde{Y}$ is highly significative (p=0.002590) while $\widetilde{T}_1$ does not reach significance (p=0.339851). Specifically, the fitted logistic model was

$$\widehat{m}_1(x,y) = \frac{1}{1+\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 x + \widehat{\gamma}_2 y)}$$

where $\widehat{\gamma}_0 = 2.97921$, $\widehat{\gamma}_1 = 0.04193$, and $\widehat{\gamma}_2 = -0.12817$. The coefficient of $\widetilde{Y}$ in the model is negative, thus censoring probability increases with the observed time up to second recurrence. With this parametric presmoothing we computed the estimator $\widehat{F}_{12}^{sp}(x,y)$ for $x = 5, 10, 15, 20, 30$ months and $y = 5, 10, 20$ months. Results are displayed in Table 2.4, top. For comparison, we also report in this Table 2.4 (bottom) the values of the estimator corresponding to no presmoothing,

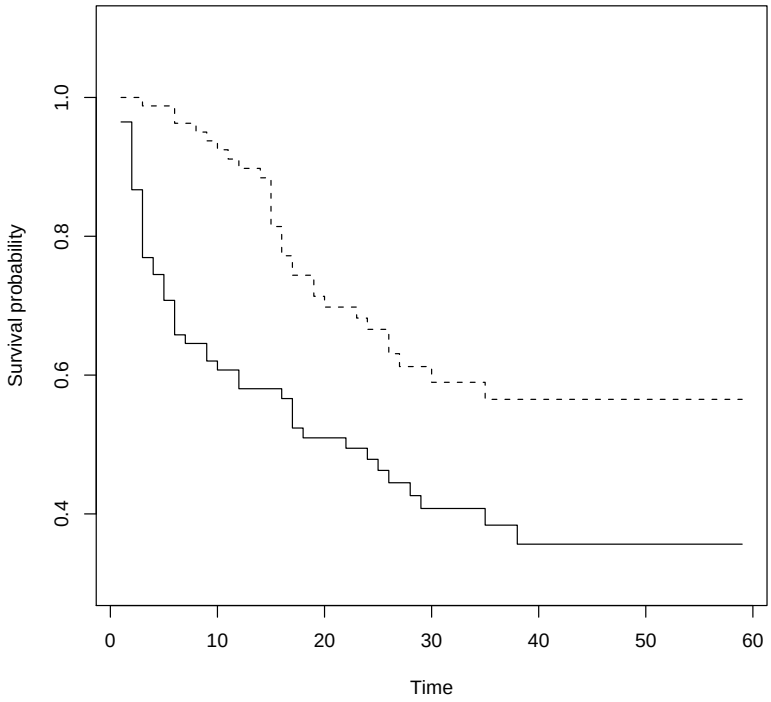

Figure 2.2: Kaplan-Meier curves for the bladder cancer data: time to first recurrence (solid line) and time to second recurrence (dashed line).

$\widehat{F}_{12}(x,y)$. From this Table we see that both methods provide similar point estimates. We estimate the standard errors for both estimators through the bootstrap method described in Section 2.3. The results in Table 2.4 (based on 5000 bootstrap resamples) reveal that: (a) the errors increase at the right corner of the joint distribution of the gap times, where the censoring effects are stronger; and (b) the semiparametric estimator has smaller standard errors, with a minimum relative efficiency of $\widehat{F}_{12}(x,y)$ of about 86% (91% when averaging the 15 cases of (x,y)). This latter point may be very important when making inferences such as e.g. group comparisons. Note also that the semiparametric estimator introduces some smoothing so one can get more reasonable plots when there are few uncensored data in the sample.

In Figure 2.3 we report the semiparametric estimator of the distribution function of T_2 for the individuals with a recurrence during the first $x = 30$ months of follow-up. Note that this conditional distribution is

$$F_{2|1}(y|x) = P\left(T_2 \leq y | T_1 \leq x\right) = \frac{F_{12}(x,y)}{F_1(x)},$$

Table 2.4: Top: Semiparametric estimator of the joint distribution function of the gap times $F_{12}(x,y)$ for the bladder cancer data (standard errors between brackets). Bottom: Same information for the estimator without presmoothing.

$\widehat{F}_{12}^{sp}(x,y)$	$y=5$	$y=10$	$y=20$
$x=5$	.0454 (.0216)	.0783(.0283)	.1896 (.0433)
$x=10$	.0906 (.0294)	.1455 (.0377)	.2568 (.0488)
$x=15$	.1133 (.0335)	.1683 (.0412)	.2796 (.0514)
$x=20$	.1482 (.0374)	.2031 (.0440)	.3144 (.0528)
$x=30$	.1965 (.0462)	.2715 (.0554)	.3828 (.0604)
$\widehat{F}_{12}(x,y)$	$y=5$	$y=10$	$y=20$
$x=5$	.0372 (.0210)	.0761 (.0298)	.1921 (.0462)
$x=10$	.0775 (.0303)	.1439 (.0401)	.2598 (.0513)
$x=15$	.1056 (.0354)	.1719 (.0436)	.2879 (.0534)
$x=20$	.1359 (.0402)	.2023 (.0469)	.3183 (.0551)
$x=30$	.1920 (.0488)	.2829 (.0574)	.3989 (.0624)

where $F_1(x) = P(T_1 \leq x)$, which can be estimated by plugging-in $\widehat{F}_{12}^{sp}(x,y)$ in the numerator and the (ordinary) Kaplan-Meier for the first gap time in the denominator. We also report in this Figure 2.3 the estimator constructed with $\widehat{F}_{12}(x,y)$. The main difference between both curves is that the semiparametric estimator has more jump points, explicitly the censored values of T_2 for which condition $T_1 \leq 30, \Delta_1 = 1$ is satisfied. This implies that the mass is more distributed, being the reason behind the variance reduction which is achieved by presmoothing. The vertical line at $y = 29$ in Figure 2.3 indicates that, according to our remarks to Theorem 2.2.1, both estimators should only be interpreted as empirical versions of $F_{2|1}^{\tau_H}(y|x) = P\left(T_2 \leq y, Y \leq \tau_H | T_1 \leq x\right)$ from that point on. Note that $\tau_H = 59$ in our application and hence $Y \leq \tau_H$ is not superfluous when $x = 30$ and $y > 29$. Finally, we give in Figures 2.4 and 2.5 two other plots which depict the joint behavior of both gap times. In Figure 2.4, two estimated distribution functions of T_2 based on the semiparametric estimator are plotted. The solid line corresponds to the subgroup $T_1 \leq 10$ months, while the dashed line refers to the subpopulation $10 < T_1 \leq 30$. This Figure suggests a negative correlation between both gap times. Figure 2.5 depicts the surface $\widehat{F}_{12}^{sp}(.,.)$, and again suggests that large times to first recurrence are connected with relatively small values of T_2.

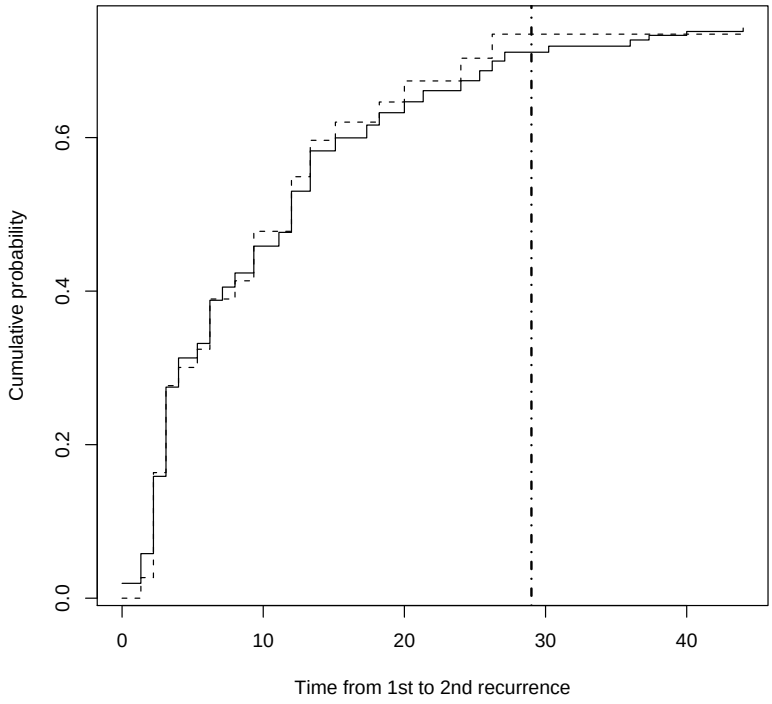

Figure 2.3: Semiparametric estimator (solid line) and no-presmoothed estimator (dashed line) of the distribution of time from first to second recurrence, for the subgroup with a recurrence in the first 30 months after randomization.

2.5 Asymptotic representation of the estimator

In this section we establish an asymptotic representation of $S_n(\varphi)$ as a sum of i.i.d. random variables. The result is similar to those obtained in Stute (1995) and Dikta et al. (2005) for Kaplan-Meier integrals and presmoothed Kaplan-Meier integrals respectively.

We use the same notation of Section 2.1, T_1 is the time up the first recurrence, Y is the time to the second recurrence, $(T_1, Y - T_1) = (T_1, T_2)$ is a pair of gap times of successive events, C is a right-censored variable, assumed to be independent of (T_1, T_2). The observable variables are $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$, $\widetilde{Y} = Y \wedge C$, $\Delta_2 = I(Y \leq C)$. The data are $\left(\widetilde{T}_{1i}, \widetilde{Y}_i, \Delta_{1i}, \Delta_{2i}\right)$, $1 \leq i \leq n$. We introduce the ordered $\widetilde{Y}$-statistics

$$\widetilde{Y}_{1:n} \leq \widetilde{Y}_{2:n} \leq \ldots \leq \widetilde{Y}_{n:n},$$

and we denote by $(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}, \Delta_{[1i:n]}, \Delta_{[2i:n]})$ the i-th concomitant (i.e., the $(\widetilde{T}_{1j}, \widetilde{T}_{2j}, \Delta_{1j}, \Delta_{2j})$

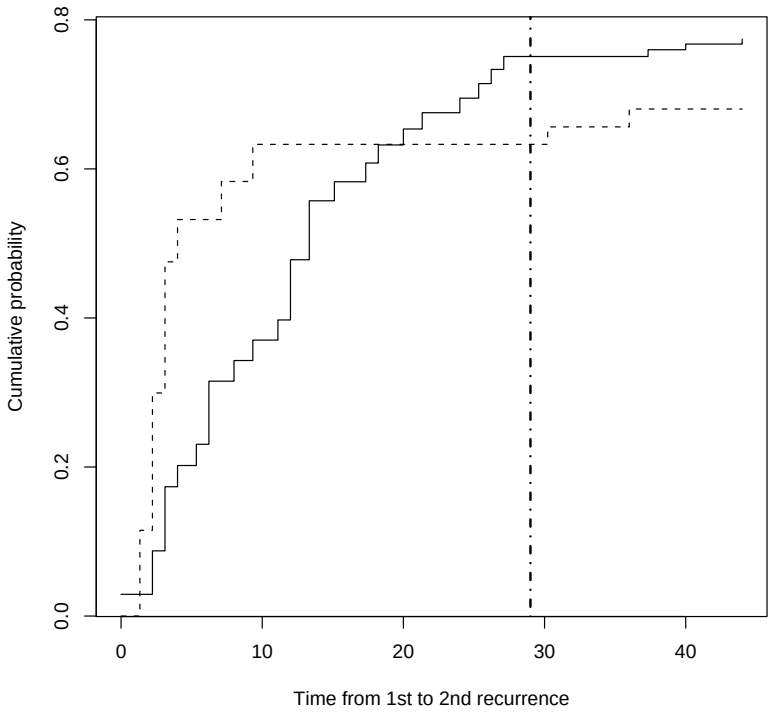

Figure 2.4: Semiparametric estimator of the distribution of time from first to second recurrence: relapse in the first 10 months (solid line) and relapse between month 10 and 30 (dashed line). Negative correlation between both gap times suggested.

pertaining to $\widetilde{Y}_j = \widetilde{Y}_{i:n}$).

Thus, the parametric presmoothed Kaplan-Meier weights are

$$W_i(\beta_n) = \frac{m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n}; \beta_n)}{n-j+1} \right],$$

where $m(x, y; \beta_n) = I(x < y) m_1(x, y; \beta_n)$ and where β_n is the maximizer of the conditional likelihood $L_1(\beta)$ introduced in Section 2.2

On the basis of these weights, the semiparametric estimator of $F_{12}(x, y)$ is

$$\widehat{F}_{12}^{sp}(x, y) = \sum_{i=1}^{n} W_i(\beta_n) I(\widetilde{T}_{[1i:n]} \leq x, \widetilde{T}_{[2i:n]} \leq y).$$

As discussed, we may be interested in the estimation of $S(\varphi) = E[\varphi(T_1, T_2)]$ for a given transfor-

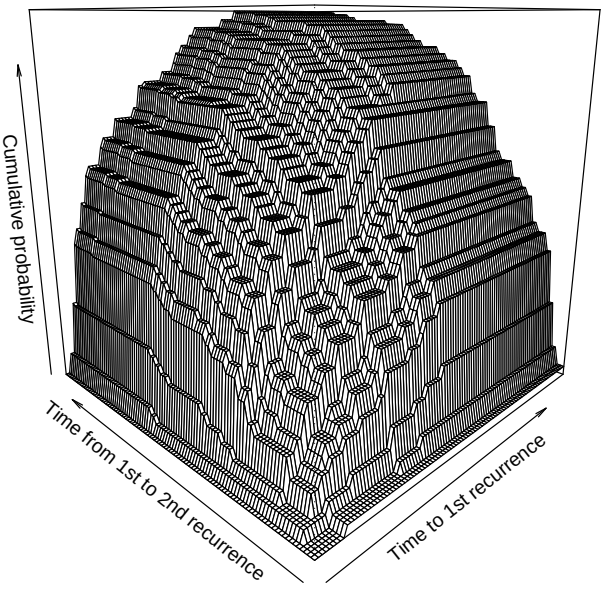

Figure 2.5: Cumulative joint distribution of the two gap times, based on the semiparametric estimator.

mation φ. Recall the estimator of $S(\varphi) = \int \varphi dF_{12}$, which is given by

$$S_n(\varphi) = \int \varphi d\widehat{F}_{12}^{sp} = \sum_{i=1}^{n} W_i(\beta_n)\varphi(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}).$$

$S_n(\varphi)$ reduces to $\widehat{F}_{12}^{sp}(x,y)$ for the special function $\varphi(u,v) = I(u \leq x, v \leq y)$.

Throughout this section we will use the following notation (note the change of notation in the distribution function of $\widetilde{Y}$ with respect to Section 2.2):

- $F(y) = P(Y \leq y)$;
- $G(x) = P(C \leq x)$;
- $\widetilde{H}(y) = P(\widetilde{Y} \leq y)$;
- $\widetilde{H}_n(y) = \dfrac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i \leq y)$;

- $H_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_{1i} \leq x, \widetilde{Y}_i \leq y)$;
- $H(x,y) = P(\widetilde{T}_1 \leq x, \widetilde{Y} \leq y)$;
- $F_{12}(x,y) = P(T_1 \leq x, T_2 \leq y)$;
- $F_{12}^0(x,y) = P(T_1 \leq x, T_2 \leq y, T_1 + T_2 \leq \tau_{\widetilde{H}})$, where $\tau_{\widetilde{H}} = \inf\{x : \widetilde{H}(x) = 1\}$;
- $H^1(x,y) = P(\widetilde{T}_1 \leq x, \widetilde{Y} \leq y, \Delta_2 = 1)$;
- $H^0(x,y) = P(\widetilde{T}_1 \leq x, \widetilde{Y} \leq y, \Delta_2 = 0)$;
- $H_n^1(x,y) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_{1i} \leq x, \widetilde{Y}_i \leq y, \Delta_{2i} = 1)$;
- $H_n^0(x,y) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_{1i} \leq x, \widetilde{Y}_i \leq y, \Delta_{2i} = 0)$.

We will refer to the following regularity conditions.

C 1. *There exists a measurable solution $\beta_n \in \mathbb{B} \subset \mathbb{R}^k$ of the equation $Grad(ln(L_1(\beta)) = 0$ which tends to β_0 in probability. The β_0 is the "true" parameter, of dimension k, and $Grad(ln(L_1(\beta)) = (D_1 lnL_1(\beta), \cdots, D_k lnL_1(\beta))$ where*

$$D_j lnL_1(\beta) = \frac{\partial}{\partial \beta_j} lnL_1(\beta), \quad j = 1, \ldots, k.$$

C 2. *Let for $0 \leq x < y$*

$$w_1(x,y;\beta) = ln(m_1(x,y;\beta)),$$

$$w_2(x,y;\beta) = ln(1 - m_1(x,y;\beta))$$

For $i = 1, 2$, $w_i(x,y;\beta)$ possesses continuous partial derivatives of second order with respect to β at each $\beta \in \mathbb{B}$ and there exists neighborhood $V(\beta_0) \subset \mathbb{B}$ of β_0 and a measurable function M such that for all $\beta \in V(\beta_0)$ with $0 \leq x < y$, and $1 \leq i, j \leq k$

$$\mid D_{i,j} w_1(x,y;\beta) \mid + |D_{i,j} w_2(x,y;\beta)| \leq M(x,y)$$

and $E(M(\widetilde{T}_1, \widetilde{Y})\Delta_1) < \infty$, where

$$D_{i,j} w_r(x,y;\beta) = \frac{\partial^2}{\partial \beta_i \partial \beta_j} w_r(x,y;\beta), \quad r = 1, 2.$$

C 3. *For* $1 \leq j \leq k$

$$[D_j m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0)\Delta_1 / m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0)]^2$$

and

$$[D_j m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0)\Delta_1 / (1 - m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0))]^2$$

have finite expectation.

C 4. *The matrix* $I(\beta_0) = (\sigma_{i,j})_{1 \leq i,j \leq k}$*, where for* $0 \leq x < y$

$$\begin{aligned} w(\Delta_2, x, y; \beta) &= \Delta_2 w_1(x, y; \beta) + (1 - \Delta_2) w_2(x, y; \beta) \\ &= \Delta_2 ln(m_1(x, y; \beta)) + (1 - \Delta_2) ln(1 - m_1(x, y; \beta)) \end{aligned}$$

and

$$\begin{aligned} \sigma_{i,j} &= -E(D_{i,j} w(\Delta_2, \widetilde{T}_1, \widetilde{Y}; \beta_0)\Delta_1) \\ &= E\left(\frac{D_i(m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0)) D_j(m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0))\Delta_1}{m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0)(1 - m_1(\widetilde{T}_1, \widetilde{Y}; \beta_0))}\right), \end{aligned} \tag{2.4}$$

is positive definite.

C 5. *There exists a neighborhood* $V(\beta_0) \subset \mathbb{B}$ *of* β_0 *such that* $m_1(x, y; \beta)$ *possesses continuous partial derivatives of second order with respect to* β *at each* $\beta \in V(\beta_0)$ *and* $0 \leq x < y$*. Furthermore,*

$$\sup_{\beta \in V(\beta_0)} \sup_{0 \leq x < y \leq T} \sum_{1 \leq i,j \leq k} |D_{i,j} m_1(x, y; \beta)| < \infty$$

and

$$\sup_{0 \leq x < y \leq T} \|Grad(m_1(x, y; \beta_0))\| < \infty,$$

where T *is a constant such that* $\widetilde{H}(T) < 1$*.*

C 6. *For* $1 \leq i \leq k$*,* $D_i m_1(x, y; \beta_0)$ *is Lipschitz on* $0 \leq x < y \leq T$ *for all* $T < \tau_{\widetilde{H}}$*.*

As noted by Dikta (1998), conditions (C1)-(C4) are needed for the asymptotic normality of β_n. Condition (C5) ensures that $m_1(x, y; \beta_n)$ is close enough to $m_1(x, y; \beta_0)$ in a uniform sense. Finally, (C6) was used in Dikta (1998) to ensure the tightness of certain processes.

The asymptotic representation of $\int \varphi d\widehat{F}_{12}^{sp}$ as a sum of i.i.d. variables will include the following functions:

$$\begin{aligned}
\gamma_0(v) &= \exp\left(\int_0^\infty \int_0^v \frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)} H(dx,dy)\right),\\
\gamma_1(v) &= \frac{1}{1-\widetilde{H}(v)} \int\int \xi^\varphi(x,y)\gamma_0(y) I(v<y) H^1(dx,dy),\\
\gamma_2(v) &= \int\int\int\int \frac{I(v>t, y>t)\xi^\varphi(x,y)\gamma_0(y)}{[1-\widetilde{H}(t)]^2} H^0(dr,dt)H^1(dx,dy),\\
\gamma_3(r,s) &= \int\int\int\int \frac{I(y>v)\alpha(u,v,r,s)\xi^\varphi(x,y)\gamma_0(y)}{1-\widetilde{H}(v)} H^1(dx,dy)H(du,dv),\\
\gamma_4(u,v) &= \int\int \xi^\varphi(x,y)\gamma_0(y)\alpha(x,y,u,v)H(dx,dy),
\end{aligned}$$

where $\xi^\varphi(x,y) = \varphi(x, y-x)$ and

$$\alpha(x,y,r,t) =< Grad(m(x,y;\beta_0))|I^{-1}(\beta_0)Grad(m(r,t;\beta_0)) >$$

where the notation $<\cdot|\cdot>$ represents the inner product in $\mathbb{R}^k$. We will also need the function

$$K(x,y,d) = \frac{d-m(x,y;\beta_0)}{m(x,y;\beta_0)(1-m(x,y;\beta_0))}$$

for $0 \le x < y$, $d = 0, 1$, with the convention $K(x,x,d) = 0$.

We will refer to the following conditions on $\xi^\varphi(x,y)$ too:

M 1. $\xi^\varphi(x,y) = 0$, for all $y > T$ where $T < \tau_{\widetilde{H}}$.

M 2. $E\left(\xi^\varphi(T_1,Y)^2\right) < \infty$

The following Theorem is the main result in this section. It establishes a representation of our estimator as a sum of i.i.d random variables. Then, the asymptotic normality of the estimator will follow from the Central Limit Theorem (CLT). Conditions (M1)-(M2) above guarantee the existence of the second moment of the leading term in the representation. These conditions (M1)-(M2) are stronger than the moment conditions on φ in the CLT of Dikta et al. (2005). Still, as it is easily seen, they are empty conditions when the interest is in the asymptotic normality of $\widehat{F}_{12}^{sp}(x,y)$ for $x+y < \tau_{\widetilde{H}}$ (just take $T = x+y$ in (M1)). A CLT for a wider class of functions φ requires a deeper investigation.

Theorem 2.5.1. *Assume that the parameter space $\mathbb{B}$ is a connected open subset of $\mathbb{R}^k$. If $\widetilde{H}$ is continuous and (C1)-(C6) and (M1)-(M2) are satisfied, then*

$$\begin{aligned} S_n(\varphi) &= n^{-1}\sum_{i=1}^{n}\xi^{\varphi}(\widetilde{T}_{1i},\widetilde{Y}_i)\gamma_0(\widetilde{Y}_i) + n^{-1}\sum_{i=1}^{n}(1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))\gamma_1(\widetilde{Y}_i) \\ &-n^{-1}\sum_{i=1}^{n}\gamma_2(\widetilde{Y}_i) - n^{-1}\sum_{i=1}^{n}K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\gamma_3(\widetilde{T}_{1i},\widetilde{Y}_i) \qquad (2.5) \\ &+n^{-1}\sum_{i=1}^{n}K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\gamma_4(\widetilde{T}_{1i},\widetilde{Y}_i) + o_p(n^{-1/2}). \end{aligned}$$

It is easily seen that $E[(1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))\gamma_1(\widetilde{Y}_i)] = E[\gamma_2(\widetilde{Y}_i)]$ and that $E[K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})(\gamma_3(\widetilde{T}_{1i},\widetilde{Y}_i)-\gamma_4(\widetilde{T}_{1i},\widetilde{Y}_i))] = 0$. Hence we obtain the following corollary.

Corollary 2.5.1. *Under the assumptions of Theorem 2.5.1*

$$\sqrt{n}\Big(\int\varphi d\widehat{F}_{12}^{sp} - \int\varphi dF_{12}^{0}\Big) \quad \xrightarrow{\mathfrak{L}} \quad \mathcal{N}(0,\sigma^2(\varphi))$$

where

$$\begin{aligned} \sigma^2(\varphi) &= Var(\xi^{\varphi}(\widetilde{T}_1,\widetilde{Y})\gamma_0(\widetilde{Y})m(\widetilde{T}_1,\widetilde{Y};\beta_0) + (1-m(\widetilde{T}_1,\widetilde{Y};\beta_0))\gamma_1(\widetilde{Y}) \\ &-\gamma_2(\widetilde{Y}) - K(\widetilde{T}_1,\widetilde{Y},\Delta_2)(\gamma_3(\widetilde{T}_1,\widetilde{Y}) - \gamma_4(\widetilde{T}_1,\widetilde{Y}))). \quad \square \end{aligned}$$

Corollary 2.5.2. *Assume that the parameter space $\mathbb{B}$ is a connected open subset of $\mathbb{R}^k$. If $\widetilde{H}$ is continuous and (C1)-(C6) are satisfied, then for $x+y<\tau_{\widetilde{H}}$*

$$\sqrt{n}(\widehat{F}_{12}^{sp}(x,y) - F_{12}(x,y)) \quad \xrightarrow{\mathfrak{L}} \quad \mathcal{N}(0,\sigma^2(x,y))$$

where $\sigma^2(x,y) = \sigma^2(\varphi_{x,y})$ and $\varphi_{x,y}(u,v) = I(u\le x, v\le y)$. $\square$

The following lemmas will be needed to prove the Theorem 2.5.1.

Lemma 2.5.1. *If assumptions (C1)-(C5) are satisfied, then $\sqrt{n}\,(\beta_n-\beta_0)$ is asymptotically normal, $\mathcal{N}(\underline{0}, I^{-1}(\beta_0))$, where $I(\beta_0)$ is the Fisher Information matrix in (C4). Furthermore,*

$$\sup_{0\le x\le y\le\tau_{\widetilde{H}}}|m(x,y;\beta_n) - m(x,y;\beta_0)| = O_p(n^{-1/2}).$$

Proof to Lemma 2.5.1

Under $(C1)$-$(C4)$, the asymptotic normality of $\sqrt{n}(\beta_n-\beta_0)$ can established as in Theorem 2.3 in Dikta (1998). Now, since $m(x,x;\beta_0)=0$, it is enough to prove

$$\sup_{0\leq x<y\leq\tau_{\widetilde{H}}}|m_1(x,y;\beta_n)-m_1(x,y;\beta_0)|=O_p(n^{-1/2}).$$

Taylor's expansion of $m_1(x,y;\beta_n)$ w.r.t. β_0 yields

$$\begin{aligned}n^{1/2}(m_1(x,y;\beta_n)-m_1(x,y;\beta_0)) &= Grad(m_1(x,y;\beta_0))n^{1/2}(\beta_n-\beta_0)\\ &+\frac{n^{1/2}}{2}\sum_{1\leq i,j\leq k}D_{i,j}(m_1(x,y;\beta^*))(\beta_{ni}-\beta_{0i})(\beta_{nj}-\beta_{0j}),\end{aligned}$$

where $\beta^*\in\mathbb{B}$ is inside the line segment connecting β_n and β_0. Given $(C1)$, $(C5)$ and the asymptotic normality of $n^{1/2}(\beta_n-\beta_0)$, we infer

$$\sup_{0\leq x<y<\infty}|m_1(x,y;\beta_n)-m_1(x,y;\beta_0)|\leq O_P(n^{-1/2})$$

which proves the lemma. □

The next lemma gives a basic representation of

$$\begin{aligned}S_n(\varphi)=\int\varphi d\widehat{F}_{12}^{sp} &= \textstyle\sum_{i=1}^n W_i(\beta_n)\varphi(\widetilde{T}_{[1i:n]},\widetilde{T}_{[2i:n]})\\ &= \textstyle\sum_{i=1}^n W_i(\beta_n)\xi^{\varphi}(\widetilde{T}_{[1i:n]},\widetilde{Y}_{i:n}).\end{aligned}\tag{2.6}$$

Lemma 2.5.2. *For a continuous $\widetilde{H}$ we have*

$$S_n(\varphi)=\int\int\xi^{\varphi}(u,v)m(u,v;\beta_n)\exp\left(n\int_0^{\infty}\int_0^{v^-}\ln\left[1+\frac{1-m(x,y;\beta_n)}{n(1-\widetilde{H}_n(y))}\right]H_n(dx,dy)\right)H_n(du,dv).$$

Proof to Lemma 2.5.2

According to (2.6) we get

$$
\begin{aligned}
S_n(\varphi) &= \sum_{i=1}^{n} \frac{\xi^{\varphi}(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n}; \beta_n)}{n-j+1} \right] \\
&= \sum_{i=1}^{n} \frac{\xi^{\varphi}(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n)}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right) \left[1 + \frac{m(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n}; \beta_n)}{n-j+1} \right] \\
&= n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) \prod_{j=1}^{i-1} \left(1 + \frac{1 - m(\widetilde{T}_{1j}, \widetilde{Y}_j; \beta_n)}{n(1 - \widetilde{H}_n(\widetilde{Y}_j)} \right)^{1_{\{\widetilde{Y}_j < \widetilde{Y}_i\}}} \\
&= n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) \times \\
&\times \exp \left(n \left[n^{-1} \sum_{j=1}^{n} 1_{\{\widetilde{Y}_j < \widetilde{Y}_i\}} ln \left(1 + \frac{1 - m(\widetilde{T}_{1j}, \widetilde{Y}_j; \beta_n)}{n(1 - \widetilde{H}_n(\widetilde{Y}_j))} \right) \right] \right) \\
&= \int \int \xi^{\varphi}(u,v) m(u,v;\beta_n) \exp \left(n \int_0^{\infty} \int_0^{v^-} \ln \left(1 + \frac{1 - m(x,y;\beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx,dy) \right) \times \\
&\times H_n(du,dv). \quad \square
\end{aligned}
$$

Now, by using $ln(1+x) \approx x$, x small, expand the exponential term of Lemma 2.5.2, at the points

$$A_i := \int_0^{\infty} \int_0^{\widetilde{Y}_i} \frac{1 - m(x,y;\beta_0)}{1 - \widetilde{H}(y)} H(dx,dy)$$

for $i = 1, \cdots, n$, to get

$$
\begin{aligned}
\exp(\cdots) &= \exp(A_i) + \frac{\exp(A_i)}{1!} \times \\
&\times \left[n \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \ln \left(1 + \frac{1 - m(x,y;\beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx,dy) - A_i \right] \\
&+ \frac{\exp(\xi_i)}{2!} \left[n \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \ln \left(1 + \frac{1 - m(x,y;\beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx,dy) - A_i \right]^2 .
\end{aligned}
$$

Here ξ_i is between the two terms in brackets. The above expression together with Lemma 2.5.2

results in

$$
\begin{aligned}
S_n(\varphi) &= \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) \exp(\cdots) \\
&= n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) \times \\
&\times \left(m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) - m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0) + m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0) \right) \times \\
&\times \left\{ \exp(A_i) \left[1 + n \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \ln \left(1 + \frac{1 - m(x, y; \beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx, dy) - A_i \right] \right. \\
&\left. + \frac{\exp(\xi_i)}{2!} \left[n \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \ln \left(1 + \frac{1 - m(x, y; \beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx, dy) - A_i \right]^2 \right\} \\
&= n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0) \gamma_0(\widetilde{Y}_i) \left(1 + B_{in}(\beta_n) + C_{in}(\beta_n) \right) \\
&+ n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) (m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) - m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0)) \gamma_0(\widetilde{Y}_i) \left(1 + B_{in}(\beta_n) + C_{in}(\beta_n) \right) \\
&+ \frac{n^{-1}}{2} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) \exp(\xi_i) \left(B_{in}(\beta_n) + C_{in}(\beta_n) \right)^2,
\end{aligned}
\tag{2.7}
$$

where

$$
C_{in}(\beta) = \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \frac{1 - m(x, y; \beta)}{1 - \widetilde{H}_n(y)} H_n(dx, dy) - \int_0^{\infty} \int_0^{\widetilde{Y}_i} \frac{1 - m(x, y; \beta_0)}{1 - \widetilde{H}(y)} H(dx, dy)
$$

and

$$
B_{in}(\beta) = n \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \ln \left(1 + \frac{1 - m(x, y; \beta_n)}{n(1 - \widetilde{H}_n(y))} \right) H_n(dx, dy) - \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \frac{1 - m(x, y; \beta)}{1 - \widetilde{H}_n(y)} H_n(dx, dy).
$$

Under continuity,

$$\ln \gamma_0(v) = \int_0^\infty \int_0^v \frac{1 - m(x, y; \beta_0)}{1 - \widetilde{H}(y)} H(dx, dy) = \int_0^\infty \int_0^v \frac{H^0(dx, dy)}{(1 - F(y))(1 - G(y))} =$$

$$E\left[\frac{(1 - \Delta_2) I(\widetilde{Y} \leq v)}{(1 - F(\widetilde{Y}))(1 - G(\widetilde{Y}))}\right] = E\left[E\left(\frac{I(Y > C) I(C \leq v)}{(1 - F(C))(1 - G(C))} | C\right)\right] =$$

$$E\left[\frac{I(C \leq v)}{(1 - F(C))(1 - G(C))} E\left(I(Y > C) | C\right)\right] = \int_0^v \frac{1}{1 - G(y)} G(dy) = -\ln(1 - G(v)).$$

Then, by the Strong Law of Large numbers (SLLN) we have w.p.1 as $n \to \infty$

$$n^{-1} \sum_{i=1}^n \xi^\varphi(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0) \gamma_0(\widetilde{Y}_i) \to \int_0^\infty \int_0^{\tau_{\widetilde{H}}} \xi^\varphi(u, v) m(u, v; \beta_0) \gamma_0(v) H(du, dv) =$$

$$\int_0^\infty \int_0^{\tau_{\widetilde{H}}} \xi^\varphi(u, v) \frac{1}{1 - G(v)} H^1(du, dv) = E\left[\frac{\xi^\varphi(\widetilde{T}_1, \widetilde{Y}) \Delta_2}{1 - G(\widetilde{Y})} I(\widetilde{Y} \leq \tau_{\widetilde{H}})\right] =$$

$$E\left[\frac{\xi^\varphi(T_1, Y) I(Y \leq C)}{1 - G(Y)} I(Y \leq \tau_{\widetilde{H}})\right] = E\left[E\left(\frac{\xi^\varphi(T_1, Y) I(Y \leq C)}{1 - G(Y)} I(Y \leq \tau_{\widetilde{H}}) | T_1, Y\right)\right] =$$

$$E\left[\frac{\xi^\varphi(T_1, Y)}{1 - G(Y)} I(Y \leq \tau_{\widetilde{H}}) E\left(I(Y \leq C) | T_1, Y\right)\right] = E\left[\frac{\xi^\varphi(T_1, Y)}{1 - G(Y)} (1 - G(Y)) I(Y \leq \tau_{\widetilde{H}})\right] =$$

$$E\left[\xi^\varphi(T_1, Y) I(Y \leq \tau_{\widetilde{H}})\right] = E\left[\varphi(T_1, T_2) I(Y \leq \tau_{\widetilde{H}})\right] = \int_0^\infty \int_0^\infty \varphi(u, v) F_{12}^0(du, dv).$$

This shows that the first term in (2.7), which is also the first term in the representation given in Theorem 2.5.1, is responsable for the limit of $S_n(\varphi)$.

In the following lemma, we give a representation for

$$n^{-1} \sum_{i=1}^n \xi^\varphi(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0) \gamma_0(\widetilde{Y}_i) C_{in}(\beta_n).$$

Later on, we will show that the integrating measure $H_n(dx, dy)$ appearing on the right-hand side of this representation can be replaced by $H(dx, dy)$. This will allow to get a representation of the above quantity in terms of i.i.d. random variables plus a remainder.

Lemma 2.5.3. *If $\widetilde{H}$ is continuous, $\int |\varphi| dF_{12}^0 < \infty$, and the assumptions (C1)-(C5) and (M1) are satisfied, then, as $n \to \infty$,*

$$n^{-1}\sum_{i=1}^{n}\xi^{\varphi}(\widetilde{T}_{1i},\widetilde{Y}_i)m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)\gamma_0(\widetilde{Y}_i)C_{in}(\beta_n)$$

$$=\int_0^\infty\int_0^\infty \xi^{\varphi}(x,y)m(x,y;\beta_0)\gamma_0(y)I_n(y)H_n(dx,dy)+o_p(n^{-1/2})$$

where

$$I_n(y)=\int_0^\infty\int_0^y\frac{1-m(u,v;\beta_0)}{1-\widetilde{H}(v)}d(H_n(u,v)-H(u,v))+$$

$$\int_0^\infty\int_0^y\frac{\widetilde{H}_n(v)-\widetilde{H}(v)}{(1-\widetilde{H}(v))^2}H^0(du,dv)-n^{-1}\sum_{i=1}^{n}K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^y\frac{\alpha(u,v,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(v)}H(du,dv).$$

Proof to Lemma 2.5.3

The expression $C_{in}(\beta_n)$ is given by

$$\begin{aligned}C_{in}(\beta_n) &= \int_0^\infty\int_0^{\widetilde{Y}_i^-}\frac{1-m(x,y;\beta_n)}{1-\widetilde{H}_n(y)}H_n(dx,dy)-\int_0^\infty\int_0^{\widetilde{Y}_i}\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)}H(dx,dy)\\ &\equiv \Lambda_n^0(\widetilde{Y}_i^-)-\Lambda^0(\widetilde{Y}_i).\end{aligned}$$

Note that, under (M1) and by the SLLN, the result follows from

$$\begin{aligned}\max_{i:\widetilde{Y}_i\le T}|C_{in}(\beta_n)-I_n(\widetilde{Y}_i)| &\le \sup_{0\le y\le T}\left|\Lambda_n^0(y^-)-\Lambda^0(y)-I_n(y)\right|\\ &= o_P(n^{-1/2}).\end{aligned}\tag{2.8}$$

Now we prove (2.8) following lines similar to those in the proof to Lemma 3.12 in Dikta (1998).

A straightforward calculation shows

$$\frac{1-m(x,y;\beta_n)}{1-\widetilde{H}_n(y)}=\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)}+\frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2}$$

$$+\frac{(1-m(x,y;\beta_n))-(1-m(x,y;\beta_0))}{1-\widetilde{H}(y)}+\frac{(1-m(x,y;\beta_n))(\widetilde{H}_n(y)-\widetilde{H}(y))^2}{(1-\widetilde{H}_n(y))(1-\widetilde{H}(y))^2}$$

$$+\frac{(1-m(x,y;\beta_n))-(1-m(x,y;\beta_0))}{(1-\widetilde{H}(y))^2}(\widetilde{H}_n(y)-\widetilde{H}(y))$$

$$\equiv I_1(x,y)+I_2(x,y)+I_3(x,y)+I_4(x,y)+I_5(x,y).$$

Now:

$$n^{1/2}\sup_{0\le t\le T}\left|\int_0^\infty\int_0^t I_5(x,y)H_n(dx,dy)\right| \le \quad n^{1/2}\frac{||\widetilde{H}_n-H||}{(1-\widetilde{H}(T))^2}$$

$$\times\int_0^\infty\int_0^T |m(x,y;\beta_n)-m(x,y;\beta_0)|H_n(dx,dy).$$

Taylor's expansion together with (C1) and (C5) yields that the right-hand side above is bounded by

$$n^{1/2}\frac{||\widetilde{H}_n-H||}{(1-\widetilde{H}(T))^2}\left\{k\sup_{0\le x,y\le T}||Grad(m_1(x,y;\beta_0))||.||\beta_n-\beta_0||\right.$$

$$\left.+\frac{||\beta_n-\beta_0||^2}{2}k^2\sup_{\beta\in V(\beta_0)}\sup_{0\le x<y<\infty}\sum_{1\le i,j\le k}|D_{i,j}m_1(x,y;\beta)|\right\}.$$

Since $n^{1/2}||\widetilde{H}_n-H||$ is bounded in probability, the consistency of β_n together with the SLLN guarantee that

$$n^{1/2}\sup_{0\le t\le T}\left|\int_0^\infty\int_0^t I_5(x,y)H_n(dx,dy)\right| = o_P(1).$$

Now, take $\varepsilon>0$ such that $\widetilde{H}(T)+\varepsilon<1$. Then we get with probability one for large n

$$n^{1/2}\sup_{0\le t\le T}\left|\int_0^\infty\int_0^t I_4(x,y)H_n(dx,dy)\right| =$$

$$= n^{1/2}\sup_{0\le t\le T}\left|\int_0^\infty\int_0^t \frac{(1-m(x,y;\beta_n))(\widetilde{H}_n(y)-\widetilde{H}(y))^2}{(1-\widetilde{H}_n(y))(1-\widetilde{H}(y))^2}H_n(dx,dy)\right|$$

$$\le n^{1/2}\frac{||\widetilde{H}_n-\widetilde{H}||^2}{(1-\widetilde{H}(T))^2}\int_0^\infty\int_0^t\frac{|m(x,y;\beta_n)|}{|1-\widetilde{H}_n(y)|}H_n(dx,dy)$$

$$\le n^{1/2}\frac{||\widetilde{H}_n-\widetilde{H}||^2}{(1-\widetilde{H}(T)-\varepsilon)^3}\int_0^\infty\int_0^t H_n(dx,dy) = o_P(1).$$

For the third term we proceed as in Lemma 3.5 in Dikta (1998) to get

$$\sup_{0\le t\le T}\quad\left|n^{1/2}\int_0^\infty\int_0^t I_3(x,y)H_n(dx,dy)+\right.$$

$$\left.n^{-1/2}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^t\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)}H_n(dx,dy)\right| = o_P(1).$$

This shows that

$$\begin{aligned} n^{1/2}(\Lambda_n^0(t)-\Lambda^0(t)) = \; & n^{1/2}\int_0^\infty\int_0^t \frac{(1-m(x,y;\beta_0))}{1-\widetilde{H}(y)} d(H_n(x,y)-H(x,y)) \\ & + n^{1/2}\int_0^\infty\int_0^t \frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2} H_n(dx,dy) \\ & - n^{-1/2}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) \\ & \times \int_0^\infty\int_0^t \frac{\alpha(u,v,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(v)} H_n(du,dv) + o_P(1) \end{aligned} \tag{2.9}$$

uniformly in $0 \le t \le T$. The next four lemmas allow to replace the empirical measure $H_n(dx,dy)$ by the theoretical one $H(dx,dy)$ in the second and the third terms of equation (2.9), and hence the proof is complete. □

Lemma 2.5.4. *If $\widetilde{H}(T) < 1$ then, as $n \to \infty$*

$$\begin{aligned} n^{1/2} \sup_{0\le t\le T} \Bigg| \int_0^\infty \; & \int_0^t \frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2} H_n(dx,dy) \\ & - \frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2} H(dx,dy) \Bigg| \end{aligned}$$

tends to 0 in probability.

Proof to Lemma 2.5.4

First observe that

$$U_n(t) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} h(\widetilde{T}_{1i},\widetilde{Y}_i,\widetilde{T}_{1j},\widetilde{Y}_j) 1_{\{\widetilde{Y}_i \le t\}}$$

is a U-statistic process as studied in Stute (1994) with kernel

$$h(x,y,u,v) = \frac{(1-m(x,y;\beta_0))(1_{\{v\le y\}}-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2}.$$

We have

$$\int_0^\infty \int_0^t \frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2} H_n(dx,dy)$$

$$= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{(1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))(I(\widetilde{Y}_j\le\widetilde{Y}_i)-\widetilde{H}(\widetilde{Y}_i))}{(1-\widetilde{H}(\widetilde{Y}_i))^2} I(\widetilde{Y}_i\le t)$$

$$= \frac{1}{n^2}\sum_{1\le i\ne j\le n}(\cdots) + \frac{1}{n^2}\sum_{i=j=1}^n(\cdots)$$

$$= \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} h(\widetilde{T}_{1i},\widetilde{Y}_i,\widetilde{T}_{1j},\widetilde{Y}_j) I(\widetilde{Y}_i\le t)$$

$$-\frac{1}{n}\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} h(\widetilde{T}_{1i},\widetilde{Y}_i,\widetilde{T}_{1j},\widetilde{Y}_j) I(\widetilde{Y}_i\le t)$$

$$+\frac{1}{n^2}\sum_{i=1}^n \frac{1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}{1-\widetilde{H}(\widetilde{Y}_i)} I(\widetilde{Y}_i\le t)$$

$$= U_n(t) - \frac{1}{n}U_n(t) + \frac{1}{n^2}\sum_{i=1}^n \frac{1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}{1-\widetilde{H}(\widetilde{Y}_i)} I(\widetilde{Y}_i\le t).$$

Obviously,

$$\sup_{0\le t\le T} \quad \frac{1}{n^2}\sum_{i=1}^n \frac{1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}{1-\widetilde{H}(\widetilde{Y}_i)} I(\widetilde{Y}_i\le t)$$

$$\le \frac{1}{n(1-\widetilde{H}(T))} = O(n^{-1})$$

and

$$\sup_{0\le t\le T}\left|\frac{1}{n}U_n(t)\right| \le \frac{1}{n(1-\widetilde{H}(T))^2} = O(n^{-1}).$$

Since $\int\int h(x,y,u,v)H(du,dv)=0$ and $h\in L_2(H\otimes H)$, Corollary 1.1 in Stute (1994) then yields

$$E\left(\sup_{0\le t\le T}\left|U_n(t) - \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h(x,y,u,v)H(dx,dy)H_n(du,dv)\right|^2\right) = O(n^{-2}).$$

Finally,

$$\int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h(x,y,u,v)H(dx,dy)H_n(du,dv) =$$

$$\frac{1}{n}\sum_{i=1}^n \frac{(1-m(x,y;\beta_0))(1_{\{\widetilde{Y}_i\le y\}}-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2}H(dx,dy) =$$

$$\int_0^\infty \int_0^t \frac{(1-m(x,y;\beta_0))(\widetilde{H}_n(y)-\widetilde{H}(y))}{(1-\widetilde{H}(y))^2}H(dx,dy). \quad \square$$

Lemma 2.5.5. *If $\widetilde{H}$ is continuous, $\widetilde{H}(T)<1$, and assumptions (C1)-(C5) are satisfied, then with probability 1 as $n\to\infty$,*

$$n^{-1/2}\textstyle\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\displaystyle\int_0^\infty\int_0^t \frac{\alpha(u,v,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(v)}H_n(du,dv) =$$

$$= n^{1/2}(U_{1n}(t)-U_{2n}(t))+O(n^{-1/2})$$

uniformly on $[0,T]$, where

$$U_{1n}(t) = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\frac{\Delta_{1i}\Delta_{2i}\alpha(\widetilde{T}_{1j},\widetilde{Y}_j,\widetilde{T}_{1i},\widetilde{Y}_i)}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)(1-\widetilde{H}(\widetilde{Y}_j))}1_{\{\widetilde{Y}_j\le t\}}$$

$$U_{2n}(t) = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\frac{\Delta_{1i}(1-\Delta_{2i})\alpha(\widetilde{T}_{1j},\widetilde{Y}_j,\widetilde{T}_{1i},\widetilde{Y}_i)}{(1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))(1-\widetilde{H}(\widetilde{Y}_j))}1_{\{\widetilde{Y}_j\le t\}}$$

Proof to Lemma 2.5.5

Since

$$K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\Delta_{1i} = \frac{\Delta_{1i}\Delta_{2i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)} - \frac{\Delta_{1i}(1-\Delta_{2i})}{1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}$$

a straightforward calculation shows that the left-hand side equals

$$n^{1/2}(U_{1n}(t)-U_{2n}(t)) - n^{-1/2}(U_{1n}(t)-U_{2n}(t)) + n^{-1/2}R_n(t)$$

where

$$R_n(t) = n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\Delta_{1i}\frac{\alpha(\widetilde{T}_{1i},\widetilde{Y}_i,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(\widetilde{Y}_i)}1_{\{\widetilde{Y}_i\le t\}}$$

According to (C5) we have for an appropriate constant $c > 0$

$$\sup_{0\le x,y\le T}\left|\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)\Delta_{1i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}\right| \le c\sum_{s=1}^{k}\frac{|D_s(m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))|\Delta_{1i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}.$$

From (C3) and the SLLN we then get with probability one, as $n\to\infty$,

$$\begin{aligned}\sup_{0\le t\le T}|U_{1n}(t)| = & \sup_{0\le t\le T}\left|\frac{1}{n(n-1)}\sum_{1\le i\ne j\le n}\frac{\Delta_{2i}\alpha(\widetilde{T}_{1j},\widetilde{Y}_j,\widetilde{T}_{1i},\widetilde{Y}_i)\Delta_{1i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)(1-\widetilde{H}(\widetilde{Y}_j))}1_{\{\widetilde{Y}_j\le t\}}\right| \\ & \le \frac{c}{n(1-\widetilde{H}(T))}\sum_{i=1}^{n}\sum_{s=1}^{k}\frac{|D_s(m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))|\Delta_{1i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)} = O(1).\end{aligned}$$

A similar approximation also holds for $U_{2n}(t)$ and $R_n(t)$ which proves the result. □

Lemma 2.5.6. *If $\widetilde{H}$ is continuous, $\widetilde{H}(T) < 1$, and assumptions (C1)-(C5) are satisfied, then as $n\to\infty$,*

$$\begin{aligned}n^{1/2}\sup_{0\le t\le T}\Big|U_{1n}(t) &- \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_1(x,y,u,v)H(dx,dy)\overline{H}_{1n}(du,dv) \\ &- \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_1(x,y,u,v)H_n(dx,dy)\overline{H}_1(du,dv) \\ &+ \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_1(x,y,u,v)H(dx,dy)\overline{H}_1(du,dv)\Big|\end{aligned}$$

tends to 0 in probability, where

$$h_1(x,y,u,v) = \frac{1_{\{u>0,v>0\}}\alpha(x,y,u,v)}{m(u,v;\beta_0)(1-\widetilde{H}(y))}1_{\{y\le T\}}$$

$\overline{H}_1$ is the d.f. of $(\overline{\widetilde{T}}_1,\overline{\widetilde{Y}}) = (\Delta_1\widetilde{T}_1,\Delta_2\widetilde{Y})$ and $\overline{H}_{1n}$ the empirical d.f. of the $(\overline{\widetilde{T}}_1,\overline{\widetilde{Y}})$-sample.

Proof to Lemma 2.5.6

Since $\widetilde{H}$ is continuous

$$\frac{\Delta_{1i}\Delta_{2i}\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)} = \frac{1_{\{\Delta_{1i}\widetilde{T}_{1i}>0,\Delta_{2i}\widetilde{Y}_i>0\}}\alpha(x,y,\Delta_{1i}\widetilde{T}_{1i},\Delta_{2i}\widetilde{Y}_i)}{m(\Delta_{1i}\widetilde{T}_{1i},\Delta_{2i}\widetilde{Y}_i;\beta_0)}$$

with probability one. Therefore,

$$U_{1n}(t) = \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} h_1(\widetilde{T}_{1j}, \widetilde{Y}_j, \Delta_{1i}\widetilde{T}_{1i}, \Delta_{2i}\widetilde{Y}_i) 1_{\{\widetilde{Y}_j \le t\}}$$

is a U-statistic process with

$$\begin{aligned}\int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_1^2(x,y,u,v) H(dx,dy)\overline{H}_1(du,dv) &= E\left(\left(\frac{\Delta_{12}\Delta_{22}\alpha(\widetilde{T}_{11},\widetilde{Y}_1,\widetilde{T}_{12},\widetilde{Y}_2)}{m(\widetilde{T}_{12},\widetilde{Y}_2;\beta_0)(1-\widetilde{H}(\widetilde{Y}_1))} I(\widetilde{Y}_1 \le T)\right)^2\right) \\ &\le cE\left(\left(\sum_{s=1}^k \frac{D_s(m(\widetilde{T}_{12},\widetilde{Y}_2;\beta_0))\Delta_{12}}{m(\widetilde{T}_{12},\widetilde{Y}_2;\beta_0)}\right)^2\right) < \infty\end{aligned}$$

according to (C3) and (C5) for an appropriate constant c. An application of Theorem 1.5 in Stute (1994) completes the proof. □

A similar result holds for $U_{2n}(t)$:

Lemma 2.5.7. *If $\widetilde{H}$ is continuous, $\widetilde{H}(T) < 1$, and assumptions (C1)-(C5) are satisfied, then as $n \to \infty$,*

$$\begin{aligned} n^{1/2} \sup_{0\le t\le T} \Big| & U_{2n}(t) - \int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h_2(x,y,u,v) H(dx,dy)\overline{H}_{2n}(du,dv) \\ & - \int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h_2(x,y,u,v) H_n(dx,dy)\overline{H}_2(du,dv) \\ & + \int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h_2(x,y,u,v) H(dx,dy)\overline{H}_2(du,dv)\Big| \end{aligned}$$

tends to zero in probability, where

$$h_2(x,y,u,v) = \frac{1_{\{u>0,v>0\}}\alpha(x,y,u,v)}{(1-m(u,v;\beta_0))(1-\widetilde{H}(y))} 1_{\{y\le T\}}$$

$\overline{H}_2$ is the d.f. of $(\overline{\overline{\widetilde{T}}}_1, \overline{\overline{\widetilde{Y}}}) = (\Delta_1\widetilde{T}_1, \Delta_1(1-\Delta_2)\widetilde{Y})$ and $\overline{H}_{2n}$ the empirical d.f. of the $(\overline{\overline{\widetilde{T}}}_1, \overline{\overline{\widetilde{Y}}})$-sample. □

Straightforward calculation shows that

$$\begin{aligned} & \int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h_1(x,y,u,v) H(dx,dy)\overline{H}_{1n}(du,dv) - \int_0^\infty \int_0^\infty \int_0^\infty \int_0^t h_2(x,y,u,v) H(dx,dy)\overline{H}_{2n}(du,dv) \\ & = n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) \int_0^\infty \int_0^t \frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)} H(dx,dy) \end{aligned}$$

and, since by conditioning on $(\widetilde{T}_1, \widetilde{Y})$, $E(\alpha(x,y,\widetilde{T}_1,\widetilde{Y})K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) = 0$, we obtain

$$\int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_1(x,y,u,v)H_n(dx,dy)\overline{H}_1(du,dv) - \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_2(x,y,u,v)H_n(dx,dy)\overline{H}_2(du,dv) = 0$$

and

$$\int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_1(x,y,u,v)H(dx,dy)\overline{H}_1(du,dv) - \int_0^\infty\int_0^\infty\int_0^\infty\int_0^t h_2(x,y,u,v)H(dx,dy)\overline{H}_2(du,dv) = 0.$$

Therefore, the equation (2.8) in Lemma 2.5.3 (and hence Lemma 2.5.3 itself) follows from the last four lemmas as announced.

The next three lemmas will demonstrate that the integrating measure H_n appearing on the right-hand side of the representation in Lemma 2.5.3 can be replaced by H, i.e. we replace the right-hand side by the corresponding projection.

Lemma 2.5.8. *If $\widetilde{H}$ is continuous and (M1)-(M2) are satisfied, then, as $n \to \infty$,*

$$\int_0^\infty\int_0^\infty\int_0^\infty\int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)}d(H_n(x,y)-H(x,y))H_n(du,dv) =$$

$$\int_0^\infty\int_0^\infty\int_0^\infty\int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)}d(H_n(x,y)-H(x,y))H(du,dv)$$

$$+O_p(n^{-1}).$$

Proof to Lemma 2.5.8

In the above equation we denote the left-hand side by V_n, and observe that

$$V_n = \int_0^\infty\int_0^\infty\int_0^\infty\int_0^v h(u,v,x,y)H_n(dx,dy)H_n(du,dv)$$

where

$$h(u,v,x,y) = \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\left(1_{\{y\le v\}}\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)} + \ln(1-G(v))\right).$$

(M1), (M2) and Lemma 5.7.3 in Serfling (1980) guarantee that

$$V_n = U_n + O_P(n^{-1})$$

where U_n denotes the associated U-statistic. Since $h \in L_2(H \otimes H)$, apply Theorem 5.3.2 in Serfling (1980) to get

$$\begin{aligned} V_n &= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h(u,v,x,y) H_n(dx,dy) H(du,dv) \\ &+ \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h(u,v,x,y) H(dx,dy) H_n(du,dv) \\ &- \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h(u,v,x,y) H(dx,dy) H(du,dv) \\ &+ O_P(n^{-1}). \end{aligned}$$

Note that the first integral on the right-hand side is identical to the integral on the right-hand side of our lemma. Since $E(h(u,v,\widetilde{T}_1,\widetilde{Y})) = 0$ and $E(h(\widetilde{T}_{11},\widetilde{Y}_1,\widetilde{T}_{12},\widetilde{Y}_2)) = 0$, the second and third integral on the right-hand side disappear, which completes the proof. Certainly,

$$\begin{aligned} E(h(u,v,\widetilde{T}_1,\widetilde{Y})) &= \xi^{\varphi}(u,v) m(u,v;\beta_0) \gamma_0(v) \left(\int_0^u \int_0^v \frac{1 - m(x,y;\beta_0)}{1 - \widetilde{H}(y)} + \ln(1 - G(v)) \right) H(dx,dy) \\ &= \xi^{\varphi}(u,v) m(u,v;\beta_0) \gamma_0(v) (\ln(1 - G(v)) - \ln(1 - G(v))) = 0 \end{aligned}$$

and

$$\begin{aligned} E(h(\widetilde{T}_{11},\widetilde{Y}_1,\widetilde{T}_{12},\widetilde{Y}_2)) &= \int_0^\infty \int_0^\infty \xi^{\varphi}(u,v) m(u,v;\beta_0) \gamma_0(v) \int_0^u \int_0^v \left(\frac{1 - m(x,y;\beta_0)}{1 - \widetilde{H}(y)} + \ln(1 - G(v)) \right) \times \\ &\times H(dx,dy) H(du,dv) \\ &= \int_0^\infty \int_0^\infty \xi^{\varphi}(u,v) m(u,v;\beta_0) \gamma_0(v) (\ln(1 - G(v)) - \ln(1 - G(v))) H(du,dv) = 0. \square \end{aligned}$$

A similar argumentation can be used to prove:

Lemma 2.5.9. *If $\widetilde{H}$ is continuous and (M1)-(M2) are satisfied, then, as $n \to \infty$,*

$$\begin{aligned} &\int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^{\varphi}(u,v) m(u,v;\beta_0) \gamma_0(v) \frac{\widetilde{H}_n(y) - \widetilde{H}(y)}{(1 - \widetilde{H}(y))^2} H^0(dx,dy) H_n(du,dv) \\ &= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^{\varphi}(u,v) m(u,v;\beta_0) \frac{\widetilde{H}_n(y) - \widetilde{H}(y)}{(1 - \widetilde{H}(y))^2} H^0(dx,dy) H(du,dv) + O_p(n^{-1}). \square \end{aligned}$$

Lemma 2.5.10. *If $\widetilde{H}$ is continuous and (C3)-(C5) and (M1)-(M2) are satisfied, then, as $n \to \infty$,*

$$n^{-1}\sum_{i=1}^{n} K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^\infty\int_0^\infty\int_0^v \xi^{\varphi}(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)}H(dx,dy)H_n(du,dv)$$

$$= n^{-1}\sum_{i=1}^{n} K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^\infty\int_0^\infty\int_0^v \xi^{\varphi}(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)}H(dx,dy)H(du,dv)$$

$$+O_p(n^{-1}).$$

Proof to Lemma 2.5.10

Firstly, note that (from our convention $K(x,x,d)=0$)

$$K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) = K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\Delta_{1i} = \frac{\Delta_{1i}\Delta_{2i}}{m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)} - \frac{\Delta_{1i}(1-\Delta_{2i})}{1-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)}.$$

We denote the left-hand side of the equation in the Lemma by V_n, and observe that

$$V_n = \int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty h_1(u,v,x,y)H_n^{11}(dx,dy)H_n(du,dv)$$

$$-\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty h_0(u,v,x,y)H_n^{10}(dx,dy)H_n(du,dv)$$

$$= V_n^1 - V_n^0$$

where

$$H_n^{11}(x,y) = n^{-1}\sum_{i=1}^{n} I(\widetilde{T}_{1i}\le x,\widetilde{Y}_i\le y,\Delta_{1i}\Delta_{2i}=1) = H_n^1(x,y),$$

$$H_n^{10}(x,y) = n^{-1}\sum_{i=1}^{n} I(\widetilde{T}_{1i}\le x,\widetilde{Y}_i\le y,\Delta_{1i}(1-\Delta_{2i})=1),$$

$$h_1(u,v,x,y) = \xi^{\varphi}(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1}{m_1(x,y;\beta_0)}\int_0^\infty\int_0^v \frac{\alpha(r,t,x,y)}{1-\widetilde{H}(t)}H(dr,dt)$$

and

$$h_0(u,v,x,y) = \xi^{\varphi}(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1}{1-m_1(x,y;\beta_0)}\int_0^\infty\int_0^v \frac{\alpha(r,t,x,y)}{1-\widetilde{H}(t)}H(dr,dt).$$

Due to $(C5)$, $Grad(m_1(.,.;\beta_0))$ is bounded and thus we obtain for an appropriate constant $c > 0$

$$\left| \frac{\Delta_{1i}}{m_1(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0)} \int_0^\infty \int_0^v \frac{\alpha(r, t, \widetilde{T}_{1i}, \widetilde{Y}_i)}{1 - \widetilde{H}(t)} H(dr, dt) \right|$$

$$\leq \frac{c}{1 - \widetilde{H}(v)} \sum_{j=1}^{k} \frac{|D_j(m_1(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0))\Delta_{1i}|}{m_1(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0)}.$$

(M1)-(M2), $(C3)$, and Lemma 5.7.3 in Serfling (1980) ensure that

$$V_n^1 = U_n^1 + O_P(n^{-1})$$

where U_n^1 represents the associated U-statistic. Note that

$$\int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_1^2(u, v, x, y) H^1(dx, dy) H(du, dv)$$

$$\leq \int_0^\infty \int_0^\infty \left(\frac{\xi^\varphi(u, v) m(u, v; \beta_0) \gamma_0(v)}{1 - \widetilde{H}(v)} \right)^2 H(du, dv)$$

$$\times c^2 \int_0^\infty \int_0^\infty \left(\sum_{j=1}^{k} \frac{|D_j(m_1(x, y; \beta_0))|}{m_1(x, y; \beta_0)} \right)^2 H^1(dx, dy) < \infty$$

and apply Theorem 5.3.2 in Serfling (1980) to obtain

$$\begin{aligned} V_n^1 &= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_1(u, v, x, y) H_n^1(dx, dy) H(du, dv) \\ &+ \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_1(u, v, x, y) H^{11}(dx, dy) H_n(du, dv) \\ &- \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_1(u, v, x, y) H^{11}(dx, dy) H(du, dv) + O_P(n^{-1}), \end{aligned}$$

where $H^{11}(x, y) = P(\widetilde{T}_1 \leq x, \widetilde{Y} \leq y, \Delta_1 \Delta_2 = 1)$. A similar result holds for V_n^0. Specifically, with $H^{10}(x, y) = P(\widetilde{T}_1 \leq x, \widetilde{Y} \leq y, \Delta_1(1 - \Delta_2) = 1)$,

$$\begin{aligned} V_n^0 &= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_0(u, v, x, y) H_n^{10}(dx, dy) H(du, dv) \\ &+ \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_0(u, v, x, y) H^{10}(dx, dy) H_n(du, dv) \\ &- \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty h_0(u, v, x, y) H^{10}(dx, dy) H(du, dv) + O_P(n^{-1}). \end{aligned}$$

Since

$$\int_0^\infty \int_0^\infty h_1(u,v,x,y)H^{11}(dx,dy) - \int_0^\infty \int_0^\infty h_0(u,v,x,y)H^{10}(dx,dy) = 0,$$

we get

$$V_n = n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) \int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)}H(dx,dy)H(du,dv)$$

$$+O_p(n^{-1}),$$

which concludes the proof of the lemma. □

Now combine the last three lemmas and Lemma 2.5.3, and observe that some of the terms cancel out to get:

Lemma 2.5.11. *If $\widetilde{H}$ is continuous, and (C1)-(C5) and (M1)-(M2) are satisfied, then, as $n \to \infty$,*

$$n^{-1}\sum_{i=1}^n \xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)\gamma_0(\widetilde{Y}_i)C_{in}(\beta_n)$$

$$= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1-m(x,y;\beta_0)}{1-\widetilde{H}(y)}H_n^0(dx,dy)H(du,dv)$$

$$- \int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\frac{1-\widetilde{H}_n(y)}{(1-\widetilde{H}(y))^2}H^0(dx,dy)H(du,dv)$$

$$-n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i}) \int_0^\infty \int_0^\infty \int_0^\infty \int_0^v \xi^\varphi(u,v)m(u,v;\beta_0)\gamma_0(v)\times$$

$$\frac{\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)}{1-\widetilde{H}(y)}H(dx,dy)H(du,dv) + o_p(n^{-1/2}). \quad \square$$

For the next quantity,

$$n^{-1}\sum_{i=1}^n \xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)\gamma_0(\widetilde{Y}_i)B_{in}(\beta_n),$$

that arises in the first term on the right-hand side of (2.7), we proceed as in Stute (1995) and use

$$x - \frac{x^2}{2} \le \ln(1+x) \le x \quad \text{for} \quad x \ge 0,$$

to get

$$-\frac{n^{-1}}{2}\int_0^\infty\int_0^{\widetilde{Y}_i^-}\frac{(1-m(x,y;\beta_0))^2}{(1-\widetilde{H}_n(y))^2}H_n(dx,dy)\le B_{in}(\beta_n)\le 0. \tag{2.10}$$

The SLLN, Glivenko - Cantelli, and $(M1)$ then guarantee

Lemma 2.5.12. *If $\widetilde{H}$ is continuous, $\int|\varphi|dF_{12}^0<\infty$, and $(M1)$ are satisfied, then, as $n\to\infty$,*

$$n^{-1}\sum_{i=1}^n\xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0)\gamma_0(\widetilde{Y}_i)B_{in}(\beta_n)=O(n^{-1})\quad w.p.1.\quad\square$$

In the next lemma we transform the second term on the right-hand side of our representation (2.7) into a sum of i.i.d. random variables.

Lemma 2.5.13. *If $\widetilde{H}$ is continuous, and $(C1)$-$(C6)$ and $(M1)$-$(M2)$ are satisfied, then, as $n\to\infty$,*

$$n^{-1}\sum_{i=1}^n\xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)(m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_n)-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))\gamma_0(\widetilde{Y}_i)\,(1+B_{in}(\beta_n)+C_{in}(\beta_n))$$

$$=n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^\infty\xi^\varphi(x,y)\gamma_0(y)\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)H(dx,dy)+o_p(n^{-1/2}).$$

Proof to Lemma 2.5.13

$(M1)$ guarantees that we can restrict our attention to $[0,T]$ with $\widetilde{H}(T)<1$. Due to Lemma 3.5 in Dikta (1998), we first observe that uniformly on $0\le x\le y\le T$.

$$m(x,y;\beta_n)-m(x,y;\beta_0)=n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)+O_P(n^{-1}). \tag{2.11}$$

Thus $(M1)$ and the SLLN guarantee

$$n^{-1}\quad\sum_{i=1}^n\xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)(m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_n)-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))\gamma_0(\widetilde{Y}_i)$$

$$=n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^\infty\xi^\varphi(x,y)\gamma_0(y)\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)H_n(dx,dy)+O_p(n^{-1})$$

$$=n^{-1}\sum_{i=1}^n K(\widetilde{T}_{1i},\widetilde{Y}_i,\Delta_{2i})\int_0^\infty\int_0^\infty\xi^\varphi(x,y)\gamma_0(y)\alpha(x,y,\widetilde{T}_{1i},\widetilde{Y}_i)H(dx,dy)+O_p(n^{-1}),$$

where the last step is supported by the same arguments used in the proof of the Lemma 2.5.10. Now $(M1)$, (2.10), Lemma 2.5.1, and the SLLN implies that

$$n^{-1}\sum_{i=1}^n\xi^\varphi(\widetilde{T}_{1i},\widetilde{Y}_i)(m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_n)-m(\widetilde{T}_{1i},\widetilde{Y}_i;\beta_0))\gamma_0(\widetilde{Y}_i)B_{in}(\beta_n)=O_P(n^{-3/2}).$$

For the final term, recall (2.8) to obtain

$$n^{-1} \quad \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i)(m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) - m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0))\gamma_0(\widetilde{Y}_i)C_{in}(\beta_n)$$

$$= n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i)(m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) - m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0))\gamma_0(\widetilde{Y}_i)I_n(\widetilde{Y}_i) + o_P(n^{-1/2}),$$

where I_n is defined in Lemma 2.5.3. Now the proof of Theorem 2.5 in Dikta (1998) shows, that a process analogous to $n^{1/2}I_n(y)$, $0 \le y \le T$, tends weakly to a centered Gaussian process which is concentrated on $C[0,T]$. Hence, similarly for $n^{1/2}I_n(y)$, we have

$$\sup_{0 \le y \le T} |I_n(y)| = O_P(n^{-1/2}) \tag{2.12}$$

and we get in summary

$$n^{-1} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i)(m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) - m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_0))\gamma_0(\widetilde{Y}_i)C_{in}(\beta_n) = o_P(n^{-1/2}),$$

which concludes the proof of the lemma. □

The following lemma demonstrates that the third term on the right-hand side of (2.7) is negligible and thus we can obtain the desired representation for our estimator as a sum of i.i.d. random variables.

Lemma 2.5.14. *If $\widetilde{H}$ is continuous, and (C1)-(C6) and (M1)-(M2) are satisfied, then, as $n \to \infty$,*

$$\frac{n^{-1}}{2} \sum_{i=1}^{n} \xi^{\varphi}(\widetilde{T}_{1i}, \widetilde{Y}_i) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) m(\widetilde{T}_{1i}, \widetilde{Y}_i; \beta_n) \exp(\xi_i) \left(B_{in}(\beta_n) + C_{in}(\beta_n)\right)^2 = O_p(n^{-1}).$$

Proof to Lemma 2.5.14

Recall from representation (2.7) that ξ_i lies between the terms

$$B_{in}(\beta_n) + \int_0^{\infty} \int_0^{\widetilde{Y}_i^-} \frac{1 - m(x,y;\beta_n)}{1 - \widetilde{H}_n(y)} H_n(dx,dy) \quad \text{and} \quad \int_0^{\infty} \int_0^{\widetilde{Y}_i} \frac{1 - m(x,y;\beta_0)}{1 - \widetilde{H}(y)} H(dx,dy).$$

Therefore, we get due to $(M1)$, Glivenko-Cantelli, and inequality (2.10)

$$\sup_{i:\widetilde{Y}_i \le T} \exp(\xi_i) < c$$

for an appropriate constant c. Now the term on the left-hand side in our lemma is bounded by

$$c \int \int |\xi^{\varphi}(x,y)| H_n(dx,dy) \times \sup_{i:\widetilde{Y}_i \le T} \left(B_{in}(\beta_n) + C_{in}(\beta_n)\right)^2$$

Due to the SLLN and the bounds given under (2.8), (2.10), and (2.12), this term is $O_P(n^{-1})$. This finally concludes the proof. □

2.6 Proof to the consistency result

In this Section we give the technical proof to the consistency result in Section 2.2 (Theorem 2.2.1). We will see that this proof is similar to that of Theorem 2.1 in Uña-Álvarez and Rodríguez-Campos (2004); here, the role of their covariate vector is played by the first gap time, while the total time Y is taken as the 'response'. Note that, since C is assumed to be independent of (T_1, T_2), the identifiability conditions H1 and H2 in de Uña-Álvarez and Rodríguez-Campos (2004) automatically hold. In our setup, these conditions read

H1. Y and C are independent

H2. $P(Y \leq C|T_1, Y) = P(Y \leq C|Y)$

which clearly follow from the independence between the censoring time and the gap times.

In order to formalize things, introduce

$$S_n(m) = \sum_{i=1}^{n} W_i(m)\varphi(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}) = \sum_{i=1}^{n} W_i(m)\xi^{\varphi}(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}),$$

where $W_i(m)$ are the presmoothed weights introduced in Section 2.2 and where $\xi^{\varphi}(u, v) = \varphi(u, v-u)$. Note that this $S_n(m)$ is an 'estimator' of $S(\varphi) = E[\varphi(T_1, T_2)] = E[\xi^{\varphi}(T_1, Y)]$ based on the true m which in practice will be unknown. Recall that the proposed semiparametric estimator of $S(\varphi)$ is

$$S_n(\varphi) = \int \varphi d\widehat{F}_{12}^{sp} = \sum_{i=1}^{n} W_i(\beta_n)\varphi(\widetilde{T}_{[1i:n]}, \widetilde{T}_{[2i:n]}) = \sum_{i=1}^{n} W_i(\beta_n)\xi^{\varphi}(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}),$$

where $W_i(\beta_n) = W_i(m_n)$ with $m_n(x, y) = m(x, y; \beta_n)$ the presmoother based on the parametric model. As in de Uña-Álvarez and Rodríguez-Campos (2004), we proceed in two steps. First, we show the convergence of $S_n(m)$ to $\int \varphi dF_{12}^0 = E[\xi^{\varphi}(T_1, Y)I(Y \leq \tau_H)]$, and then we prove that the difference $S_n(\varphi) - S_n(m)$ goes to zero under appropriate conditions.

For proving the consistency of $S_n(m)$ we need three Lemmas. The first one states the supermartingale structure of $S_n(m)$, which enables us to apply powerful convergence results. The other two Lemmas allow for the identification of the limit.

Introduce the sequence $(\mathcal{F}_n)_{n\geq 1}$, where

$$\mathcal{F}_n = \sigma\left(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}, 1 \leq i \leq n, \widetilde{T}_{1,n+1,}\widetilde{Y}_{n+1,\ldots}\right).$$

Note that $S_n(m)$ is adapted to $\mathcal{F}_n$. Note also that $\mathcal{F}_n \downarrow$ and set $\mathcal{F}_\infty = \cap_{n\geq 1}\mathcal{F}_n$ for the limit of $\mathcal{F}_n$.

Lemma 2.6.1. *Assume that H is continuous. Then,*

$$E\left[S_n(m) \mid \mathcal{F}_{n+1}\right] = S_{n+1}(m) - \frac{\xi^{\varphi}(\widetilde{T}_{[1,n+1:n+1]}, \widetilde{Y}_{n+1:n+1})}{n+1} \times$$

$$\times m(\widetilde{T}_{[1,n+1:n+1]}, \widetilde{Y}_{n+1:n+1})(1 - m(\widetilde{T}_{[1n:n+1]}, \widetilde{Y}_{n:n+1}) \prod_{j=1}^{n-1} \left[1 - \frac{m(\widetilde{T}_{[1j:n+1]}, \widetilde{Y}_{j:n+1})}{n-j+1}\right].$$

In particular, for $\varphi \geq 0$, $(S_n(m), \mathcal{F}_n)_{n\geq 1}$ is a reverse-time supermartingale.

Proof to Lemma 2.6.1 The proof follows exactly the same steps as in the proof to Lemma 4.1 in de Uña-Álvarez and Rodríguez-Campos (2004), which in its turn is a consequence of Lemma 2.1 in Stute (1993), Lemma 2.2 in Stute and Wang (1993), and Lemma 2.1 in Dikta (2000). □

Lemma 2.6.1 allows for the application of the convergence result in Neveu (1975) , Proposition V-3-11. Indeed, the Hewitt-Savage 0-1 law ensures that the limit S of $S_n(m)$ is constant with probability 1. In order to determine $S = \lim_{n\to\infty} E\left[S_n(m)\right]$, we will need the following lemma. This is a proper adaptation to our context of Lemma 2.3 in Stute (1993). Introduce the notation

$$\varphi_n(t) = \prod_{i=1}^{n} \left[1 + \frac{1 - \widetilde{m}(\widetilde{Y}_{i:n})}{n-i+1}\right]^{I(\widetilde{Y}_{i:n}<t)}, \qquad \text{where } \widetilde{m}(z) = E(\Delta_2 \mid \widetilde{Y} = z),$$

and

$$g_n(t) = E\left[\varphi_n(t)\right]; \qquad g_0(t) \equiv 1.$$

Finally,

$$\widetilde{\xi}(z) = E\left[\xi^{\varphi}(\widetilde{T}_1, \widetilde{Y})\Delta_2 \mid \widetilde{Y} = z\right].$$

Lemma 2.6.2. *Under the assumptions of Lemma 2.6.1 , we have*

$$E\left[S_n(m)\right] = E\left[\widetilde{\xi}(\widetilde{Y})g_{n-1}(\widetilde{Y})\right].$$

Proof of Lemma 2.6.2

Similar to that in Stute (1993), Lemma 2.3, after noting that

$$E\left[m(\widetilde{T}_1, \widetilde{Y}) \mid \widetilde{Y} = z\right] = \widetilde{m}(z), \qquad E\left[\xi^{\varphi}(\widetilde{T}_1, \widetilde{Y})m(\widetilde{T}_1, \widetilde{Y}) \mid \widetilde{Y} = z\right] = \widetilde{\xi}(z).$$

Note that the fact that the 'covariate' $\widetilde{T}_1$ is a censored version of the 'true covariate' T_1 is not an issue here, since the outer expectation integrate this variable out. □

Now, by Stute and Wang (1993), we have

$$g_n(t) \uparrow \frac{1}{1-G(t)} \qquad \text{for each } t \text{ such that } H(t) < 1.$$

This fact together with Lemma 2.6.2 will allow for the identification of S.

Lemma 2.6.3. *Under the assumptions of Lemma 2.6.1, we have with probability 1*

$$S_n(m) \to S = \lim_{n\to\infty} E\left[S_n(m)\right] = \int \varphi dF_{12}^0.$$

Proof of Lemma 2.6.3

Assume $\varphi \geq 0$ w.l.o.g. The general case is obtained by decomposing φ into its positive and negative part. Lemma 2.6.2 and the monotone convergence theorem give

$$S = E\left[\widetilde{\xi}(\widetilde{Y})\frac{1}{1-G(\widetilde{Y})}\right] = E\left[\frac{\xi^{\varphi}(\widetilde{T}_1,\widetilde{Y})\Delta_2}{1-G(\widetilde{Y})}\right] = E\left[\frac{\xi^{\varphi}(T_1,\widetilde{Y})\Delta_2}{1-G(\widetilde{Y})}\right] = \int \varphi dF_{12}^0,$$

where for the last equality we have used the independence between C and (T_1, Y). □

For proving that the difference $S_n(\varphi) - S_n(m)$ goes to zero, we need the following result, which is a proper adaptation of Lemma 2.2 in Dikta (2000) to our setup. Introduce for any pair of functions $p(x,z)$ and $q(x,z)$ with $0 \leq q \leq 1$ the quantity

$$\overline{S}_n(p,q) = \sum_{i=1}^{n} \overline{W}_{i,n}(p,q)\varphi(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n})$$

where

$$\overline{W}_{i,n}(p,q) = \frac{p(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n})}{n-i+1}\prod_{j=1}^{i-1}\left[1 - \frac{q(\widetilde{T}_{[1j:n]}, \widetilde{Y}_{j:n})}{n-j+1}\right].$$

The proof, which we omit, is based on martingale properties (as those described in Lemma 2.6.1) of both $\overline{S}_n(p,q)$ and

$$\varphi_{q,n}(t) = \prod_{i=1}^{n}\left[1 + \frac{1-\widetilde{q}(\widetilde{Y}_{i:n})}{n-i+1}\right]^{I(\widetilde{Y}_{i:n}<t)}, \qquad \text{where } \widetilde{q}(z) = E(q(\widetilde{T}_1,\widetilde{Y}) \mid \widetilde{Y} = z).$$

Lemma 2.6.4. *Under assumptions of Lemma 2.6.1, we have with probability 1*

$$\overline{S}_n(p,q) \to \overline{S}(p,q) \equiv E\left[\varphi(\widetilde{T}_1,\widetilde{Y})p(\widetilde{T}_1,\widetilde{Y})\exp\left\{\int_0^{\widetilde{Y}} \frac{1-\widetilde{q}}{1-H}dH\right\}\right].$$

Assume now that condition U holds. Then, since both $m(x,y)$ and $m(x,y;\beta_n)$ are zero for $x = y$, we have

$$\sup_{x,y}|m(x,y;\beta_n) - m(x,y)| \to 0 \qquad \text{w. p. 1.}$$

We have, for a given $\varepsilon > 0$,

$$0 \leq m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n};\beta_n) \leq \left|m_n(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n};\beta_n) - m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n})\right| + m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) \leq$$

$$\leq \varepsilon + m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n})$$

eventually. Similarly, since $a+b \geq |a|-|b|$ whenever $a+b \geq 0$, we eventually have

$$m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n) \geq m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) - \left| m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}; \beta_n) - m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) \right| \geq$$

$$\geq m(\widetilde{T}_{[1i:n]}, \widetilde{Y}_{i:n}) - \varepsilon.$$

Introduce the functions

$$M_{1,\varepsilon}(x,z) = \max(0, m(x,z) - \varepsilon), \qquad M_{2,\varepsilon}(x,z) = \min(1, m(x,z) + \varepsilon).$$

Assume $\varphi \geq 0$ w.l.o.g. Since $M_{2,\varepsilon}(x,z) \leq M_{1,\varepsilon}(x,z) + 2\varepsilon$, we get (with $m_n = m(.,.;\beta_n)$)

$$S_n(m_n) \leq \overline{S}_n(M_{2,\varepsilon}, M_{1,\varepsilon}) \leq S_n(M_{1,\varepsilon}) + 2\varepsilon \overline{S}_n(1, M_{1,\varepsilon})$$

where we use the obvious notation $S_n(q) = \overline{S}_n(q,q)$. We also have

$$S_n(m_n) \geq \overline{S}_n(M_{1,\varepsilon}, M_{2,\varepsilon}) \geq S_n(M_{2,\varepsilon}) - 2\varepsilon \overline{S}_n(1, M_{2,\varepsilon}).$$

Use Lemma 2.6.4 to obtain

$$\begin{aligned} S(M_{2,\varepsilon}) - 2\varepsilon \overline{S}(1, M_{2,\varepsilon}) &\leq \liminf_{n\to\infty} S_n(m_n) \leq \limsup_{n\to\infty} S_n(m_n) \leq \\ &\leq S(M_{1,\varepsilon}) + 2\varepsilon \overline{S}(1, M_{1,\varepsilon}) \end{aligned}$$

where we put $S(q) = \overline{S}(q,q)$. Bounds for $S(M_{2,\varepsilon}) - 2\varepsilon\overline{S}(1, M_{2,\varepsilon})$ and $S(M_{1,\varepsilon}) + 2\varepsilon\overline{S}(1, M_{1,\varepsilon})$ can be easily found as in Dikta (2000):

$$\begin{aligned} S(M_{1,\varepsilon}) + 2\varepsilon \overline{S}(1, M_{1,\varepsilon}) &\leq \int\int \frac{\varphi(x,y)}{(1-H(x+y))^{\varepsilon}} F_{12}^0(dx,dy) + \\ &\quad + 2\varepsilon \int\int \frac{\varphi(x,y)}{m(x,x+y)(1-H(x+y))^{\varepsilon}} F_{12}^0(dx,dy), \end{aligned}$$

$$\begin{aligned} S(M_{2,\varepsilon}) - 2\varepsilon \overline{S}(1, M_{2,\varepsilon}) &\geq \int\int \varphi(x,y)(1-H(x+y))^{\varepsilon} F_{12}^0(dx,dy) - \\ &\quad - 2\varepsilon \int\int \frac{\varphi(x,y)}{m(x,x+y)} F_{12}^0(dx,dy). \end{aligned}$$

Note that $m(x,x+y) = m_1(x,x+y)$ unless T_2 has positive mass at zero, a situation excluded by assumption $P(T_2 = 0) = 0$. Let $\varepsilon \downarrow 0$ and apply the monotone convergence theorem to end with the proof of Theorem 2.2.1. $\square$

Chapter 3

Presmoothing the transition probabilities in the illness-death model

Contents

3.1 Introduction 54
3.2 The presmoothed estimator. Consistency 55
3.3 Simulation study 60
3.4 Real data illustration 64

3.1 Introduction

Multi-state models (Andersen et al. (1993); Meira-Machado et al. (2009)) are the most common models used for the description of longitudinal survival data. A multi-state model is a model for a stochastic process, which is characterized by a set of states and the possible transitions among them. The states represent different situations of the individual (healthy, diseased, etc) along a follow-up. Special multi-state models that have been widely used in biomedical applications are the three-state progressive model, the illness-death model, or the bivariate model (Hougaard (2000)).

Let $X(t)$ represent the state occupied by the process at time $t \geq 0$. For two states i,j and $s < t$, introduce the transition probability

$$p_{ij}(s,t) = P\left(X(t) = j | X(s) = i\right).$$

There has been much interest in the estimation of $p_{ij}(s,t)$ since it allows for long-term predictions of the process. Aalen and Johansen (1978) introduced a nonparametric estimator of $p_{ij}(s,t)$ for Markov models. The Markov assumption states that the future evolution of the process is independent of the previously visited states and the times of transition amongst them given the present state of the process. This simplifying assumption allows for the construction of simple estimators, since individuals with different past histories become comparable. However, it has been quoted that the Markov assumption is violated in some applications (e.g. Andersen et al., 2000). This is a relevant remark, since Aalen-Johansen estimator may be inconsistent if the process is non-Markov. Estimators of $p_{ij}(s,t)$ which are consistent in non-Markov situations are hardly found in literature.

Meira-Machado et al. (2006) introduced a substitute for the Aalen-Johansen estimator in the case of a non-Markov illness-death model. They showed that when the Markov assumption does not hold, the new estimator may behave much better than the Aalen-Johansen which may be systematically biased. However, by removing the Markov condition, the proposed substitute for the Aalen-Johansen estimator provides undesirable large standard errors. This problem becomes worse when there is a large proportion of censored data. In order to overcome this issue, we propose here a modification of Meira-Machado et al. (2006)'s estimator based on presmoothing, which allows for a variance reduction in the presence of censoring.

In order to illustrate our estimators using real data, we consider data from one of the first successful trials of adjuvant chemotherapy for colon cancer, which is freely available from the `R survival` package. In this study, 929 patients affected by colon cancer underwent a potential curative surgery. Unfortunately, some of these patients had residual cancer, which lead to the recurrence of disease and death (in some cases). Therefore, we may consider the recurrence as an associated state of risk, and use the so-called illness-death model with states "alive and disease-

free", "alive with recurrence" (local-regional or metastases) and "dead". See Section 3.2 for a more formal definition of the model.

3.2 The presmoothed estimator. Consistency

In this Chapter we consider the illness-death model depicted in Figure 3.1. In this model, all the subjects are in State 1 ('healthy') at time $t = 0$. At some future time, they will arrive at State 3 ('dead'), which is absorbing. In the meanwhile they may visit State 2 ('diseased') at some time point; or not, passing directly to State 3 without visiting State 2. Note that this multi-state model is progressive (Hougaard, 2000), in the sense that past states can not be revisited. For this model the set of states is $\mathcal{S} = \{1, 2, 3\}$, and the transitions allowed are $1 \to 2$, $1 \to 3$, and $2 \to 3$. Given two time points $s < t$, there are in essence three different transition probabilities to estimate: $p_{11}(s,t)$, $p_{13}(s,t)$, and $p_{23}(s,t)$. The two other transition probabilities ($p_{12}(s,t)$ and $p_{22}(s,t)$) are easily obtained from $p_{12}(s,t) = 1 - p_{11}(s,t) - p_{13}(s,t)$ and $p_{22}(s,t) = 1 - p_{23}(s,t)$.

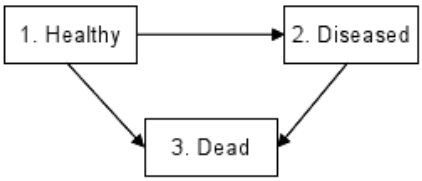

Figure 3.1: Illness-death model: the three states (boxes) and the possible transition among them (arrows).

Let T_{ij} be the potential transition time from state i to state j. This means that a subject not visiting state 2 will reach the 'dead' state at time T_{13}, while this time will be $T_{12} + T_{23}$ if he/she passes through state 2 before. We denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting state 2 at some time. Let $Z = T_{12} \wedge T_{13}$ be the sojourn time in state 1, and let $T = Z + \rho T_{23}$ be the total survival time of the process (up to reaching the absorbing state). We denote the censoring variable by C which is assumed to be independent of the process; finally, we put $\widetilde{Z} = Z \wedge C$ and $\widetilde{T} = T \wedge C$ for the censored versions of Z and T, and $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$ for the respective censoring indicators. With this notation, the transition probabilities are written as

$$
\begin{aligned}
p_{11}(s,t) &= \frac{P\left(Z > t\right)}{P\left(Z > s\right)}, \qquad p_{13}(s,t) = \frac{P\left(s < Z, T \leq t\right)}{P(Z > s)}, \\
p_{23}(s,t) &= \frac{P\left(Z \leq s, s < T \leq t\right)}{P\left(Z \leq s < T\right)}.
\end{aligned}
$$

All these quantities involve expectations of particular transformations of the pair (Z, T),

$S(\varphi) = E[\varphi(Z,T)]$ say. Thus we now discuss how these expectations can be empirically approximated from the data

$$\left\{\left(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i, \Delta_{1i}\rho_i\right), 1 \leq i \leq n\right\},$$

which are assumed to form a random sample of the vector $\left(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta, \Delta_1\rho\right)$. Note that $p_{11}(s,t)$ and the denominator of $p_{13}(s,t)$ only involve the Z variable, and that they can be estimated by the ordinary Kaplan-Meier estimator of the sojourn time distribution in state 1. However, the remaining quantities cannot be estimated so simply.

Let $\widetilde{T}_{1:n} \leq \cdots \leq \widetilde{T}_{n:n}$ denote the ordered $\widetilde{T}_i$'s, and let W_i be the Kaplan-Meier weight attached to $\widetilde{T}_{i:n}$ when estimating the marginal distribution of T from $(\widetilde{T}_i, \Delta_i)$'s. That is,

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

where $\Delta_{[i:n]}$ is the ith concomitant of $\widetilde{T}_{i:n}$. Here, ties within the censored or within the uncensored times are ordered arbitrarily, and ties among the uncensored and censored times are treated as if the former precede the latter.

In the uncensored case we have $W_i = n^{-1}$ for each i. In Meira-Machado et al (2006) the following estimator of $S(\varphi)$ was proposed:

$$S_n(\varphi) = \sum_{i=1}^{n} W_i \varphi(\widetilde{Z}_{[i:n]}, \widetilde{T}_{i:n}).$$

where $\widetilde{Z}_{[i:n]}$ is the concomitant of $\widetilde{T}_{i:n}$. Consider now the presmoothed version of $S_n(\varphi)$ given by

$$S_n(\varphi; m_n) = \sum_{i=1}^{n} W_i(m_n) \varphi(\widetilde{Z}_{[i:n]}, \widetilde{T}_{i:n})$$

where

$$W_i(m_n) = \frac{m_n\left(\widetilde{Z}_{[i:n]}, \widetilde{T}_{i:n}\right)}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m_n\left(\widetilde{Z}_{[j:n]}, \widetilde{T}_{j:n}\right)}{n-j+1}\right]$$

and where $m_n(z,t)$ stands for an estimator of the binary regression function

$$m(z,t) = P\left(\Delta = 1 | \widetilde{Z} = z, \widetilde{T} = t\right).$$

Since $\left(\widetilde{Z}, \widetilde{T}, \Delta\right)$ are observable, the function $m(z,t)$ can be estimated by standard methods. However, the naive construction of a smooth estimator for $m(z,t)$ will generally fail. This is because the function $m(z,t)$ will typically be discontinuous along the line $t = z$, that is, for those covariate values $\left(\widetilde{Z}, \widetilde{T}\right)$ corresponding to individuals who are censored while being in state 1 or who suffer a direct transition $1 \to 3$ to the absorbing state.

In order to see this, note that for $z < t$ we have

$$m(z,t) = P\left(\Delta = 1 | \widetilde{Z} = z, \widetilde{T} = t, \Delta_1\rho = 1\right) \equiv m_1(z,t),$$

that is, $m_1\left(\widetilde{Z}, \widetilde{T}\right)$ is the conditional probability of censoring on T given $\left(\widetilde{Z}, \widetilde{T}\right)$ and given that transition $1 \to 2$ is observed ($\Delta_1\rho = 1$). However, when $z = t$ we get

$$m(t,t) = P\left(\Delta_1 = 1 | \widetilde{Z} = t, \Delta_1\rho = 0\right) \equiv m_2(t),$$

which is the conditional probability of observing $1 \to 3$ given $\widetilde{Z} = t$ (or $\widetilde{T} = t$) and given that transition $1 \to 2$ is never observed. We implicitly assume that the events $\left\{\widetilde{Z} = \widetilde{T}\right\}$ and $\Delta_1\rho = 0$ are the same. This is reasonable unless there is a significative proportion of individuals with zero sojourn time in state 2. These formulae show that the functions m_1 and m_2 represent different binary regression problems and that they are based on disjoint subpopulations (according to the value of $\Delta_1\rho$). Furthermore, the limit of $m_1(z,t)$ as z approaches to t does not coincide with $m_2(t)$ in reality. Figure 3.2 displays these functions for the colon cancer data, when estimated separately by two logistic models. The noise around $m_1(z,t)$ comes from the fact that the variable z is omitted from the plot while it is present in the model (although without reaching statistical significance, p-value=0.285). Both functions are clearly separated.

In summary, in order to construct $m_n(z,t)$ we propose to estimate the functions $m_1(z,t)$ and $m_2(t)$ independently by fitting some smooth models, $m_{1n}(z,t)$ and $m_{2n}(t)$ say, so we finally have

$$m_n(z,t) = m_{1n}(z,t)\, I(z < t) + m_{2n}(t) I(z = t),$$

or

$$\begin{aligned} m_n\left(\widetilde{Z}_i, \widetilde{T}_i\right) &= m_{1n}\left(\widetilde{Z}_i, \widetilde{T}_i\right) I(\widetilde{Z}_i < \widetilde{T}_i) + m_{2n}(\widetilde{Z}_i) I(\widetilde{Z}_i = \widetilde{T}_i) \\ &= m_{1n}\left(\widetilde{Z}_i, \widetilde{T}_i\right) \Delta_{1i}\rho_i + m_{2n}(\widetilde{Z}_i)(1 - \Delta_{1i}\rho_i). \end{aligned}$$

The estimator $m_{1n}(z,t)$ is based on the subsample $\{i : \Delta_{1i}\rho_i = 1\}$, while $m_{2n}(t)$ is computed from $\{i : \Delta_{1i}\rho_i = 0\}$. The only condition we assume on these two functions is that they should approximate well their targets in a uniform sense; more specifically, set

$$U_1 : \sup_{z,t} |m_{1n}(z,t) - m_1(z,t)| \to 0 \qquad \text{w. p. 1},$$

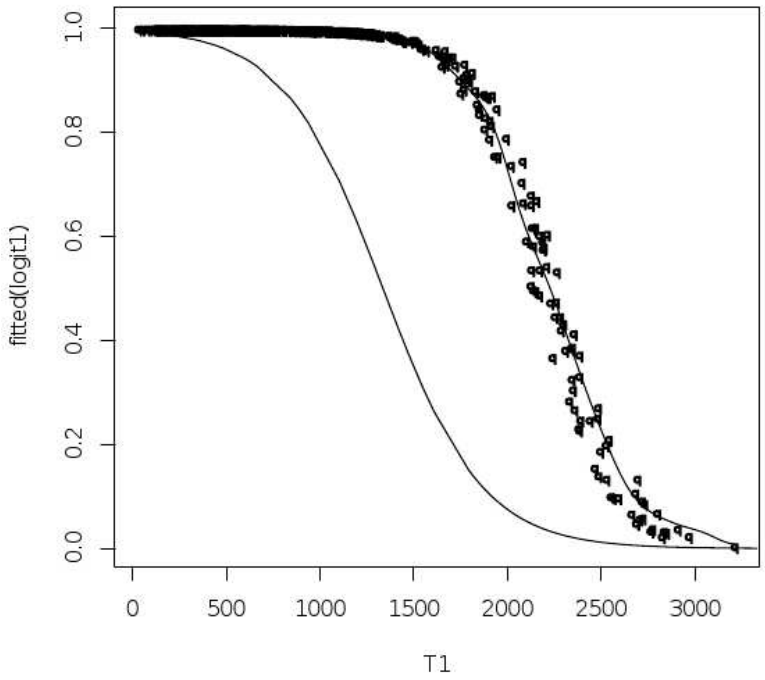

Figure 3.2: Presmoothing functions m_1 and m_2 estimated by logistic models vs. $\widetilde{T}$ (variable $\widetilde{Z}$ not shown). Colon cancer data.

and

$$U_2: \ \sup_t |m_{2n}(t) - m_2(t)| \to 0 \qquad \text{w. p. 1.}$$

Since $m(z,t) = m_1(z,t)\, I(z < t) + m_2(t) I(z = t)$, under U_1 and U_2 we have

$$U: \ \sup_{z,t} |m_n(z,t) - m(z,t)| \to 0 \qquad \text{w. p. 1.}$$

and hence Theorem 2.1 in de Uña-Álvarez and Rodríguez-Campos (2004) can be applied with some adaptation to the present context. Conditions under which U_1 and U_2 hold are investigated in a number of papers, including Devroye (1978a,b), Mack and Silverman (1982) , and Härdle and Luckhaus (1984). See also Dikta (1998) for the parametric setup. Now we state our main result and the corresponding corollaries. Let H be the distribution function of $\widetilde{T}$ and let $\tau_H = \inf\{t : H(t) = 1\}$.

Theorem 3.2.1. *Assume that H is continuous, that U_1 and U_2 hold, and that*

$$E\left[\frac{|\varphi(Z,T)|\, I(T \leq \tau_H)}{m(Z,T)(1-H(T))^{\rho}}\right] < \infty$$

is satisfied for some $\rho > 0$. Then, $S_n(\varphi; m_n) \to S^{\tau}(\varphi)$ with probability 1, where $S^{\tau}(\varphi) = E[\varphi(Z,T)I(T \leq \tau_H)]$.

The Theorem 3.2.1 is a proper adaptation of the Strong Law in Dikta (2000) to our scenario. We note that the result is not restricted to parametric presmoothing; the only thing one should have in mind is that condition U must be verified by the chosen estimator $m_n(z,t)$. Note also that, in general, one can not ensure that $S^{\tau}(\varphi)$ and $S(\varphi)$ will coincide; indeed, as always with censored data, one should not expect consistency beyond the upper bound of the censoring distribution, because there is no sampling information regarding the lifetime there. As a particular case, we have $S^{\tau}(\varphi) = S(\varphi)$ if the support of T is contained in that of C.

The proof to Theorem 3.2.1 is similar to that of Theorem 2.2.1 in Chapter 2, see Section 2.6 for the details. Here, the sojourn time in State 1 (Z) plays the role of 'covariate', while the total time T up to reaching the absorbing State 3 is taken as the 'response'. Due to the existing similarities we do not repeat the details corresponding to the illness-death model.

Now, we come back to our initial goal of estimating the transition probabilities $p_{ij}(s,t)$. Recall that $p_{11}(s,t)$ can be estimated by the ordinary Kaplan-Meier based on the $\left(\widetilde{Z}_i, \Delta_{1i}\right)$'s. In order to introduce some presmoothing, we recommend to replace the Δ_{1i}'s by some smooth fit to the binary regression function $m_0(z) = P\left(\Delta_1 = 1 | \widetilde{Z} = z\right)$. Now, we focus on the estimation of $p_{13}(s,t)$ and $p_{23}(s,t)$. Write

$$p_{13}(s,t) = \frac{P(s < Z, T \leq t)}{P(s < Z)} = \frac{E[\varphi_{s,t}(Z,T)]}{P(s < Z)},$$

where $\varphi_{s,t}(u,v) = I(u > s, v \leq t)$. Introduce the presmoothed estimator

$$\widehat{p}_{13}(s,t) = \frac{S_n(\varphi_{s,t}; m_n)}{\widehat{P}(s < Z)}$$

where $\widehat{P}(s < Z)$ stands for a consistent estimator (e.g. Kaplan-Meier) of $P(s < Z)$.

Similarly, we have

$$p_{23}(s,t) = \frac{P(Z \leq s, s < T \leq t)}{P(Z \leq s < T)} = \frac{E[\widetilde{\varphi}_{s,t}(Z,T)]}{E[\overline{\varphi}_s(Z,T)]}$$

where $\widetilde{\varphi}_{s,t}(u,v) = I(u \leq s, s < v \leq t)$ and $\overline{\varphi}_s(u,v) = I(u \leq s < v)$. Therefore, in this case, we estimate the transition probability through

$$\widehat{p}_{23}(s,t) = \frac{S_n\left(\widetilde{\varphi}_{s,t}; m_n\right)}{S_n\left(\overline{\varphi}_s; m_n\right)}.$$

We have the following Corollary.

Corollary 3.2.1. *Assume that the conditions in Theorem 3.2.1 hold for the special φ-functions $\varphi_{s,t}$, $\widetilde{\varphi}_{s,t}$ and $\overline{\varphi}_s$. Then, for any consistent estimator $\widehat{P}(Z > s)$ of $P(Z > s)$ we have with probability 1 $\widehat{p}_{13}(s,t) \to p^{\tau}_{13}(s,t)$ and $\widehat{p}_{23}(s,t) \to p^{\tau}_{23}(s,t)$, where $p^{\tau}_{13}(s,t) = P\left(T \leq t, T \leq \tau_H \ / \ Z > s\right)$ and $p^{\tau}_{23}(s,t) = P(T \leq t \ / \ Z \leq s < T, T \leq \tau_H)$.* □

Corollary 3.2.1 is an immediate consequence of Theorem 3.2.1. As for the Theorem, consistency can not be ensured in general. However, when the support of C contains that of T we have $p^{\tau}_{13}(s,t) = p_{13}(s,t)$ and $p^{\tau}_{23}(s,t) = p_{23}(s,t)$. In particular this will happen whenever $\tau_H = \infty$.

Remark In practice, when n is small, it may happen $\widehat{p}_{13}(s,t) > 1$ and/or $\widehat{p}_{11}(s,t) + \widehat{p}_{13}(s,t) > 1$. When any of these inequalities occurs, we propose the modification $\widehat{p}_{13}(s,t) = 1 - \widehat{p}_{11}(s,t)$, which ensures $\widehat{p}_{13}(s,t) \leq 1$ and $\widehat{p}_{11}(s,t) + \widehat{p}_{13}(s,t) \leq 1$. With this remark in mind, we always have $\widehat{p}_{12}(s,t) = 1 - \widehat{p}_{11}(s,t) - \widehat{p}_{13}(s,t) \geq 0$. For moderate or large sample sizes this problem disappears.

3.3 Simulation study

In this Section we investigate the performance of the proposed estimators $\widehat{p}_{ij}\left(s,t\right)$ through simulations. More specifically, the estimators $\widehat{p}_{11}\left(s,t\right)$, $\widehat{p}_{13}\left(s,t\right)$ and $\widehat{p}_{23}\left(s,t\right)$ introduced in Section 3.2 are considered.

To simulate the data in the illness-death model, we separately consider the subjects passing through State 2 at some time (that is, those cases with $\rho = 1$), and those who directly go to the absorbing State 3 ($\rho = 0$). For the first subgroup of individuals ($\rho = 1$), the successive gap times $(Z, T - Z)$ are simulated according to the bivariate distribution

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

where the marginal distribution functions F_1 and F_2 are exponential with rate parameter 1. This corresponds to the so-called Farlie-Gumbel-Morgenstern copula, where the single parameter θ controls for the amount of dependency between the gap times. The parameter θ was set to 0 for simulating independent gap times, and also to 1, corresponding to 0.25 correlation between Z and $T - Z$. This simulated scenario is the same as that described in Section 2.3. For the second subgroup of individuals ($\rho = 0$), the value of Z is simulated according to an exponential with rate parameter 1. In summary, the simulation procedure is as follows:

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

Step 2. If $\rho = 1$ then:

(2.1) $V_1 \sim U(0,1), V_2 \sim U(0,1)$ are independently generated;

(2.2) $U_1 = V_1, A = \theta(2U_1 - 1) - 1, B = (1 - \theta(2U_1 - 1))^2 + 4\theta V_2(2U_1 - 1)$

(2.3) $U_2 = 2V_2 / \left(\sqrt{B} - A\right)$

(2.4) $Z = \ln(1/(1 - U_1)), T = \ln(1/(1 - U_2)) + Z$

If $\rho = 0$ then $Z = \ln(1/(1 - U(0,1)))$.

Situations with $p = 1$ corresponds to the three-state progressive model, in which a direct transition $1 \to 3$ is not allowed. In our simulation we consider $p = 0.7$. An independent uniform censoring time C is generated, according to models $U[0,4]$ and $U[0,3]$. The first model results in 24% of censoring on the first gap time Z, and in 47% of censoring on the second gap time $T - Z$, for those individuals with $\rho = 1$. The second model increases these censoring levels to 32% and about 57%, respectively.

After some algebra, it is seen that the function

$$m_1(z,t) = P\left(\Delta = 1 | \widetilde{Z} = z, \widetilde{T} = t, \Delta_1 \rho = 1\right)$$

is written as

$$m_1(z,t) = \frac{1}{1 + \eta_1(z,t)}, \qquad \text{where } \eta_1(z,t) = \frac{\lambda_G(t)}{\lambda^1_{T|Z=z}(t|z)}$$

and where $\lambda_G(.)$ and $\lambda^1_{T|Z=z}(.|z)$ stand respectively for the hazard rate of the censoring variable and the hazard rate of T given $Z = z$ under restriction $\rho = 1$. Note that $\lambda_G(t) = 1/(\tau_G - t)$ when $C \sim U[0, \tau_G]$ and that $\lambda^1_{T|Z=z}(t|z)$ is given by

$$\lambda^1_{T|Z=z}(t|z) = \frac{2 + 4\exp(-t) - 2\exp(-z) - 2\exp(-t+z)}{2 + 2\exp(-t) - 2\exp(-z) - \exp(-t+z)} \qquad \text{if } \theta = 1,$$

being 1 when $\theta = 0$. The function $m_1(z,t)$ belongs to the logistic family with some preliminary transformation of the conditioning variables. To be more specific we have (for $\beta_0 = 0$ and $\beta_1 = 1$)

$$m_1(z,t;\beta) = \frac{1}{1 + \exp(\beta_0 + \beta_1 \ln(\eta_1(z,t)))}.$$

This is the parametric model we fit to $m_1(z,t)$ in the simulations. The β parameter in model $m_1(.;\beta)$ is estimated via maximization of the conditional likelihood of the Δ_i's given the

$(\widetilde{Z}_i, \widetilde{T}_i)$'s, for those subjects with $\Delta_1\rho = 1$ (see e.g. Dikta, 1998, 2000). The same estimation criterium is used for the other presmoothing functions (m_0 and m_2) in this section. For $m_2(t) = P\left(\Delta_1 = 1|\widetilde{Z} = t, \Delta_1\rho = 0\right)$, we have

$$m_2(t) = \frac{1}{1+\eta_2(t)}, \qquad \text{where } \eta_2(t) = \frac{\lambda_G(t)}{\lambda_Z^0(t)}$$

and where $\lambda_Z^0(t)$ stands for the sub-hazard function of Z restricted to $\rho = 0$, namely

$$\lambda_Z^0(t) = P\left(Z = t, \rho = 0|Z \geq t\right) = 1 - p.$$

Similarly as above, we fit the logistic model

$$m_2(t;\gamma) = \frac{1}{1+\exp(\gamma_0 + \gamma_1 \ln(\eta_2(t)))}$$

to estimate the function $m_2(t)$ in the simulations. As before, this logistic model has the true presmoothing function m_2 as a special case ($\gamma_0 = 0$,$\gamma_1 = 1$).

The aim of this simulation study is to compare the estimator by Meira-Machado et al. (2006) and the new estimator based on presmoothing ideas. In order to measure the estimates' relative performance, we computed the integrated absolute bias, integrated variance and the integrated Mean Square Error (MSE) of the estimates. For each simulated setting we derived the analytic expression of $p_{11}(s,t)$, $p_{13}(s,t)$ and $p_{23}(s,t)$ so that the bias and the MSE of the estimator could be examined. Sample sizes 50, 100 and 200 were considered. In each simulation, $K = 1000$ samples were generated.

Let $\hat{p}_{ij}^k(s,t)$ denote the estimated transition probability based on the kth generated data set. For each fixed (s,t) we obtained the mean for all generated data sets, $\overline{\hat{p}_{ij}(s,t)} = \frac{1}{K}\sum_{k=1}^K \hat{p}_{ij}^k(s,t)$. We then computed the pointwise estimates of the bias, variance and MSE as:

$$\widehat{bias}(s,t) = p_{ij}(s,t) - \overline{\hat{p}_{ij}(s,t)}$$

$$\widehat{var}(\hat{p}_{ij}(s,t)) = \frac{1}{K-1}\sum_{k=1}^K [\hat{p}_{ij}^k(s,t) - \overline{\hat{p}_{ij}(s,t)}]^2$$

$$\widehat{MSE}(\hat{p}_{ij}(s,t)) = \frac{1}{K}\sum_{k=1}^K [\hat{p}_{ij}^k(s,t) - p_{ij}(s,t)]^2$$

To summarize the results we also calculated the integrated absolute bias, integrated variance and the integrated MSE, defined in Table 3.1 . We fixed the values of s using the quantiles 0.25, 0.5 and 0.75 of the exponential distribution with rate 1. The results obtained in Table 3.2 and

Statistic	Definition	Estimator
Integrated absolute bias	$\int_s^{t_1} \lvert bias(s,t)\rvert dt$	$\sum_{t=s}^{t_1} \lvert\widehat{bias}(s,t)\rvert\delta$
Integrated variance	$\int_s^{t_1} var(\hat{p}_{ij}(s,t))dt$	$\sum_{t=s}^{t_1} \widehat{var}(\hat{p}_{ij}(s,t))\delta$
Integrated MSE	$\int_s^{t_1} MSE(\hat{p}_{ij}(s,t))dt$	$\sum_{t=s}^{t_1} \widehat{MSE}(\hat{p}_{ij}(s,t))\delta$

Table 3.1: Summary statistics measuring bias, variance and mean square error.

3.3 were obtained by numerical integration on the interval $[s,t_1]$ with $t_1 = 3$, taking a grid of step $\delta = 0.05$.

In Tables 3.2 to 3.5 we report the results for the integrated absolute bias, integrated variance and the integrated MSE attained by the proposed estimators for $p_{ij}(s,t)$ when based on several presmoothing functions. The row labeled m corresponds to presmoothing with the true presmoothing function, which is unrealistic because this function will generally be unknown. However, as in Chapter 2 this row represents a 'gold standard' the other methods can be compared to. The row labeled with $m(.;\beta,\gamma)$ corresponds to a semiparametric estimator which is obtained using a presmoothing based on a parametric family which contains the true m. Specifically, we consider a logistic model with the preliminary transformation of the conditioning variables $\tilde{Z} = z, \tilde{T} = t$ shown before. Similarly, for $p_{11}(s,t)$ and for the denominator of $p_{13}(s,t)$ we also perform logistic presmoothing for the function $m_0(z) = P\left(\Delta_1 = 1 | \widetilde{Z} = z\right)$, with the variable $\widetilde{Z}$ transformed by $-\ln(\tau_G - \widetilde{Z})$ (so the parametric family contains the true $m_0(z)$).

In order to investigate the robustness of the proposed estimator with respect to miss-specifications of the binary regression family, we considered also presmoothing via standard logistic models, without any preliminary transformation of the transition times. This is labeled with $m(.,\xi)$ in Tables. Note that the true m and the true m_0 do not belong to this parametric family. Finally, we also report the results pertaining to the estimator in Meira-Machado et al. (2006), which corresponds to the situation with no presmoothing at all. This is labeled in the Tables as KM.

Some expected features are clearly seen in Tables. For example, we see that the (integrated) MSE, bias and variance of $\hat{p}_{ij}(s,t)$ decrease with an increasing sample size, while they increase with the censoring degree. The best performance is attained by the estimator which makes use of the true m (resp. m_0), which was expected. However, in practice one must estimate the function m. The lowest errors among the realistic versions of the estimators correspond to the estimator based on the correctly specified parametric family, $m(.;\beta,\gamma)$. Finally, we see that the presmoothed estimator based on the wrong parametric model $m(.;\xi)$ is still (much) better than KM; as in Chapter 2 the practical message is that it is worthwhile doing some presmoothing even when we are not completely sure about the parametric family.

Compared to the estimator without any presmoothing (KM), in the simulations the relative efficiency of the estimators based on presmoothing is always above 1. In special cases, the relative

deficiency of the Kaplan-Meier estimator is below 50%; this occurs for larger values of s, where the censoring effects are stronger. This supports the belief that the relative benefits of presmoothing will be seen more clearly in the presence of large censoring degrees.

Although we restrict the integrated bias, variance and mean square error (MSE) to the interval $[s, 3]$, we verified that, in both settings, the enlargement of this interval favors the estimator based on presmoothing. This happens because higher levels of censoring are expected at the right tail of the distribution.

3.4 Real data illustration

For illustration, we apply the proposed methods of Section 3.2 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 (Moertel et al. (1990)). In this study, from the total of 929 patients, 468 developed 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. As mentioned in the Introduction recurrence can be expressed as an intermediate event which can be modeled using an illness-death model.

Using the Cox proportional hazards model, we verified that the transition rate from state 2 to state 3 is affected by the time spent in the previous state (Kay (1986)). This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set. In this section we will present estimated transition probabilities calculated using the new approach, based on presmoothed Kaplan-Meier weights and the estimators of Meira-Machado et al. (2006). Neither one of the approaches assume the process as being Markovian.

In Figure 3.3 we illustrate differences between the estimated transition probabilities $p_{ij}(s,t)$, $1 \leq i \leq j \leq 3$ based on presmoothing the Kaplan-Meier weights and the estimator corresponding to no presmoothing (KM; Meira-Machado et al. (2006)). The presmoothed estimator was obtained by standard logistic regression for both m_1 and m_2. The value s was chosen to be the 75th percentile of the sojourn time in state 1 ($s = 1549$ days). From this figure we conclude that the new estimator have more jump points (corresponding to patients with censored values of the total time) but with smaller steps. The number of jump points and the size of the steps are related to the censoring degree and to the sample size. We can also verify that both methods provide similar point estimates for small time values. Departures between both estimated curves can be more appreciated for larger time values where the censoring effects are stronger. In summary, the new approach provides more reliable curves with less variability, specially at the right tail of the lifetime distribution. Other values of s reported similar results.

Table 3.2: Integrated absolute bias, integrated variance and the integrated MSE of $\hat{p}_{ij}(s,.)$ along 1,000 trials, case $\theta = 1$ and $C \backsim U[0,4]$.

	n	50			100			200		
$P_{ij}(s,t)$	$Method$	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR
$P_{11}(0.2877,t)$	$m(.;\beta,\gamma)$	0.01492	0.00907	0.01486	0.00718	0.00918	0.00715	0.00359	0.00287	0.00359
	$m(.;\xi)$	0.01537	0.01339	0.01527	0.00756	0.01601	0.00745	0.00380	0.00987	0.00376
	KM	0.01749	0.00426	0.01750	0.00855	0.00595	0.00854	0.00418	0.00118	0.00418
	m	0.01175	0.00361	0.01175	0.00609	0.00418	0.00609	0.00304	0.00205	0.00304
$P_{13}(0.2877,t)$	$m(.;\beta,\gamma)$	0.01917	0.01996	0.01898	0.00904	0.01302	0.00897	0.00440	0.00392	0.00440
	$m(.;\xi)$	0.01959	0.01866	0.01942	0.00928	0.01291	0.00921	0.00451	0.00474	0.00450
	KM	0.02318	0.00925	0.02317	0.01186	0.00519	0.01186	0.00565	0.00164	0.00566
	m	0.01273	0.00317	0.01273	0.00662	0.00300	0.00662	0.00314	0.00315	0.00314
$P_{23}(0.2877,t)$	$m(.;\beta,\gamma)$	0.11278	0.33928	0.06426	0.08010	0.33575	0.03299	0.06054	0.32289	0.01665
	$m(.;\xi)$	0.11800	0.34077	0.06895	0.08311	0.33317	0.03632	0.06179	0.31516	0.01950
	KM	0.13334	0.35142	0.08173	0.09356	0.34754	0.04299	0.06818	0.32733	0.02322
	m	0.11267	0.33787	0.06466	0.07973	0.33510	0.03271	0.06017	0.32254	0.01636
$P_{11}(0.6931,t)$	$m(.;\beta,\gamma)$	0.02315	0.01308	0.02300	0.01082	0.01042	0.01076	0.00529	0.00339	0.00528
	$m(.;\xi)$	0.02419	0.01967	0.02395	0.01167	0.02104	0.01145	0.00574	0.01397	0.00565
	KM	0.02858	0.00531	0.02859	0.01332	0.00635	0.01332	0.00639	0.00118	0.00640
	m	0.01824	0.00457	0.01824	0.00933	0.00371	0.00933	0.00463	0.00130	0.00464
$P_{13}(0.6931,t)$	$m(.;\beta,\gamma)$	0.02720	0.02753	0.02677	0.01133	0.01821	0.01116	0.00537	0.00611	0.00535
	$m(.;\xi)$	0.02908	0.02946	0.02856	0.01240	0.02462	0.01205	0.00572	0.01202	0.00561
	KM	0.03709	0.01125	0.03706	0.01601	0.00949	0.01598	0.00789	0.00120	0.00789
	m	0.01652	0.00575	0.01651	0.00809	0.00299	0.00810	0.00394	0.00147	0.00395
$P_{23}(0.6931,t)$	$m(.;\beta,\gamma)$	0.07035	0.21315	0.04952	0.04205	0.19120	0.02526	0.02784	0.17532	0.01370
	$m(.;\xi)$	0.07708	0.20687	0.05734	0.04515	0.17437	0.03111	0.02845	0.14607	0.01846
	KM	0.09932	0.23282	0.07454	0.05878	0.19997	0.04046	0.03876	0.18576	0.02297
	m	0.06874	0.20093	0.05025	0.04073	0.18794	0.02452	0.02698	0.17190	0.01337
$P_{11}(1.3863,t)$	$m(.;\beta,\gamma)$	0.04543	0.01990	0.04498	0.02034	0.01166	0.02022	0.00980	0.00557	0.00979
	$m(.;\xi)$	0.04639	0.02281	0.04590	0.02166	0.01967	0.02139	0.01059	0.01478	0.01045
	KM	0.06285	0.00915	0.06283	0.02668	0.00571	0.02668	0.01307	0.00273	0.01308
	m	0.03552	0.00895	0.03547	0.01742	0.00296	0.01743	0.00878	0.00156	0.00879
$P_{13}(1.3863,t)$	$m(.;\beta,\gamma)$	0.03601	0.01591	0.03581	0.01446	0.01751	0.01421	0.00673	0.00772	0.00669
	$m(.;\xi)$	0.03778	0.01910	0.03746	0.01545	0.02518	0.01490	0.00718	0.01700	0.00694
	KM	0.05982	0.00558	0.05985	0.02463	0.00834	0.02457	0.01180	0.00177	0.01181
	m	0.02015	0.00429	0.02015	0.00912	0.00414	0.00911	0.00476	0.00089	0.00477
$P_{23}(1.3863,t)$	$m(.;\beta,\gamma)$	0.06433	0.11788	0.05405	0.03299	0.09703	0.02598	0.01696	0.07198	0.01304
	$m(.;\xi)$	0.07135	0.10612	0.06325	0.03598	0.07266	0.03221	0.01724	0.03337	0.01645
	KM	0.10830	0.15046	0.09176	0.05788	0.11084	0.04867	0.03064	0.07963	0.02579
	m	0.06038	0.09580	0.05336	0.03057	0.08890	0.02462	0.01590	0.06653	0.01252

Table 3.3: Integrated absolute bias, integrated variance and the integrated MSE of $\hat{p}_{ij}(s,.)$ along 1,000 trials, case $\theta = 1$ and $C \backsim U[0,3]$.

	n	50			100			200		
$P_{ij}(s,t)$	$Method$	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR
$P_{11}(0.2877,t)$	$m(.;\beta,\gamma)$	0.02832	0.09375	0.02331	0.01531	0.07357	0.01202	0.00764	0.04701	0.00607
	$m(.;\xi)$	0.02573	0.07579	0.02249	0.01273	0.05904	0.01113	0.00586	0.03274	0.00536
	KM	0.03225	0.06921	0.02948	0.01790	0.05395	0.01566	0.00969	0.03988	0.00822
	m	0.01498	0.11368	0.01059	0.00948	0.11017	0.00550	0.00738	0.12113	0.00280
$P_{13}(0.2877,t)$	$m(.;\beta,\gamma)$	0.04984	0.10192	0.04249	0.02472	0.08005	0.02082	0.01185	0.05872	0.01006
	$m(.;\xi)$	0.05045	0.09333	0.04335	0.02468	0.07204	0.02096	0.01185	0.04932	0.01026
	KM	0.06527	0.11023	0.05562	0.03778	0.08992	0.02980	0.02187	0.07670	0.01540
	m	0.03516	0.24886	0.01407	0.03131	0.26380	0.00736	0.03132	0.28168	0.00370
$P_{23}(0.2877,t)$	$m(.;\beta,\gamma)$	0.12218	0.40839	0.06439	0.09002	0.39064	0.03768	0.06763	0.36926	0.02120
	$m(.;\xi)$	0.12935	0.43071	0.06386	0.09588	0.40568	0.03868	0.07262	0.38376	0.02171
	KM	0.15415	0.45271	0.08112	0.11203	0.43394	0.04620	0.08408	0.41669	0.02407
	m	0.12240	0.41434	0.06244	0.09067	0.40199	0.03473	0.06909	0.38448	0.01825
$P_{11}(0.6931,t)$	$m(.;\beta,\gamma)$	0.05402	0.13757	0.04220	0.02905	0.10863	0.02159	0.01431	0.07022	0.01061
	$m(.;\xi)$	0.04761	0.11442	0.03996	0.02293	0.08387	0.01943	0.01019	0.04623	0.00911
	KM	0.06148	0.09688	0.05507	0.03424	0.08026	0.02908	0.01870	0.05956	0.01534
	m	0.02296	0.12529	0.01691	0.01515	0.13079	0.00869	0.01218	0.14753	0.00442
$P_{13}(0.6931,t)$	$m(.;\beta,\gamma)$	0.08327	0.15366	0.06796	0.04161	0.11016	0.03376	0.02022	0.07845	0.01666
	$m(.;\xi)$	0.08753	0.14999	0.07193	0.04419	0.11328	0.03544	0.02130	0.07471	0.01762
	KM	0.11946	0.14723	0.10037	0.06925	0.12959	0.05311	0.04044	0.10629	0.02795
	m	0.05162	0.27132	0.02238	0.04906	0.30923	0.01160	0.05071	0.33635	0.00603
$P_{23}(0.6931,t)$	$m(.;\beta,\gamma)$	0.09040	0.32719	0.05399	0.05770	0.28552	0.03070	0.04003	0.25597	0.01879
	$m(.;\xi)$	0.10076	0.35268	0.05780	0.06438	0.29891	0.03461	0.04519	0.26506	0.02230
	KM	0.13399	0.38893	0.08022	0.08955	0.35406	0.04595	0.06444	0.33193	0.02684
	m	0.08702	0.32171	0.05196	0.05645	0.29396	0.02767	0.04075	0.27371	0.01620
$P_{11}(1.3863,t)$	$m(.;\beta,\gamma)$	0.14605	0.25396	0.10265	0.08101	0.19992	0.05367	0.04209	0.13535	0.02790
	$m(.;\xi)$	0.11180	0.18588	0.08698	0.05251	0.11868	0.04214	0.02381	0.06612	0.02065
	KM	0.19191	0.19282	0.16447	0.10504	0.15280	0.08499	0.05940	0.11772	0.04605
	m	0.04883	0.14550	0.03879	0.03121	0.16072	0.01797	0.02765	0.19280	0.00967
$P_{13}(1.3863,t)$	$m(.;\beta,\gamma)$	0.13218	0.13472	0.11114	0.07285	0.11363	0.06033	0.04016	0.08880	0.03354
	$m(.;\xi)$	0.13648	0.13063	0.11627	0.07760	0.11575	0.06395	0.04159	0.08713	0.03489
	KM	0.26640	0.20827	0.22730	0.15874	0.18311	0.12502	0.09843	0.15913	0.07022
	m	0.08307	0.27007	0.04429	0.08372	0.34069	0.02452	0.08976	0.38495	0.01436
$P_{23}(1.3863,t)$	$m(.;\beta,\gamma)$	0.11456	0.35334	0.06039	0.07326	0.29224	0.03377	0.04903	0.24384	0.01876
	$m(.;\xi)$	0.12753	0.37918	0.06622	0.08083	0.31168	0.03800	0.05308	0.25858	0.02137
	KM	0.19063	0.45280	0.09973	0.14137	0.41735	0.06497	0.10093	0.37783	0.03788
	m	0.10898	0.33798	0.05898	0.07232	0.29907	0.03151	0.05035	0.26725	0.01641

Table 3.4: Integrated absolute bias, integrated variance and the integrated MSE of $\hat{p}_{ij}(s,.)$ along 1,000 trials, case $\theta = 0$ and $C \backsim U[0,4]$.

	n	50			100			200		
$P_{ij}(s,t)$	$Method$	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR
$P_{11}(0.2877,t)$	$m(.;\beta,\gamma)$	0.01532	0.11222	0.01074	0.00894	0.09896	0.00520	0.00600	0.09067	0.00258
	$m(.;\xi)$	0.01534	0.11245	0.01073	0.00901	0.09912	0.00522	0.00609	0.09085	0.00260
	km	0.01668	0.09741	0.01303	0.00975	0.09246	0.00635	0.00643	0.08632	0.00321
	m	0.00994	0.07850	0.00756	0.00583	0.06953	0.00383	0.00393	0.06676	0.00196
$P_{13}(0.2877,t)$	$m(.;\beta,\gamma)$	0.11164	0.50138	0.01592	0.10178	0.49699	0.00766	0.09800	0.49615	0.00367
	$m(.;\xi)$	0.11214	0.50025	0.01678	0.10253	0.49819	0.00808	0.09917	0.49924	0.00392
	km	0.11684	0.49578	0.02018	0.10625	0.49499	0.01013	0.10206	0.49695	0.00488
	m	0.11227	0.51111	0.01182	0.10607	0.51014	0.00617	0.10342	0.51071	0.00288
$P_{23}(0.2877,t)$	$m(.;\beta,\gamma)$	0.05968	0.04123	0.05885	0.03289	0.02444	0.03266	0.01744	0.02777	0.01708
	$m(.;\xi)$	0.06164	0.04577	0.06078	0.03319	0.01932	0.03305	0.01844	0.02994	0.01804
	km	0.07459	0.12855	0.06830	0.04104	0.08767	0.03805	0.02214	0.06851	0.02031
	m	0.05757	0.03180	0.05709	0.03150	0.03474	0.03101	0.01792	0.04589	0.01694
$P_{11}(0.6931,t)$	$m(.;\beta,\gamma)$	0.02415	0.13026	0.01757	0.01305	0.10973	0.00828	0.00802	0.09710	0.00404
	$m(.;\xi)$	0.02422	0.13088	0.01757	0.01320	0.11021	0.00834	0.00818	0.09753	0.00409
	km	0.02675	0.11019	0.02192	0.01480	0.10125	0.01061	0.00888	0.09090	0.00523
	m	0.01470	0.08832	0.01161	0.00834	0.07453	0.00601	0.00532	0.07030	0.00311
$P_{13}(0.6931,t)$	$m(.;\beta,\gamma)$	0.14120	0.54445	0.02318	0.12545	0.53671	0.01080	0.11941	0.53474	0.00507
	$m(.;\xi)$	0.14353	0.54345	0.02577	0.12778	0.53904	0.01210	0.12208	0.53928	0.00584
	km	0.14938	0.54047	0.03169	0.13250	0.54116	0.01496	0.12566	0.54170	0.00738
	m	0.12932	0.53037	0.01596	0.11889	0.52430	0.00836	0.11520	0.52472	0.00395
$P_{23}(0.6931,t)$	$m(.;\beta,\gamma)$	0.03220	0.02955	0.03162	0.01789	0.02749	0.01756	0.00953	0.02983	0.00909
	$m(.;\xi)$	0.03381	0.02671	0.03336	0.01847	0.02407	0.01823	0.01029	0.03348	0.00974
	km	0.04463	0.11132	0.03956	0.02526	0.08104	0.02248	0.01384	0.07224	0.01164
	m	0.02980	0.03232	0.02930	0.01660	0.03851	0.01592	0.00974	0.04687	0.00864
$P_{11}(1.3863,t)$	$m(.;\beta,\gamma)$	0.05255	0.16692	0.03954	0.02623	0.13456	0.01811	0.01434	0.10954	0.00888
	$m(.;\xi)$	0.05173	0.16619	0.03895	0.02609	0.13377	0.01809	0.01439	0.10882	0.00892
	km	0.06081	0.13203	0.05300	0.03098	0.11948	0.02462	0.01685	0.10075	0.01210
	m	0.02955	0.10694	0.02437	0.01564	0.08591	0.01234	0.00940	0.07645	0.00665
$P_{13}(1.3863,t)$	$m(.;\beta,\gamma)$	0.18084	0.55608	0.03966	0.16394	0.56344	0.01883	0.14948	0.55487	0.00852
	$m(.;\xi)$	0.18756	0.55879	0.04488	0.16901	0.56781	0.02152	0.15401	0.56004	0.01029
	km	0.20325	0.57184	0.05432	0.17905	0.57654	0.02785	0.16143	0.57054	0.01308
	m	0.15013	0.52142	0.02573	0.13825	0.52375	0.01284	0.12947	0.51863	0.00608
$P_{23}(1.3863,t)$	$m(.;\beta,\gamma)$	0.02775	0.03867	0.02671	0.01444	0.02518	0.01404	0.00754	0.02475	0.00720
	$m(.;\xi)$	0.02864	0.03212	0.02791	0.01506	0.02148	0.01481	0.00801	0.03064	0.00749
	km	0.04298	0.12011	0.03606	0.02274	0.09149	0.01859	0.01271	0.07901	0.00958
	m	0.02496	0.03210	0.02428	0.01309	0.03088	0.01260	0.00735	0.03739	0.00656

Table 3.5: Integrated absolute bias, integrated variance and the integrated MSE of $\hat{p}_{ij}(s,.)$ along 1,000 trials, case $\theta = 0$ and $C \backsim U[0,3]$.

	n	50			100			200		
$P_{ij}(s,t)$	$Method$	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR	MSE	$BIAS$	VAR
$P_{11}(0.2877,t)$	$m(.;\beta,\gamma)$	0.02533	0.17916	0.01379	0.01538	0.15282	0.00685	0.01091	0.14005	0.00342
	$m(.;\xi)$	0.02506	0.17822	0.01361	0.01519	0.15098	0.00675	0.01084	0.13762	0.00337
	km	0.02680	0.15533	0.01802	0.01621	0.13952	0.00891	0.01155	0.13276	0.00465
	m	0.01286	0.11926	0.00768	0.00804	0.10377	0.00395	0.00589	0.09757	0.00208
$P_{13}(0.2877,t)$	$m(.;\beta,\gamma)$	0.12879	0.53879	0.01965	0.11432	0.52547	0.00984	0.10744	0.51743	0.00459
	$m(.;\xi)$	0.12985	0.53560	0.02123	0.11531	0.52441	0.01039	0.10943	0.51938	0.00493
	km	0.13043	0.51819	0.02738	0.11441	0.51102	0.01357	0.10954	0.51502	0.00671
	m	0.11430	0.51893	0.01113	0.10729	0.51199	0.00616	0.10369	0.50750	0.00301
$P_{23}(0.2877,t)$	$m(.;\beta,\gamma)$	0.06438	0.10146	0.05970	0.03678	0.06041	0.03448	0.02206	0.05091	0.02057
	$m(.;\xi)$	0.06821	0.09515	0.06430	0.03825	0.04967	0.03645	0.02342	0.04493	0.02228
	km	0.08730	0.20600	0.07115	0.05159	0.16975	0.04040	0.03379	0.14987	0.02501
	m	0.06075	0.08568	0.05692	0.03329	0.06214	0.03092	0.02085	0.05535	0.01923
$P_{11}(0.6931,t)$	$m(.;\beta,\gamma)$	0.04093	0.21343	0.02311	0.02289	0.17416	0.01108	0.01487	0.15437	0.00543
	$m(.;\xi)$	0.04034	0.21229	0.02279	0.02244	0.17173	0.01090	0.01463	0.15108	0.00532
	km	0.04360	0.18053	0.03095	0.02486	0.15663	0.01520	0.01620	0.14432	0.00780
	m	0.01926	0.13991	0.01166	0.01151	0.11558	0.00623	0.00787	0.10579	0.00331
$P_{13}(0.6931,t)$	$m(.;\beta,\gamma)$	0.17470	0.60555	0.02940	0.15028	0.58540	0.01418	0.13851	0.57441	0.00643
	$m(.;\xi)$	0.17729	0.59965	0.03419	0.15242	0.58370	0.01628	0.14217	0.57751	0.00760
	km	0.17720	0.57943	0.04401	0.15180	0.57299	0.02154	0.14260	0.57638	0.01014
	m	0.13713	0.55184	0.01509	0.12515	0.53886	0.00836	0.11936	0.53234	0.00412
$P_{23}(0.6931,t)$	$m(.;\beta,\gamma)$	0.03916	0.09798	0.03405	0.02343	0.06271	0.02065	0.01364	0.05524	0.01164
	$m(.;\xi)$	0.04079	0.08802	0.03666	0.02354	0.05280	0.02142	0.01384	0.05188	0.01218
	km	0.06005	0.20827	0.04219	0.03793	0.17421	0.02511	0.02706	0.16486	0.01558
	m	0.03454	0.08476	0.03005	0.02024	0.06607	0.01739	0.01255	0.05950	0.01039
$P_{11}(1.3863,t)$	$m(.;\beta,\gamma)$	0.08796	0.27925	0.05076	0.04771	0.22165	0.02511	0.02791	0.18594	0.01237
	$m(.;\xi)$	0.08334	0.27252	0.04843	0.04460	0.21323	0.02400	0.02579	0.17640	0.01175
	km	0.09636	0.22418	0.07311	0.05438	0.19447	0.03730	0.03196	0.17021	0.01894
	m	0.03982	0.18166	0.02424	0.02191	0.14304	0.01263	0.01412	0.12523	0.00707
$P_{13}(1.3863,t)$	$m(.;\beta,\gamma)$	0.23182	0.63266	0.04758	0.20590	0.62815	0.02404	0.18433	0.61352	0.01099
	$m(.;\xi)$	0.23549	0.62112	0.05819	0.20700	0.62166	0.02873	0.18685	0.61176	0.01394
	km	0.25719	0.62767	0.07769	0.21916	0.62794	0.03947	0.19605	0.62309	0.01894
	m	0.16726	0.56019	0.02361	0.15240	0.55236	0.01247	0.14050	0.54015	0.00628
$P_{23}(1.3863,t)$	$m(.;\beta,\gamma)$	0.03802	0.12475	0.02882	0.02197	0.08587	0.01642	0.01274	0.06413	0.00910
	$m(.;\xi)$	0.03794	0.10625	0.03080	0.02091	0.06663	0.01691	0.01174	0.05618	0.00908
	km	0.06668	0.23691	0.03966	0.04242	0.20345	0.02200	0.03008	0.18463	0.01313
	m	0.03332	0.11220	0.02503	0.01915	0.08305	0.01366	0.01182	0.06660	0.00802

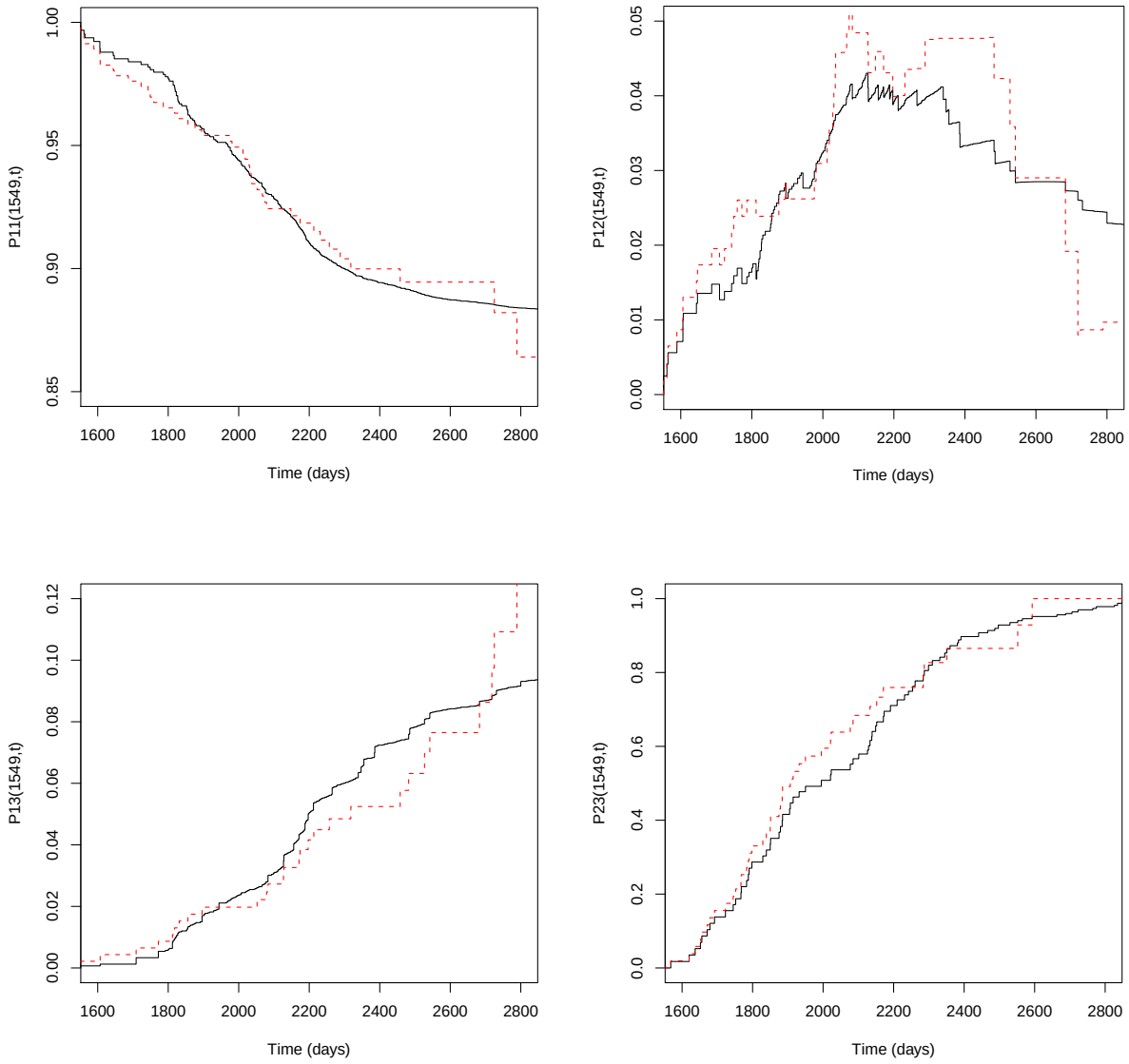

Figure 3.3: Estimated transition probabilities for $P_{ij}(s,t)$ with s=1549 based on the Kaplan-Meier weights (dashed line) and based on presmoothed Kaplan-Meier weights (solid line). Colon cancer data.

Chapter 4

R code and further examples

Contents

4.1 Introduction 72
4.2 R code for the simulations in Section 2.3 72
4.3 A simple example in the illness-death model 89
4.4 Leukaemia data 102

4.1 Introduction

In this Chapter we provide the R code used for the simulation study in Chapter 2. This includes R code for the computation of mean squared errors of several estimators of the gap times joint distribution function, along a number of Monte Carlo trials. The list of estimators includes the new semiparametric estimator, its 'gold standard' (which is based in the true presmoothing function), and the Kaplan-Meier based estimator of de Uña-Álvarez and Meira-Machado (2008). Besides, the R code needed for the study of the performance of the bootstrap estimator of the estimator's standard error is also provided. All this information is contained in Section 4.2.

We also provide in Section 4.3 a simple example of the computation of the presmoothed transition probabilities in the illness-death model, as defined in Chapter 3. To this end, we simulate one trial of hypothetical data coming from one of the models described in Section 3.3, and we compute the presmoothed estimators of the transition probabilities $p_{11}(s,t)$ and $p_{23}(s,t)$ for specific pairs (s,t). For comparison, the Kaplan-Meier-based estimators in Meira-Machado et al (2006) are also evaluated.

Finally, in Section 4.4 we analyze the leukaemia data introduced in Section 1.2. For this data, we give a number of plots representing the transition probabilities when estimated with the ordinary Kaplan-Meier weights and also via logistic presmoothing. A number of comments are provided too.

4.2 R code for the simulations in Section 2.3

In the following, we give the R code for the computation of the mean squared errors in Tables 2.1 (independent gap times) and 2.2 (dependent gap times) in Section 2.3. A correct and a miss-specified parametric model for the presmoothing function are included. To avoid repetitions, we only consider the model $U[0,4]$ for the censoring distribution.

- Simulation results given in Table 2.1 ($\theta = 0$), left. `M` is the number of trials, `n` is the sample size, and `a` is the θ parameter. The following R objects are used to contained the variables of interest $\widetilde{T}_1$, Δ_1, $\widetilde{Y}$, Δ_2: `t1tilde`, `d1`, `ytilde`, `d`

```
M=1000;n=500;a=0

t1=matrix(nrow=n,ncol=M)
t2=matrix(nrow=n,ncol=M)
y=matrix(nrow=n,ncol=M)
c=matrix(nrow=n,ncol=M)
```

```
ytilde=matrix(nrow=n,ncol=M)
t1tilde=matrix(nrow=n,ncol=M)
d=matrix(nrow=n,ncol=M)
d1=matrix(nrow=n,ncol=M)

for (j in 1:M) {

v1=runif(n,0,1);v2=runif(n,0,1)
u1=v1
A=a*(2*u1-1)-1;B=(1-a*(2*u1-1))^2+4*a*v2*(2*u1-1)
u2=2*v2/(sqrt(B)-A)
t1[,j]=log(1/(1-u1))
t2[,j]=log(1/(1-u2))

y[,j]=t1[,j]+t2[,j]

c[,j]=runif(n,0,4)}

for (j in 1:M) for (i in 1:n)

{ytilde[i,j]=min(y[i,j],c[i,j])
d[i,j]=1
d1[i,j]=1}

for (j in 1:M) for (i in 1:n){if (ytilde[i,j]<y[i,j]) d[i,j]=0}

for (j in 1:M) for (i in 1:n){t1tilde[i,j]=min(t1[i,j],c[i,j])}

for (j in 1:M) for (i in 1:n){if (t1tilde[i,j]<t1[i,j]) d1[i,j]=0}
```

- Re-ordering so cases with `d1==1` come first

```
for (j in 1:M) {
t1tilde[,j]=c(t1tilde[d1[,j]==1,j],t1tilde[d1[,j]==0,j])
ytilde[,j]=c(ytilde[d1[,j]==1,j],ytilde[d1[,j]==0,j])
d[,j]=c(d[d1[,j]==1,j],d[d1[,j]==0,j])
d1[,j]=c(d1[d1[,j]==1,j],d1[d1[,j]==0,j])}
```

- Computing the presmoothing function. `Mlogitstar11` and `Mlogit1` stand for the correct

and the miss-specified parametric model, respectively. `M0` contains the values of the true presmoothing function (gold standard).

```
Mlogitstar11=matrix(nrow=n,ncol=M)
Mlogit1=matrix(nrow=n,ncol=M)
M0=matrix(nrow=n,ncol=M)

ytildestar=log(1/(4-ytilde))
t1tildestar=log(1/(4-t1tilde))

for (j in 1:M) {

n1=length(d[d1[,j]==1,j])

fittedzstar11=rep(1,n)

fittedzstar11=fitted(glm(d[d1[,j]==1,j]~t1tildestar[d1[,j]==1,j]
                    +ytildestar[d1[,j]==1,j],family=binomial))

fittedz1=fitted(glm(d[d1[,j]==1,j]~t1tilde[d1[,j]==1,j]
                   +ytilde[d1[,j]==1,j],family=binomial))

Mlogitstar11[,j]=c(fittedzstar11,rep(0,n-n1))
Mlogit1[,j]=c(fittedz1,rep(0,n-n1))
M0[,j]=c((4-ytilde[1:n1,j])/(5-ytilde[1:n1,j]),rep(0,n-n1))
```

- Calculating presmoothed and non-presmoothed KM weights: `W` (presmoothed KM, correctly specified model); `W1` (premoothed KM under miss-specification); `W0` (premoothed KM, true presmoothing function) and `Wkm` (no presmoothed KM).

```
W=matrix(nrow=n,ncol=M)
W1=matrix(nrow=n,ncol=M)
W0=matrix(nrow=n,ncol=M)
Wkm=matrix(nrow=n,ncol=M)
for (k in 1:M) {

P=rep(1,n)
R=rank(ytilde[,k])

for (i in 1:n){
```

```
for (j in 1:n){ if (R[j]<R[i])
P[i]<-P[i]*(1-Mlogitstar11[j,k]/(n-R[j]+1))}
W[i,k]<-P[i]*Mlogitstar11[i,k]/(n-R[i]+1)}

Pkm=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
Pkm[i]<-Pkm[i]*(1-d[j,k]/(n-R[j]+1))}
Wkm[i,k]<-Pkm[i]*d[i,k]/(n-R[i]+1)}

P1=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P1[i]<-P1[i]*(1-Mlogit1[j,k]/(n-R[j]+1))}
W1[i,k]<-P1[i]*Mlogit1[i,k]/(n-R[i]+1)}

P0=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P0[i]<-P0[i]*(1-M0[j,k]/(n-R[j]+1))}
W0[i,k]<-P0[i]*M0[i,k]/(n-R[i]+1)}}
```

- Calculating the values $\hat{F}(x_1, x_2)$

```
t2tilde=ytilde-t1tilde

F=matrix(nrow=16,ncol=M)
Fkm=matrix(nrow=16,ncol=M)
F1=matrix(nrow=16,ncol=M)
F0=matrix(nrow=16,ncol=M)

x1=rep(c(0.2231,0.5108,0.9163,1.6094),4)
x2=c(rep(0.2231,4),rep(0.5108,4),rep(0.9163,4),rep(1.6094,4))

I1=matrix(nrow=n,ncol=16)

for (j in 1:M) {

for (k in 1:16) { for (i in 1:n) { if (t1tilde[i,j]<=x1[k] & t2tilde[i,j]<=x2[k])
```

```
I1[i,k]<-1 else I1[i,k]=0}}
```

- Table of $F(x_1, x_2)$ estimated values (organized by columns). Calculation of MSE.

```
F[,j]<-t(as.matrix(W[,j]))%*%I1[,]
Fkm[,j]=t(as.matrix(Wkm[,j]))%*%I1[,]
F1[,j]=t(as.matrix(W1[,j]))%*%I1[,]
F0[,j]=t(as.matrix(W0[,j]))%*%I1[,]}

mseF=vector(length=16)
mseFkm=vector(length=16)
mseF1=vector(length=16)
mseF0=vector(length=16)

trueF=c(0.0400,0.0800,0.1200,0.1600,0.0800,0.1600,0.2400,0.3200,
        0.1200,0.2400,0.3600,0.4800,0.1600,0.3200,0.4800,0.6400)
for (i in 1:16) {mseF[i]=mean((F[i,]-trueF[i])^2)}
for (i in 1:16) {mseFkm[i]=mean((Fkm[i,]-trueF[i])^2)}
for (i in 1:16) {mseF1[i]=mean((F1[i,]-trueF[i])^2)}
for (i in 1:16) {mseF0[i]=mean((F0[i,]-trueF[i])^2)}
```

- Simulation results given in Table 2.1 ($\theta = 1$), left. `M` is the number of trials, `n` is the sample size, and `a` is the θ parameter.

```
M=1000;n=500;a=1

t1=matrix(nrow=n,ncol=M)
t2=matrix(nrow=n,ncol=M)
y=matrix(nrow=n,ncol=M)
c=matrix(nrow=n,ncol=M)

ytilde=matrix(nrow=n,ncol=M)
t1tilde=matrix(nrow=n,ncol=M)
d=matrix(nrow=n,ncol=M)
d1=matrix(nrow=n,ncol=M)

for (j in 1:M) {
```

```
v1=runif(n,0,1);v2=runif(n,0,1)
u1=v1
A=a*(2*u1-1)-1;B=(1-a*(2*u1-1))^2+4*a*v2*(2*u1-1)
u2=2*v2/(sqrt(B)-A)
t1[,j]=log(1/(1-u1))
t2[,j]=log(1/(1-u2))

y[,j]=t1[,j]+t2[,j]

c[,j]=runif(n,0,4)}

for (j in 1:M) for (i in 1:n)

{ytilde[i,j]=min(y[i,j],c[i,j])
d[i,j]=1
d1[i,j]=1}

for (j in 1:M) for (i in 1:n){if (ytilde[i,j]<y[i,j]) d[i,j]=0}

for (j in 1:M) for (i in 1:n){t1tilde[i,j]=min(t1[i,j],c[i,j])}

for (j in 1:M) for (i in 1:n){if (t1tilde[i,j]<t1[i,j]) d1[i,j]=0}
```

- Re-ordering so cases with `d1=1` come first

```
for (j in 1:M) {
t1tilde[,j]=c(t1tilde[d1[,j]==1,j],t1tilde[d1[,j]==0,j])
ytilde[,j]=c(ytilde[d1[,j]==1,j],ytilde[d1[,j]==0,j])
d[,j]=c(d[d1[,j]==1,j],d[d1[,j]==0,j])
d1[,j]=c(d1[d1[,j]==1,j],d1[d1[,j]==0,j])}
```

- Computing the presmoothing function. `Mlogitstar11` and `Mlogit1` stand for the correct and the miss-specified parametric model, respectively. `M0` contains the values of the true presmoothing function (gold standard).

```
Mlogitstar11=matrix(nrow=n,ncol=M)
Mlogit1=matrix(nrow=n,ncol=M)
M0=matrix(nrow=n,ncol=M)

ytildestar=log((1/(4-ytilde))*(2+2*exp(-ytilde)-2*exp(-t1tilde)-exp(-ytilde+t1tilde))
```

```
            /(2+4*exp(-ytilde)-2*exp(-t1tilde)-2*exp(-ytilde+t1tilde)))

for (j in 1:M) {

n1=length(d[d1[,j]==1,j])

fittedzstar11=rep(1,n)

fittedzstar11=fitted(glm(d[d1[,j]==1,j]~
ytildestar[d1[,j]==1,j],family=binomial))
fittedz1=fitted(glm(d[d1[,j]==1,j]~
t1tilde[d1[,j]==1,j]+ytilde[d1[,j]==1,j],family=binomial))

Mlogitstar11[,j]=c(fittedzstar11,rep(0,n-n1))
Mlogit1[,j]=c(fittedz1,rep(0,n-n1))

X=t1tilde[1:n1,j]
Y=ytilde[1:n1,j]-t1tilde[1:n1,j]

M0[,j]=c(1/(1+(1/(4-(X+Y)))*((2+2*exp(-(X+Y))-2*exp(-X)-exp(-Y))
       /(2+4*exp(-(Y+X))-2*exp(-X)-2*exp(-Y)))),rep(0,n-n1))}
```

- Calculating presmoothed and non-presmoothed KM weights: `W` (presmoothed KM, correctly specified model); `W1` (premoothed KM under miss-specification); `W0` (premoothed KM, true presmoothing function) and `Wkm` (no presmoothed KM)

```
W=matrix(nrow=n,ncol=M)
W1=matrix(nrow=n,ncol=M)
W0=matrix(nrow=n,ncol=M)
Wkm=matrix(nrow=n,ncol=M)

for (k in 1:M) {

P=rep(1,n)
R=rank(ytilde[,k])

for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P[i]<-P[i]*(1-Mlogitstar11[j,k]/(n-R[j]+1))}
```

```
W[i,k]<-P[i]*Mlogitstar11[i,k]/(n-R[i]+1)}

Pkm=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
Pkm[i]<-Pkm[i]*(1-d[j,k]/(n-R[j]+1))}
Wkm[i,k]<-Pkm[i]*d[i,k]/(n-R[i]+1)}

P1=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P1[i]<-P1[i]*(1-Mlogit1[j,k]/(n-R[j]+1))}
W1[i,k]<-P1[i]*Mlogit1[i,k]/(n-R[i]+1)}

P0=rep(1,n)
for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P0[i]<-P0[i]*(1-M0[j,k]/(n-R[j]+1))}
W0[i,k]<-P0[i]*M0[i,k]/(n-R[i]+1)}}
```

- Calculating the values $\hat{F}(x_1, x_2)$

```
t2tilde=ytilde-t1tilde

F=matrix(nrow=16,ncol=M)
Fkm=matrix(nrow=16,ncol=M)
F1=matrix(nrow=16,ncol=M)
F0=matrix(nrow=16,ncol=M)

x1=rep(c(0.2231,0.5108,0.9163,1.6094),4)
x2=c(rep(0.2231,4),rep(0.5108,4),rep(0.9163,4),rep(1.6094,4))

I1=matrix(nrow=n,ncol=16)

for (j in 1:M) {

for (k in 1:16) { for (i in 1:n) { if (t1tilde[i,j]<=x1[k] & t2tilde[i,j]<=x2[k])
I1[i,k]<-1 else I1[i,k]=0}}
```

- Table of $F(x_1, x_2)$ estimated values (organized by columns). Calculation of MSE.

```
F[,j]<-t(as.matrix(W[,j]))%*%I1[,]
Fkm[,j]=t(as.matrix(Wkm[,j]))%*%I1[,]
F1[,j]=t(as.matrix(W1[,j]))%*%I1[,]
F0[,j]=t(as.matrix(W0[,j]))%*%I1[,]}

mseF=vector(length=16)
mseFkm=vector(length=16)
mseF1=vector(length=16)
mseF0=vector(length=16)

trueF=c(0.0656,0.1184,0.1584,0.1856,0.1184,0.2176,0.2976,
0.3584,0.1584,0.2976,0.4176,0.5184,0.1856,0.3584,0.5184,0.6656)
for (i in 1:16) {mseF[i]=mean((F[i,]-trueF[i])^2)}
for (i in 1:16) {mseFkm[i]=mean((Fkm[i,]-trueF[i])^2)}
for (i in 1:16) {mseF1[i]=mean((F1[i,]-trueF[i])^2)}
for (i in 1:16) {mseF0[i]=mean((F0[i,]-trueF[i])^2)}
```

Now we give the R code corresponding to the simulation results in Table 2.3. We compute the mean and standard deviations of the quotient (`sd bootstrap/sd montecarlo`), for the cases $\theta = 0, 1$ (independent and dependent gap times) and $U[0, 4]$ for censoring (Models 1 and 3). The parametric model for the presmoothing function is correctly specified. `M` is the number of trials, `n` the sample size, and `B` the number of bootstrap resamples.

- Simulating the data, Model 1 ($\theta = 0$)

```
M=500;n=100;B=100;a=0

t1=matrix(nrow=n,ncol=M)
t2=matrix(nrow=n,ncol=M)
c=matrix(nrow=n,ncol=M)

ytilde=matrix(nrow=n,ncol=M)
t1tilde=matrix(nrow=n,ncol=M)
d=matrix(nrow=n,ncol=M)
d1=matrix(nrow=n,ncol=M)

for (j in 1:M) {
```

```
v1=runif(n,0,1);v2=runif(n,0,1)
u1=v1
A=a*(2*u1-1)-1;B0=(1-a*(2*u1-1))^2+4*a*v2*(2*u1-1)
u2=2*v2/(sqrt(B0)-A)
t1[,j]=log(1/(1-u1))
t2[,j]=log(1/(1-u2))
c[,j]=runif(n,0,4)}

y=t1+t2

for (j in 1:M) for (i in 1:n)

{ytilde[i,j]=min(y[i,j],c[i,j])
d[i,j]=1
d1[i,j]=1}

for (j in 1:M) for (i in 1:n){if (ytilde[i,j]<y[i,j]) d[i,j]=0}

for (j in 1:M) for (i in 1:n){t1tilde[i,j]=min(t1[i,j],c[i,j])}

for (j in 1:M) for (i in 1:n){if (t1tilde[i,j]<t1[i,j]) d1[i,j]=0}
```

- Re-ordering so the cases `d1==1` come first; computing the presmoothing function

```
for (j in 1:M) {
t1tilde[,j]=c(t1tilde[d1[,j]==1,j],t1tilde[d1[,j]==0,j])
ytilde[,j]=c(ytilde[d1[,j]==1,j],ytilde[d1[,j]==0,j])
d[,j]=c(d[d1[,j]==1,j],d[d1[,j]==0,j])
d1[,j]=c(d1[d1[,j]==1,j],d1[d1[,j]==0,j])}

M0=matrix(nrow=n,ncol=M)

ytildestar=log(1/(4-ytilde))
t1tildestar=log(1/(4-t1tilde))

for (j in 1:M) {

n1=length(d[d1[,j]==1,j])
```

```
fit0=fitted(glm(d[d1[,j]==1,j]~t1tildestar[d1[,j]==1,j]
             +ytildestar[d1[,j]==1,j],family=binomial))

M0[,j]=c(fit0,rep(0,n-n1))}
```

- Calculating presmoothed KM weights

```
W=matrix(nrow=n,ncol=M)

for (k in 1:M) {

P=rep(1,n)
R=rank(ytilde[,k])

for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P[i]<-P[i]*(1-M0[j,k]/(n-R[j]+1))}
W[i,k]<-P[i]*M0[i,k]/(n-R[i]+1)}}
```

- Calculating $\hat{F}(x_1, x_2)$ (`F`), and the Monte Carlo standard deviation (`sdF`)

```
t2tilde=ytilde-t1tilde

F=matrix(nrow=16,ncol=M)

x1=rep(c(0.2231,0.5108,0.9163,1.6094),4)
x2=c(rep(0.2231,4),rep(0.5108,4),rep(0.9163,4),rep(1.6094,4))

I1=matrix(nrow=n,ncol=16)

for (j in 1:M) {

for (k in 1:16) { for (i in 1:n) { if (t1tilde[i,j]<=x1[k] & t2tilde[i,j]<=x2[k])
I1[i,k]<-1 else I1[i,k]=0}}

F[,j]<-t(as.matrix(W[,j]))%*%I1[,]}

sdF=vector(length=16)
```

```
for (i in 1:16) { sdF[i]=sd(F[i,]) }
```

- Calculating the bootstrap standard deviation and summarizing the rates bootstrap vs. Monte Carlo

```
sdFb=matrix(nrow=16,ncol=M)
Fboot=matrix(nrow=B,ncol=16)

ind<-seq(1,n,by=1)

for (j in 1:M) {

tt1tilde<-t1tilde[,j]
yytilde<-ytilde[,j]
dd1<-d1[,j]
dd<-d[,j]

C=cbind(tt1tilde,yytilde,dd1,dd)

for (b in 1:B) {

indb<-sample(ind,n,replace=TRUE)
M1b<-C[indb,]

M2b0<-matrix(0,nrow=nrow(M1b),ncol=ncol(M1b))
ord<-order(M1b[,3],decreasing=T)
M2b0[,3]<-sort(M1b[,3],decreasing=T)
M2b0[,-3]<-M1b[ord,-3]

M2b<-matrix(0,nrow=nrow(M1b),ncol=ncol(M1b))
ord<-order(M2b0[,4],decreasing=T)
M2b[,4]<-sort(M2b0[,4],decreasing=T)
M2b[,-4]<-M2b0[ord,-4]

n1=sum(M2b[,3])
fit0b=fitted(glm(M2b[1:n1,4]~I(log(1/(4-M2b[1:n1,1])))
          +I(log(1/(4-M2b[1:n1,2]))),family=binomial))
```

```
M0b=c(fit0b,rep(0,n-n1))

Rb=rank(M2b[,2],ties.method="first")

Pb=rep(1,n)
Wb=rep(0,n)

for (i in 1:n){
for (k in 1:n){ if (Rb[k]<Rb[i])
Pb[i]<-Pb[i]*(1-M0b[k]/(n-Rb[k]+1))}
Wb[i]<-Pb[i]*M0b[i]/(n-Rb[i]+1)}

t2tildeb=M2b[,2]-M2b[,1]

I1b=matrix(nrow=n,ncol=16)

for (k in 1:16) { for (i in 1:n) { if (M2b[i,1]<=x1[k] & t2tildeb[i]<=x2[k])
I1b[i,k]<-1 else I1b[i,k]=0}}

Fb<-t(as.matrix(Wb))%*%I1b

Fboot[b,]<-Fb}

for (i in 1:16) { sdFb[i,j]=sd(Fboot[,i]) }}

R=matrix(nrow=16,ncol=M)
Rm=vector(length=16)
Rsd=vector(length=16)

for (i in 1:16) for (j in 1:M) {

R[i,j]=sdFb[i,j]/sdF[i]}

for (i in 1:16) {

Rm[i]=mean(R[i,]);Rsd[i]=sd(R[i,])}

cbind(Rm,Rsd)
```

- Simulating the data, Model 3 ($\theta = 1$)

```
M=500;n=100;a=1;B=100

t1=matrix(nrow=n,ncol=M)
t2=matrix(nrow=n,ncol=M)
c=matrix(nrow=n,ncol=M)

ytilde=matrix(nrow=n,ncol=M)
t1tilde=matrix(nrow=n,ncol=M)
d=matrix(nrow=n,ncol=M)
d1=matrix(nrow=n,ncol=M)

for (j in 1:M) {

v1=runif(n,0,1);v2=runif(n,0,1)
u1=v1
A=a*(2*u1-1)-1;B0=(1-a*(2*u1-1))^2+4*a*v2*(2*u1-1)
u2=2*v2/(sqrt(B0)-A)
t1[,j]=log(1/(1-u1))
t2[,j]=log(1/(1-u2))

c[,j]=runif(n,0,4)}

y=t1+t2

for (j in 1:M) for (i in 1:n)

{ytilde[i,j]=min(y[i,j],c[i,j])
d[i,j]=1
d1[i,j]=1}

for (j in 1:M) for (i in 1:n){if (ytilde[i,j]<y[i,j]) d[i,j]=0}

for (j in 1:M) for (i in 1:n){t1tilde[i,j]=min(t1[i,j],c[i,j])}

for (j in 1:M) for (i in 1:n){if (t1tilde[i,j]<t1[i,j]) d1[i,j]=0}
```

- Re-ordering so the cases `d1==1` come first; calculating the presmoothing function

```
for (j in 1:M) {
t1tilde[,j]=c(t1tilde[d1[,j]==1,j],t1tilde[d1[,j]==0,j])
ytilde[,j]=c(ytilde[d1[,j]==1,j],ytilde[d1[,j]==0,j])
d[,j]=c(d[d1[,j]==1,j],d[d1[,j]==0,j])
d1[,j]=c(d1[d1[,j]==1,j],d1[d1[,j]==0,j])}

M0=matrix(nrow=n,ncol=M)

ytildestar=log((1/(4-ytilde))*(2+2*exp(-ytilde)-2*exp(-t1tilde)-exp(-ytilde+t1tilde))
                          /(2+4*exp(-ytilde)-2*exp(-t1tilde)-2*exp(-ytilde+t1tilde)))

for (j in 1:M) {

n1=length(d[d1[,j]==1,j])

fit0=fitted(glm(d[d1[,j]==1,j]~ytildestar[d1[,j]==1,j],family=binomial))

M0[,j]=c(fit0,rep(0,n-n1))
```

- Calculating presmoothed KM weights

```
W=matrix(nrow=n,ncol=M)

for (k in 1:M) {

P=rep(1,n)
R=rank(ytilde[,k])

for (i in 1:n){
for (j in 1:n){ if (R[j]<R[i])
P[i]<-P[i]*(1-M0[j,k]/(n-R[j]+1))}
W[i,k]<-P[i]*M0[i,k]/(n-R[i]+1)}}
```

- Calculating the values $\hat{F}(x_1, x_2)$ (`F`), and the Monte Carlo standard deviation (`sdF`)

```
t2tilde=ytilde-t1tilde

F=matrix(nrow=16,ncol=M)

x1=rep(c(0.2231,0.5108,0.9163,1.6094),4)
```

```
x2=c(rep(0.2231,4),rep(0.5108,4),rep(0.9163,4),rep(1.6094,4))

I1=matrix(nrow=n,ncol=16)

for (j in 1:M) {

for (k in 1:16) { for (i in 1:n) { if (t1tilde[i,j]<=x1[k] & t2tilde[i,j]<=x2[k])
I1[i,k]<-1 else I1[i,k]=0}}

F[,j]<-t(as.matrix(W[,j]))%*%I1[,]}

sdF=vector(length=16)

for (i in 1:16) { sdF[i]=sd(F[i,]) }
```

- Calculating the bootstrap standard deviation and summarizing the rates bootstrap vs. Monte Carlo

```
sdFb=matrix(nrow=16,ncol=M)
Fboot=matrix(nrow=B,ncol=16)

ind<-seq(1,n,by=1)

for (j in 1:M) {

tt1tilde<-t1tilde[,j]
yytilde<-ytilde[,j]
dd1<-d1[,j]
dd<-d[,j]

C=cbind(tt1tilde,yytilde,dd1,dd)

for (b in 1:B) {

indb<-sample(ind,n,replace=TRUE)
M1b<-C[indb,]

M2b0<-matrix(0,nrow=nrow(M1b),ncol=ncol(M1b))
```

```
ord<-order(M1b[,3],decreasing=T)
M2b0[,3]<-sort(M1b[,3],decreasing=T)
M2b0[,-3]<-M1b[ord,-3]

M2b<-matrix(0,nrow=nrow(M1b),ncol=ncol(M1b))
ord<-order(M2b0[,4],decreasing=T)
M2b[,4]<-sort(M2b0[,4],decreasing=T)
M2b[,-4]<-M2b0[ord,-4]

n1=sum(M2b[,3])
fit0b=fitted(glm(M2b[1:n1,4]~I(log((1/(4-M2b[1:n1,2]))*(2+2*exp(-M2b[1:n1,2])
-2*exp(-M2b[1:n1,1])-exp(-M2b[1:n1,2]+M2b[1:n1,1]))/(2+4*exp(-M2b[1:n1,2])
-2*exp(-M2b[1:n1,1])-2*exp(-M2b[1:n1,2]+M2b[1:n1,1])))
),family=binomial))

M0b=c(fit0b,rep(0,n-n1))

Rb=rank(M2b[,2],ties.method="first")

Pb=rep(1,n)
Wb=rep(0,n)

for (i in 1:n){
for (k in 1:n){ if (Rb[k]<Rb[i])
Pb[i]<-Pb[i]*(1-M0b[k]/(n-Rb[k]+1))}
Wb[i]<-Pb[i]*M0b[i]/(n-Rb[i]+1)}

t2tildeb=M2b[,2]-M2b[,1]

I1b=matrix(nrow=n,ncol=16)

for (k in 1:16) { for (i in 1:n) { if (M2b[i,1]<=x1[k] & t2tildeb[i]<=x2[k])
I1b[i,k]<-1 else I1b[i,k]=0}}

Fb<-t(as.matrix(Wb))%*%I1b

Fboot[b,]<-Fb}

for (i in 1:16) { sdFb[i,j]=sd(Fboot[,i])}}
```

```
R=matrix(nrow=16,ncol=M)
Rm=vector(length=16)
Rsd=vector(length=16)

for (i in 1:16) for (j in 1:M) {

R[i,j]=sdFb[i,j]/sdF[i]}

for (i in 1:16) {

Rm[i]=mean(R[i,]);Rsd[i]=sd(R[i,])}

cbind(Rm,Rsd)
```

4.3 A simple example in the illness-death model

This section aims to illustrate the use of presmoothing for computing transition probabilities in the scope of the illness-death model (see Figure 3.1). A random sample with n=50 observations was generated from the Farlie-Gumbel-Morgenstern copula family, with $Exp(1)$ marginal and $\theta = 1$ (correlation of 0.25), when $\rho = 1$ (where $\rho \sim Ber(0.7)$); and directly from $Exp(1)$ when $\rho = 0$. This dataset is presented in Table 4.3. In this data input each individual is represented by one line of data. The variable $\widetilde{Z}$ represents the observed time in state 1; $\widetilde{T}$ is the observed total time. The variable Δ_1 denotes the status indicator of $\widetilde{Z}$ (taking value 1 if the individual is observed to leave state 1 and 0 otherwise); the variable Δ denotes the status of the total time (taking value 1 if a transition to state 3 is observed and 0 otherwise). The variable $\Delta_1\rho$ takes the value 1 if the individual is observed to experience a transition from state 1 into state 2. Note that possible courses for the individual include: $1 \to 1$ (the individual remains in state 1 until the end of the study; if $\Delta_1 = 0$); $1 \to 3$ (a direct transition from state 1 into state 3 is observed; if $\Delta_1 = 1$ and $\Delta_1\rho = 0$); $1 \to 2 \to 2$ (if $\Delta_1\rho = 1$ and $\Delta = 0$); and $1 \to 2 \to 3$ (if $\Delta_1\rho = 1$ and $\Delta = 1$). The variable W_i^Z denotes the Kaplan-Meier weight attached to $\widetilde{Z}$ when estimating the marginal distribution of Z from the $(\widetilde{Z}_i, \Delta_{1i})$'s. Similarly, the variable W_i^T denotes the Kaplan-Meier weight attached to $\widetilde{T}$ when estimating the marginal distribution of T from the $(\widetilde{T}_i, \Delta)$'s. Neither one uses presmoothing. On the other hand, the variables $W_i^Z(m_{0n})$ and $W_i^T(m_n)$ are both based on presmoothing. In order to obtain these weights we need to introduce (recall) the

presmoothing functions. The presmoothing function for the $W_i^T(m_n)$ weights is

$$m_n(z,t) = m_{1n}(z,t)\, I(z < t) + m_{2n}(t) I(z = t).$$

The estimator for $m_{1n}(z,t)$ is based on the sub-sample $\{i : \Delta_{1i}\rho_i = 1\}$ (i.e. all individuals passing through state 2); whereas, the $m_{2n}(t)$ function is computed from the sub-sample $\{i : \Delta_{1i}\rho_i = 0\}$ (i.e. individuals that never underwent state 2). In Table 4.1 we present the summary (coefficients, standard errors between brackets and p-value) of the three presmoothing functions m_{0n}, m_{1n} and m_{2n} based on logistic models, where $m_{0n}(z) = \hat{P}(\Delta_1 = 1|\widetilde{Z} = z)$, $m_{1n}(z,t) = \hat{P}(\Delta = 1|\widetilde{Z} = z, \widetilde{T} = t, \Delta_1\rho = 1)$ and $m_{2n}(t) = \hat{P}(\Delta_1 = 1|\widetilde{Z} = z, \Delta_1\rho = 0)$. In this case the influence of $\widetilde{T}$ on $m_{1n}(z,t)$ does not reach statistical significance and, on the other hand, $\widetilde{Z}$ is statistically significant for the presmoothing functions m_{0n} and m_{2n}.

Presmoothing function	Estimated Coefficients	p-value
$m_{0n}(z) = (1+\exp(\widehat{\eta}_0 + \widehat{\eta}_1 z))^{-1}$	$\widehat{\eta}_0 = 1.8116 \quad (0.5470)$	0.0010
	$\widehat{\eta}_1 = -1.3500 \quad (0.5509)$	0.0143
$m_{1n}(z,t) = (1+\exp(\widehat{\beta}_0 + \widehat{\beta}_1 z + \widehat{\beta}_2 t))^{-1}$	$\widehat{\beta}_0 = 4.0639 \quad (1.7938)$	0.0235
	$\widehat{\beta}_1 = 0.9782 \quad (1.3645)$	0.4735
	$\widehat{\beta}_2 = -2.8881 \quad (1.4850)$	0.0518
$m_{2n}(z) = (1+\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 z))^{-1}$	$\widehat{\gamma}_0 = 1.2156 \quad (0.6606)$	0.0657
	$\widehat{\gamma}_1 = -1.7595 \quad (0.8165)$	0.0312

Table 4.1: Summary of the three presmoothing functions m_{0n}, m_{1n} and m_{2n} based on logistic models.

In Table 4.2 we present the estimated transition probabilities $\hat{p}_{11}(s,t)$ and $\hat{p}_{23}(s,t)$, based on presmoothed Kaplan-Meier weights ($\hat{p}_{11}^{pkm}(s,t)$ and $\hat{p}_{23}^{pkm}(s,t)$) and without presmoothing ($\hat{p}_{11}^{km}(s,t)$ and $\hat{p}_{23}^{km}(s,t)$) for three pairs of (s,t) values. To obtain the estimator $\hat{p}_{11}^{km}(s,t)$ we only need to use the W_i^Z weights; whereas the estimator $\hat{p}_{23}^{km}(s,t)$ are obtained using the W_i^T weights. The presmoothed transition probability estimates $\hat{p}_{11}^{pkm}(s,t)$ and $\hat{p}_{23}^{pkm}(s,t)$ are obtained using the weights $W_i^Z(m_{0n})$ and $W_i^T(m_n)$, respectively. From Table 4.2 we see that the estimated transition probabilities change when introducing the presmoothing.

Now we give the `R` code needed to perform this example.

- Simulating the data. `t23` stands for the transition time from state 2 to state 3. `delta1` actually contains the value of $\Delta_1\rho$, while status stands for Δ. For computing the estimator of m_0 a new indicator `delta2` is needed, which contains the value of Δ_1

$(s,t) =$	$(0.2877, 0.7877)$	$(0.6931, 1.1931)$	$(1.3863, 1.8863)$
$\widehat{p}_{11}^{pkm}(s,t)$	0.4394	0.6809	0.7261
$\widehat{p}_{11}^{km}(s,t)$	0.4182	0.6420	0.6667
$\widehat{p}_{23}^{pkm}(s,t)$	0.6831	0.4733	0.8240
$\widehat{p}_{23}^{km}(s,t)$	0.4300	0.5931	1.0000

Table 4.2: Estimated transition probabilities with and without any presmoothing.

```
M=1; n=50; a=1

z=matrix(nrow=n,ncol=M)
t23=matrix(nrow=n,ncol=M)
c=matrix(nrow=n,ncol=M)
ztilde=matrix(nrow=n,ncol=M)
t23tilde=matrix(nrow=n,ncol=M)
delta1=matrix(nrow=n,ncol=M)
status=matrix(nrow=n,ncol=M)
ttilde=matrix(nrow=n,ncol=M)
for (j in 1:M) {

delta1[,j]=rbinom(n,1,0.7)
v1=runif(n,0,1);v2=runif(n,0,1)
u1=v1
A=a*(2*u1-1)-1;B=(1-a*(2*u1-1))^2+4*a*v2*(2*u1-1)
u2=2*v2/(sqrt(B)-A)
z[,j]=log(1/(1-u1))
t23[,j]=log(1/(1-u2))*delta1[,j]
c[,j]=runif(n,0,3)}

for (j in 1:M) for (i in 1:n)

{ztilde[i,j]=min(z[i,j],c[i,j])
 t23tilde[i,j]=min(t23[i,j],max(c[i,j]-z[i,j],0))
 status[i,j]=1}

for (j in 1:M) for (i in 1:n){if (t23tilde[i,j]<t23[i,j]) status[i,j]=0}
```

```
for (j in 1:M) for (i in 1:n){if (ztilde[i,j]<z[i,j])
{delta1[i,j]=0; status[i,j]=0; t23tilde[i,j]=0}}

delta2=matrix(nrow=n,ncol=M)
ttilde=matrix(nrow=n,ncol=M)
for (j in 1:M) for (i in 1:n){
ttilde[i,j]=ztilde[i,j]+t23tilde[i,j]
delta2[i,j]=delta1[i,j]+(1-delta1[i,j])*status[i,j]}
```

- Re-ordering so the cases `d1==1` come first

```
for (j in 1:M) {
o<-order(delta1[,j],decreasing = TRUE)
ztilde[,j]=ztilde[o,j]
delta1[,j]=delta1[o,j]
t23tilde[,j]=t23tilde[o,j]
status[,j]=status[o,j]
ttilde[,j]=ttilde[o,j]
delta2[,j]=delta2[o,j]}
```

- Computing the logistic presmoothing function

```
Mo_km=matrix(nrow=n,ncol=M)
Mo_logit=matrix(nrow=n,ncol=M)
for (j in 1:M){
Mo_logit[,j]=fitted(glm(delta2[,j]~ztilde[,j],family=binomial))}
```

- The following defines a function (named `pesosKM`) to compute the Kaplan-Meier weights, with and without presmoothing

```
pesosKM<-function(time,status){
M1<-cbind(time,status)
n=nrow(M1)
M2<-matrix(0,nrow=n,ncol=ncol(M1))
ord<-order(M1[,2],decreasing=TRUE)
M2[,2]<-sort(M1[,2],decreasing=TRUE)
M2[,-2]<-M1[ord,-2]
R=rank(M2[,1],ties.method="first")
Pkm2=rep(1,n)
Wkm<-rep(0,n)
```

```
Pkm2<-1-M2[,2]/(n-R+1)
count<-outer(R,R,"<")
Pkm2_aux<-matrix(Pkm2,nrow=n,ncol=n,byrow=FALSE)
Pkm2_2<-count*Pkm2_aux
Pkm2_2[Pkm2_2[,]==0]<-1
Pkm2_cum<-apply(Pkm2_2,2,prod)
Pkm3<-M2[,2]/(n-R+1)
Wkm<-Pkm3*Pkm2_cum
ord2<-order(M2[,1],decreasing=FALSE)
Wkm<-Wkm[ord2]
ord2<-rank(time,ties.method="first")
Wkm<-Wkm[ord2]
return(Wkm)}
```

- Weights of two estimators: presmoothed with logistic model (`Wo_logit`) and ordinary Kaplan-Meier (`Wo_km`) for the transition probability p_{11}.

```
Wo_km=matrix(nrow=n,ncol=M)
Wo_logit=matrix(nrow=n,ncol=M)

for (k in 1:M) {
Wo_km[,k]<-pesosKM(ztilde[,k],delta2[,k])
Wo_logit[,k]<-pesosKM(ztilde[,k],Mo_logit[,k])}
```

- Matrices `I11_ni` (i=1,2,3) used to take the 'numerator restriction' $I(Z > t)$ into account, for the three quartiles. Matrices `I11_di` (i=1,2,3) corresponding to the 'denominator restriction' $I(Z > s)$. Calculation of $\hat{p}_{11}(s,t)$. The values of `s`, are sample quartiles $s_1 \approx q1/4$, $s_2 \approx q1/2$ and $s_3 \approx q3/4$

```
s1=0.2877
s2=0.6931
s3=1.3863

t1=seq(from=s1,to=3,0.25)
t2=seq(from=s2,to=3,0.25)
t3=seq(from=s3,to=3,0.25)

I11_n1=matrix(nrow=n,ncol=length(t1))
I11_N1=array(0,dim=c(n,length(t1),M))
I11_d1=matrix(nrow=n,ncol=M)
```

```
I11_n2=matrix(nrow=n,ncol=length(t2))
I11_N2=array(0,dim=c(n,length(t2),M))
I11_d2=matrix(nrow=n,ncol=M)
I11_n3=matrix(nrow=n,ncol=length(t3))
I11_N3=array(0,dim=c(n,length(t3),M))
I11_d3=matrix(nrow=n,ncol=M)

hat_p11_km_n1=matrix(nrow=length(t1),ncol=M)
hat_p11_km_d1=matrix(nrow=M,ncol=length(s1))
hat_p11_km_1=matrix(nrow=length(t1),ncol=M)
hat_p11_km_n2=matrix(nrow=length(t2),ncol=M)
hat_p11_km_d2=matrix(nrow=M,ncol=length(s2))
hat_p11_km_2=matrix(nrow=length(t2),ncol=M)
hat_p11_km_n3=matrix(nrow=length(t3),ncol=M)
hat_p11_km_d3=matrix(nrow=M,ncol=length(s3))
hat_p11_km_3=matrix(nrow=length(t3),ncol=M)

hat_p11_logit_n1=matrix(nrow=length(t1),ncol=M)
hat_p11_logit_d1=matrix(nrow=M,ncol=length(s1))
hat_p11_logit_1=matrix(nrow=length(t1),ncol=M)
hat_p11_logit_n2=matrix(nrow=length(t2),ncol=M)
hat_p11_logit_d2=matrix(nrow=M,ncol=length(s2))
hat_p11_logit_2=matrix(nrow=length(t2),ncol=M)
hat_p11_logit_n3=matrix(nrow=length(t3),ncol=M)
hat_p11_logit_d3=matrix(nrow=M,ncol=length(s3))
hat_p11_logit_3=matrix(nrow=length(t3),ncol=M)

for (j in 1:M){

for( k in 1:length(t1)){for (i in 1:n)
        {if (ztilde[i,j] <= t1[k]) I11_N1[i,k,j]<-1}}

for( k in 1:length(t2)){for (i in 1:n)
        {if (ztilde[i,j] <= t2[k]) I11_N2[i,k,j]<-1}}

for( k in 1:length(t3)){for (i in 1:n)
        {if (ztilde[i,j] <= t3[k]) I11_N3[i,k,j]<-1}}

for (i in 1:n){
```

```
{if (ztilde[i,j] <= s1) I11_d1[i,j]<-1 else I11_d1[i,j]=0}
{if (ztilde[i,j] <= s2) I11_d2[i,j]<-1 else I11_d2[i,j]=0}
{if (ztilde[i,j] <= s3) I11_d3[i,j]<-1 else I11_d3[i,j]=0}}}
```

- Calculating $\hat{p}_{11}(s,t)$ with `Wo_km`

```
for (j in 1:M){
I11_n1<- I11_N1[,,j]
hat_p11_km_d1[j,1]=1-t(as.matrix(Wo_km[,j]))%*%I11_d1[,j]
for( k in 1:length(t1)){
hat_p11_km_n1[k,j]=1-t(as.matrix(Wo_km[,j]))%*%I11_n1[,k]
hat_p11_km_1[k,j]=hat_p11_km_n1[k,j]/hat_p11_km_d1[j,1]}}

for (j in 1:M){
I11_n2<- I11_N2[,,j]
hat_p11_km_d2[j,1]=1-t(as.matrix(Wo_km[,j]))%*%I11_d2[,j]
for( k in 1:length(t2)){
hat_p11_km_n2[k,j]=1-t(as.matrix(Wo_km[,j]))%*%I11_n2[,k]
hat_p11_km_2[k,j]=hat_p11_km_n2[k,j]/hat_p11_km_d2[j,1]}}

for (j in 1:M){
I11_n3<- I11_N3[,,j]
hat_p11_km_d3[j,1]=1-t(as.matrix(Wo_km[,j]))%*%I11_d3[,j]
for( k in 1:length(t3)){
hat_p11_km_n3[k,j]=1-t(as.matrix(Wo_km[,j]))%*%I11_n3[,k]
hat_p11_km_3[k,j]=hat_p11_km_n3[k,j]/hat_p11_km_d3[j,1]}}
```

- Calculating $\hat{p}_{11}(s,t)$ with `Wo_logit`

```
for (j in 1:M){
I11_n1<- I11_N1[,,j]
hat_p11_logit_d1[j,1]=1-t(as.matrix(Wo_logit[,j]))%*%I11_d1[,j]
for( k in 1:length(t1)){
hat_p11_logit_n1[k,j]=1-t(as.matrix(Wo_logit[,j]))%*%I11_n1[,k]
hat_p11_logit_1[k,j]=hat_p11_logit_n1[k,j]/hat_p11_logit_d1[j,1]}}

for (j in 1:M){1
I11_n2<- I11_N2[,,j]
hat_p11_logit_d2[j,1]=1-t(as.matrix(Wo_logit[,j]))%*%I11_d2[,j]
for( k in 1:length(t2)){
```

```
hat_p11_logit_n2[k,j]=1-t(as.matrix(Wo_logit[,j]))%*%I11_n2[,k]
hat_p11_logit_2[k,j]=hat_p11_logit_n2[k,j]/hat_p11_logit_d2[j,1]}}

for (j in 1:M){
I11_n3<- I11_N3[,,j]
hat_p11_logit_d3[j,1]=1-t(as.matrix(Wo_logit[,j]))%*%I11_d3[,j]
for( k in 1:length(t3)){
hat_p11_logit_n3[k,j]=1-t(as.matrix(Wo_logit[,j]))%*%I11_n3[,k]
hat_p11_logit_3[k,j]=hat_p11_logit_n3[k,j]/hat_p11_logit_d3[j,1]}}
```

- Matrix `M12_logit` of estimated $\hat{m}_n(Z,T)$ values

```
M12_logit=matrix(nrow=n,ncol=M)
M1_logit=matrix(nrow=n,ncol=M)
M2_logit=matrix(nrow=n,ncol=M)

for ( j in 1:M) {
ztilde[,j]=c(ztilde[delta1[,j]==1,j],ztilde[delta1[,j]==0,j])
delta1[,j]=c(delta1[delta1[,j]==1,j],delta1[delta1[,j]==0,j])
t23tilde[,j]=c(t23tilde[delta1[,j]==1,j],t23tilde[delta1[,j]==0,j])
status[,j]=c(status[delta1[,j]==1,j],status[delta1[,j]==0,j])
ttilde[,j]=c(ttilde[delta1[,j]==1,j],ttilde[delta1[,j]==0,j])
delta2[,j]=c(delta2[delta1[,j]==1,j],delta2[delta1[,j]==0,j])

n1=length(status[delta1[,j]==1,j])

m1_logit=fitted(glm(status[delta1[,j]==1,j]~ztilde[delta1[,j]==1,j]
                        +ttilde[delta1[,j]==1,j],family=binomial))
m2_logit=fitted(glm(status[delta1[,j]==0,j]~ztilde[delta1[,j]==0,j]
                                                ,family=binomial))
M1_logit[,j]=c(m1_logit,rep(0,n-n1))
M2_logit[,j]=c(rep(0,n1),m2_logit)
M12_logit[,j]=M1_logit[,j]+M2_logit[,j]}
```

- Calculating presmoothed and non-presmoothed KM weights, for m_n; `W12_logit` (presmoothed KM, logit model); `W12_km` (no presmoothed KM, KM ordinary)

```
W12_logit=matrix(nrow=n,ncol=M)
W12_km=matrix(nrow=n,ncol=M)
```

```
for (k in 1:M) {
W12_km[,k]<-pesosKM(ttilde[,k],status[,k])
W12_logit[,k]<-pesosKM(ttilde[,k],M12_logit[,k])}
```

- Matrices with the indicator-type restrictions, needed for computing $\hat{p}_{13}$ with `W12_logit` and `W12_km`

```
hat_p13_logit_n1=matrix(nrow=length(t1),ncol=M)
hat_p13_logit_d1=matrix(nrow=M,ncol=length(s1))
hat_p13_logit_1=matrix(nrow=length(t1),ncol=M)
hat_p13_logit_n2=matrix(nrow=length(t2),ncol=M)
hat_p13_logit_d2=matrix(nrow=M,ncol=length(s2))
hat_p13_logit_2=matrix(nrow=length(t2),ncol=M)
hat_p13_logit_n3=matrix(nrow=length(t3),ncol=M)
hat_p13_logit_d3=matrix(nrow=M,ncol=length(s3))
hat_p13_logit_3=matrix(nrow=length(t3),ncol=M)

hat_p13_km_n1=matrix(nrow=length(t1),ncol=M)
hat_p13_km_d1=matrix(nrow=M,ncol=length(s1))
hat_p13_km_1=matrix(nrow=length(t1),ncol=M)
hat_p13_km_n2=matrix(nrow=length(t2),ncol=M)
hat_p13_km_d2=matrix(nrow=M,ncol=length(s2))
hat_p13_km_2=matrix(nrow=length(t2),ncol=M)
hat_p13_km_n3=matrix(nrow=length(t3),ncol=M)
hat_p13_km_d3=matrix(nrow=M,ncol=length(s3))
hat_p13_km_3=matrix(nrow=length(t3),ncol=M)

I13_n1=matrix(nrow=n,ncol=length(t1))
I13_N1=array(0,dim=c(n,length(t1),M))
I13_d1=matrix(nrow=n,ncol=M)
I13_n2=matrix(nrow=n,ncol=length(t2))
I13_N2=array(0,dim=c(n,length(t2),M))
I13_d2=matrix(nrow=n,ncol=M)
I13_n3=matrix(nrow=n,ncol=length(t3))
I13_N3=array(0,dim=c(n,length(t3),M))
I13_d3=matrix(nrow=n,ncol=M)

for (j in 1:M){
```

```
for( k in 1:length(t1)){for (i in 1:n)
        {if (ztilde[i,j] > s1 & ttilde[i,j]<= t1[k])
             I13_N1[i,k,j]<-1}}
for( k in 1:length(t2)){for (i in 1:n)
        {if (ztilde[i,j] > s2 & ttilde[i,j]<= t2[k])
             I13_N2[i,k,j]<-1}}
for( k in 1:length(t3)){for (i in 1:n)
        {if (ztilde[i,j] > s3 & ttilde[i,j]<= t3[k])
             I13_N3[i,k,j]<-1}}

for (i in 1:n){
        {if (ztilde[i,j] <= s1) I13_d1[i,j]<-1 else I13_d1[i,j]=0}
        {if (ztilde[i,j] <= s2) I13_d2[i,j]<-1 else I13_d2[i,j]=0}
        {if (ztilde[i,j] <= s3) I13_d3[i,j]<-1 else I13_d3[i,j]=0}}}
```

- Calculating $\hat{p}_{13}$ with `W12_km` and `W12_logit`

```
for (j in 1:M){
I13_n1<-I13_N1[,,j]
hat_p13_km_d1[j,1]=hat_p11_km_d1[j,1]
for( k in 1:length(t1)){
hat_p13_km_n1[k,j]=t(as.matrix(W12_km[,j]))%*%I13_n1[,k]
hat_p13_km_1[k,j]=hat_p13_km_n1[k,j]/hat_p13_km_d1[j,1]}}

for (j in 1:M){
I13_n2<-I13_N2[,,j]
hat_p13_km_d2[j,1]=hat_p11_km_d2[j,1]
for( k in 1:length(t2)){
hat_p13_km_n2[k,j]=t(as.matrix(W12_km[,j]))%*%I13_n2[,k]
hat_p13_km_2[k,j]=hat_p13_km_n2[k,j]/hat_p13_km_d2[j,1]}}

for (j in 1:M){
I13_n3<-I13_N3[,,j]
hat_p13_km_d3[j,1]=hat_p11_km_d3[j,1]
for( k in 1:length(t3)){
hat_p13_km_n3[k,j]=t(as.matrix(W12_km[,j]))%*%I13_n3[,k]
hat_p13_km_3[k,j]=hat_p13_km_n3[k,j]/hat_p13_km_d3[j,1]}}
```

```
for (j in 1:M){
I13_n1<-I13_N1[,,j]
hat_p13_logit_d1[j,1]=hat_p11_logit_d1[j,1]
for( k in 1:length(t1)){
hat_p13_logit_n1[k,j]=t(as.matrix(W12_logit[,j]))%*%I13_n1[,k]
hat_p13_logit_1[k,j]=hat_p13_logit_n1[k,j]/hat_p13_logit_d1[j,1]}}

for (j in 1:M){
I13_n2<-I13_N2[,,j]
hat_p13_logit_d2[j,1]=hat_p11_logit_d2[j,1]
for( k in 1:length(t2)){
hat_p13_logit_n2[k,j]=t(as.matrix(W12_logit[,j]))%*%I13_n2[,k]
hat_p13_logit_2[k,j]=hat_p13_logit_n2[k,j]/hat_p13_logit_d2[j,1]}}

for (j in 1:M){
I13_n3<-I13_N3[,,j]
hat_p13_logit_d3[j,1]=hat_p11_logit_d3[j,1]
for( k in 1:length(t3)){
hat_p13_logit_n3[k,j]=t(as.matrix(W12_logit[,j]))%*%I13_n3[,k]
hat_p13_logit_3[k,j]=hat_p13_logit_n3[k,j]/hat_p13_logit_d3[j,1]}}
```

- Calculating $\hat{p}_{12}$ with logistic presmoothing and without any presmoothing

```
hat_p12_logit_n1=matrix(nrow=length(t1),ncol=M)
hat_p12_logit_d1=matrix(nrow=M,ncol=length(s1))
hat_p12_logit_1=matrix(nrow=length(t1),ncol=M)
hat_p12_logit_n2=matrix(nrow=length(t2),ncol=M)
hat_p12_logit_d2=matrix(nrow=M,ncol=length(s2))
hat_p12_logit_2=matrix(nrow=length(t2),ncol=M)
hat_p12_logit_n3=matrix(nrow=length(t3),ncol=M)
hat_p12_logit_d3=matrix(nrow=M,ncol=length(s3))
hat_p12_logit_3=matrix(nrow=length(t3),ncol=M)

hat_p12_km_n1=matrix(nrow=length(t1),ncol=M)
hat_p12_km_d1=matrix(nrow=M,ncol=length(s1))
hat_p12_km_1=matrix(nrow=length(t1),ncol=M)
hat_p12_km_n2=matrix(nrow=length(t2),ncol=M)
hat_p12_km_d2=matrix(nrow=M,ncol=length(s2))
hat_p12_km_2=matrix(nrow=length(t2),ncol=M)
```

```
hat_p12_km_n3=matrix(nrow=length(t3),ncol=M)
hat_p12_km_d3=matrix(nrow=M,ncol=length(s3))
hat_p12_km_3=matrix(nrow=length(t3),ncol=M)

hat_p12_logit_1=1-hat_p11_logit_1-hat_p13_logit_1
hat_p12_logit_2=1-hat_p11_logit_2-hat_p13_logit_2
hat_p12_logit_3=1-hat_p11_logit_3-hat_p13_logit_3
hat_p12_km_1=1-hat_p11_km_1-hat_p13_km_1
hat_p12_km_2=1-hat_p11_km_2-hat_p13_km_2
hat_p12_km_3=1-hat_p11_km_3-hat_p13_km_3
```

- Calculating $\hat{p}_{23}$ with logistic presmoothing and without any presmoothing

```
I23_n1=matrix(nrow=n,ncol=length(t1))
I23_N1=array(0,dim=c(n,length(t1),M))
I23_d1=matrix(nrow=n,ncol=M)
I23_n2=matrix(nrow=n,ncol=length(t2))
I23_N2=array(0,dim=c(n,length(t2),M))
I23_d2=matrix(nrow=n,ncol=M)
I23_n3=matrix(nrow=n,ncol=length(t3))
I23_N3=array(0,dim=c(n,length(t3),M))
I23_d3=matrix(nrow=n,ncol=M)

hat_p23_logit_n1=matrix(nrow=length(t1),ncol=M)
hat_p23_logit_d1=matrix(nrow=M,ncol=length(s1))
hat_p23_logit_1=matrix(nrow=length(t1),ncol=M)
hat_p23_logit_n2=matrix(nrow=length(t2),ncol=M)
hat_p23_logit_d2=matrix(nrow=M,ncol=length(s2))
hat_p23_logit_2=matrix(nrow=length(t2),ncol=M)
hat_p23_logit_n3=matrix(nrow=length(t3),ncol=M)
hat_p23_logit_d3=matrix(nrow=M,ncol=length(s3))
hat_p23_logit_3=matrix(nrow=length(t3),ncol=M)

hat_p23_km_n1=matrix(nrow=length(t1),ncol=M)
hat_p23_km_d1=matrix(nrow=M,ncol=length(s1))
hat_p23_km_1=matrix(nrow=length(t1),ncol=M)
hat_p23_km_n2=matrix(nrow=length(t2),ncol=M)
hat_p23_km_d2=matrix(nrow=M,ncol=length(s2))
hat_p23_km_2=matrix(nrow=length(t2),ncol=M)
```

```
hat_p23_km_n3=matrix(nrow=length(t3),ncol=M)
hat_p23_km_d3=matrix(nrow=M,ncol=length(s3))
hat_p23_km_3=matrix(nrow=length(t3),ncol=M)

for (j in 1:M){
for( k in 1:length(t1)){for (i in 1:n)
        {if (ztilde[i,j] <= s1 & ttilde[i,j]>s1 & ttilde[i,j]<= t1[k])
             I23_N1[i,k,j]<-1}}
for( k in 1:length(t2)){for (i in 1:n)
        {if (ztilde[i,j] <= s2 & ttilde[i,j]>s2 & ttilde[i,j]<= t2[k])
             I23_N2[i,k,j]<-1}}
for( k in 1:length(t3)){for (i in 1:n)
        {if (ztilde[i,j] <= s3 & ttilde[i,j]>s3 & ttilde[i,j]<= t3[k])
             I23_N3[i,k,j]<-1}}

for (i in 1:n){
        {if (ztilde[i,j]<= s1 & ttilde[i,j]>s1) I23_d1[i,j]<-1 else I23_d1[i,j]=0}
        {if (ztilde[i,j]<= s2 & ttilde[i,j]>s2) I23_d2[i,j]<-1 else I23_d2[i,j]=0}
        {if (ztilde[i,j]<= s3 & ttilde[i,j]>s3) I23_d3[i,j]<-1 else I23_d3[i,j]=0}}}

for (j in 1:M){
I23_n1<-I23_N1[,,j]
hat_p23_logit_d1[j,1]=t(as.matrix(W12_logit[,j]))%*%I23_d1[,j]
for( k in 1:length(t1)){
hat_p23_logit_n1[k,j]=t(as.matrix(W12_logit[,j]))%*%I23_n1[,k]
hat_p23_logit_1[k,j]=hat_p23_logit_n1[k,j]/hat_p23_logit_d1[j,1]}}

for (j in 1:M){
I23_n2<-I23_N2[,,j]
hat_p23_logit_d2[j,1]=t(as.matrix(W12_logit[,j]))%*%I23_d2[,j]
for( k in 1:length(t2)){
hat_p23_logit_n2[k,j]=t(as.matrix(W12_logit[,j]))%*%I23_n2[,k]
hat_p23_logit_2[k,j]=hat_p23_logit_n2[k,j]/hat_p23_logit_d2[j,1]}}

for (j in 1:M){
I23_n3<-I23_N3[,,j]
hat_p23_logit_d3[j,1]=t(as.matrix(W12_logit[,j]))%*%I23_d3[,j]
for( k in 1:length(t3)){
hat_p23_logit_n3[k,j]=t(as.matrix(W12_logit[,j]))%*%I23_n3[,k]
```

```
hat_p23_logit_3[k,j]=hat_p23_logit_n3[k,j]/hat_p23_logit_d3[j,1]}}

for (j in 1:M){
I23_n1<-I23_N1[,,j]
hat_p23_km_d1[j,1]=t(as.matrix(W12_km[,j]))%*%I23_d1[,j]
for( k in 1:length(t1)){
hat_p23_km_n1[k,j]=t(as.matrix(W12_km[,j]))%*%I23_n1[,k]
hat_p23_km_1[k,j]=hat_p23_km_n1[k,j]/hat_p23_km_d1[j,1]}}

for (j in 1:M){
I23_n2<-I23_N2[,,j]
hat_p23_km_d2[j,1]=t(as.matrix(W12_km[,j]))%*%I23_d2[,j]
for( k in 1:length(t2)){
hat_p23_km_n2[k,j]=t(as.matrix(W12_km[,j]))%*%I23_n2[,k]
hat_p23_km_2[k,j]=hat_p23_km_n2[k,j]/hat_p23_km_d2[j,1]}}

for (j in 1:M){
I23_n3<-I23_N3[,,j]
hat_p23_km_d3[j,1]=t(as.matrix(W12_km[,j]))%*%I23_d3[,j]
for( k in 1:length(t3)){
hat_p23_km_n3[k,j]=t(as.matrix(W12_km[,j]))%*%I23_n3[,k]
hat_p23_km_3[k,j]=hat_p23_km_n3[k,j]/hat_p23_km_d3[j,1]}}
```

4.4 Leukaemia data

In this section we report some results of our analysis of the leukaemia data provided by the IPO, which were briefly described in Section 1.2. Recall that we use an illness-death model for this data set, where state 1 represents the first transplant, state 2 is reserved for the second transplant, while state 3 represents the death of the patient. The 251 data points are presented in Figure 4.1, where different symbols are used according to the censoring status of each individual.

The computation of the semiparametric transition probabilities requires the preliminary estimation of three parametric models, one for each of the presmoothing functions involved in the problem. To this end we used a logistic model in all the cases. The results corresponding to the fitting of these logistic models are reported in Table 6.1. From this Table we see that the impact of the total survival time is statistically significative in the three cases (indeed, larger observed survival times correspond to larger probabilities of censoring), while the sojourn time in state 1 is not significative for the presmoothing function m_1. The fitted models are displayed in Figure 6.2

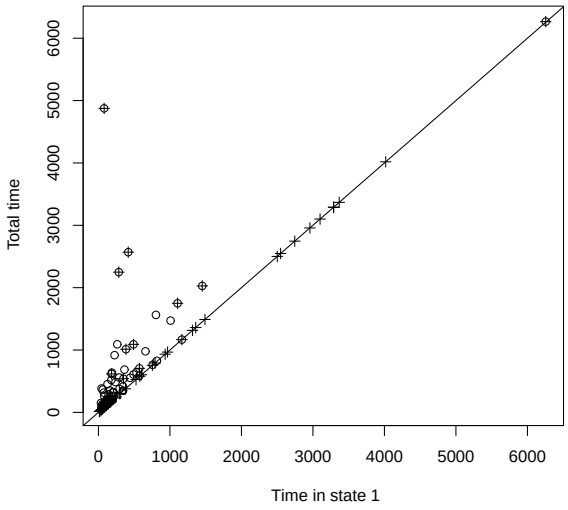

Figure 4.1: Leukaemia data: uncensored pairs ($\circ$), second gap time censored ($\oplus$), and both times censored ($+$).

together with the sampling information in which they are based. In Figure 6.2, right, the values of the estimated presmoothing function m_1 seem to be wiggly, as a result of the (non-significative) hidden influence of the time spent in state 1.

Figure 6.3 displays the transition probabilities $p_{i,j}(s,t)$ when estimated by the semiparametric estimator proposed in Chapter 3 or the (non-presmoothed) estimator proposed by Meira-Machado et al. (2006). As values of s we took the three sampling quartiles pertaining to $\widetilde{Z}$: 130, 335 and 1240 days. When comparing both estimators, we see that they are almost equal for t close to s; however, they become more different to each other as t grows. This is because of the redistribution of the mass attached to censored transition times which is achieved by the semiparametric estimator.

Interestingly, the first row in Figure 6.3 suggests that the probability of having a relapse decreases as the time passes by; the same is true for the probability of dying (third row). Similarly, the last two rows in Figure 6.3 indicate that the risk of dying is higher just after having the second transplant, and that then it decreases with time. Numerical results reported in Table 6.2 also support these comments. Specifically, according to the semiparametric estimator, the probability of staying in the initial ('healthy') state at time t=2000 days is increased 87% when the time elapsed after the first transplant (s) increases from 130 to 1240 days. Respectively, the probability of having died at time t=2000 days decreases 21% 1240 days after the second transplant when compared to 130 days after this second surgery. Interestingly, note that this

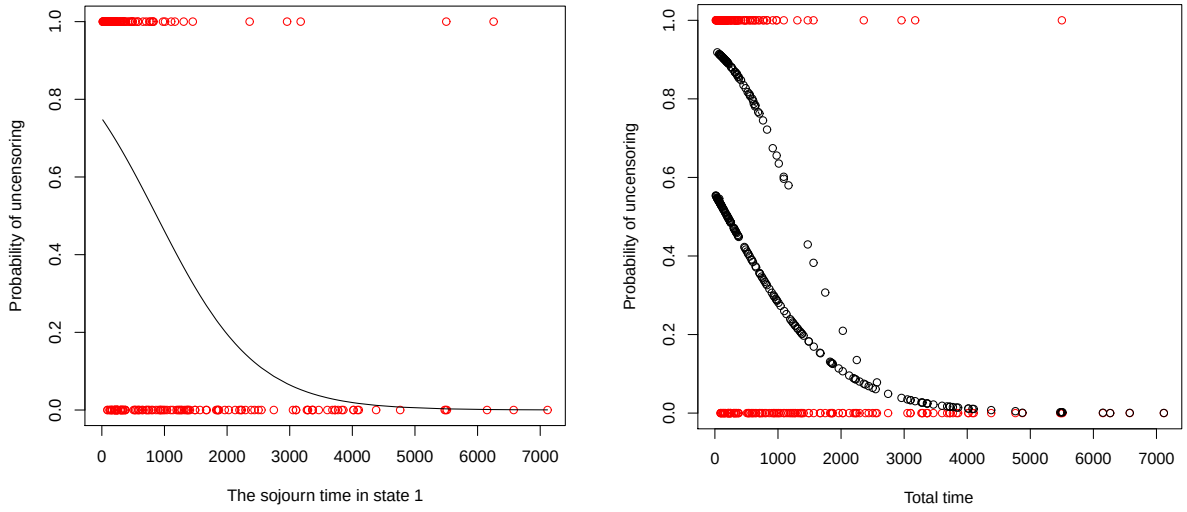

Figure 4.2: Presmoothing functions m_0 (left), m_1 (top) and m_2 (bottom) estimated by logistic models. Leukaemia data.

latter information is not available from the Kaplan-Meier-based estimator, which concentrates its mass in the uncensored transition times.

$\widetilde{Z}$	$\widetilde{T}$	Δ_1	Δ	$\Delta_1\rho$	W_i^Z	$W_i^Z(m_{0n})$	W_i^T	$W_i^T(m_n)$
0.3876	0.5303	1	1	1	0.0209	0.0169	0.0236	0.0216
0.2324	0.4423	1	0	1	0.0210	0.0169	0.0000	0.0212
0.5091	0.9360	1	1	1	0.0231	0.0171	0.0277	0.0225
2.1315	2.1871	1	0	1	0.0974	0.0233	0.0000	0.0434
0.0526	0.3043	1	1	1	0.0200	0.0171	0.0209	0.0203
1.7732	1.8987	1	1	1	0.0487	0.0165	0.0510	0.0314
0.7747	2.4381	1	0	1	0.0270	0.0183	0.0000	0.0118
0.4862	1.2153	1	1	1	0.0231	0.0169	0.0277	0.0214
0.1291	1.1969	1	1	1	0.0200	0.0170	0.0277	0.0192
0.3630	0.3983	1	1	1	0.0210	0.0166	0.0209	0.0211
0.6979	0.7626	1	1	1	0.0253	0.0178	0.0265	0.0233
0.5939	1.7032	1	0	1	0.0241	0.0173	0.0000	0.0183
0.6360	0.9510	1	1	1	0.0253	0.0175	0.0277	0.0229
0.4862	0.6681	1	1	1	0.0231	0.0171	0.0254	0.0223
0.7791	0.8939	1	1	1	0.0270	0.0187	0.0277	0.0234
0.5239	1.8041	1	0	1	0.0231	0.0172	0.0000	0.0170
0.7193	1.7355	1	1	1	0.0253	0.0180	0.0437	0.0199
0.6850	0.7380	1	1	1	0.0253	0.0176	0.0254	0.0228
0.0226	0.1385	1	1	1	0.0200	0.0171	0.0200	0.0197
0.6773	0.6773	1	1	0	0.0253	0.0174	0.0254	0.0122
1.0028	1.0028	1	1	0	0.0292	0.0183	0.0277	0.0100
0.4531	0.4531	0	0	0	0.0000	0.0168	0.0000	0.0134
1.6254	1.6254	1	1	0	0.0487	0.0167	0.0388	0.0063
1.3773	1.3773	0	0	0	0.0000	0.0151	0.0000	0.0068
1.4130	1.4130	0	0	0	0.0000	0.0156	0.0000	0.0069
0.1398	0.1398	1	1	0	0.0200	0.0170	0.0200	0.0147
0.5694	0.5694	1	1	0	0.0231	0.0171	0.0236	0.0126
0.1862	0.1862	1	1	0	0.0205	0.0169	0.0204	0.0145
0.7466	0.7466	0	0	0	0.0000	0.0181	0.0000	0.0117
0.3795	0.3795	1	1	0	0.0210	0.0167	0.0209	0.0135
0.1788	0.1787	0	0	0	0.0000	0.0170	0.0000	0.0145
0.8822	0.8822	0	0	0	0.0000	0.0182	0.0000	0.0105
1.4451	1.4451	0	0	0	0.0000	0.0163	0.0000	0.0070
0.4128	0.4128	0	0	0	0.0000	0.0169	0.0000	0.0136
0.1271	0.1271	1	1	0	0.0200	0.0169	0.0200	0.0147
0.3274	0.3274	1	1	0	0.0210	0.0167	0.0209	0.0138
0.3817	0.3817	1	1	0	0.0210	0.0168	0.0209	0.0136
0.3824	0.3824	1	1	0	0.0210	0.0169	0.0209	0.0137
1.5884	1.5884	0	0	0	0.0000	0.0157	0.0000	0.0061
2.0226	2.0226	0	0	0	0.0000	0.0153	0.0000	0.0052
0.4551	0.4551	0	0	0	0.0000	0.0170	0.0000	0.0136
2.1188	2.1188	0	0	0	0.0000	0.0172	0.0000	0.0054
0.2885	0.2885	1	1	0	0.0210	0.0167	0.0209	0.0139
0.5941	0.5941	0	0	0	0.0000	0.0175	0.0000	0.0128
0.5930	0.5930	0	0	0	0.0000	0.0171	0.0000	0.0126
0.2214	0.2214	0	0	0	0.0000	0.0169	0.0000	0.0143
0.3020	0.3020	1	1	0	0.0209	0.0167	0.0209	0.0139
2.7304	2.7304	0	0	0	0.0000	0.0211	0.0000	0.0062
0.0523	0.0523	1	1	0	0.0200	0.0171	0.0200	0.0151
0.9738	0.9738	1	1	0	0.0292	0.0180	0.0277	0.0100

Table 4.3: Simulated data.

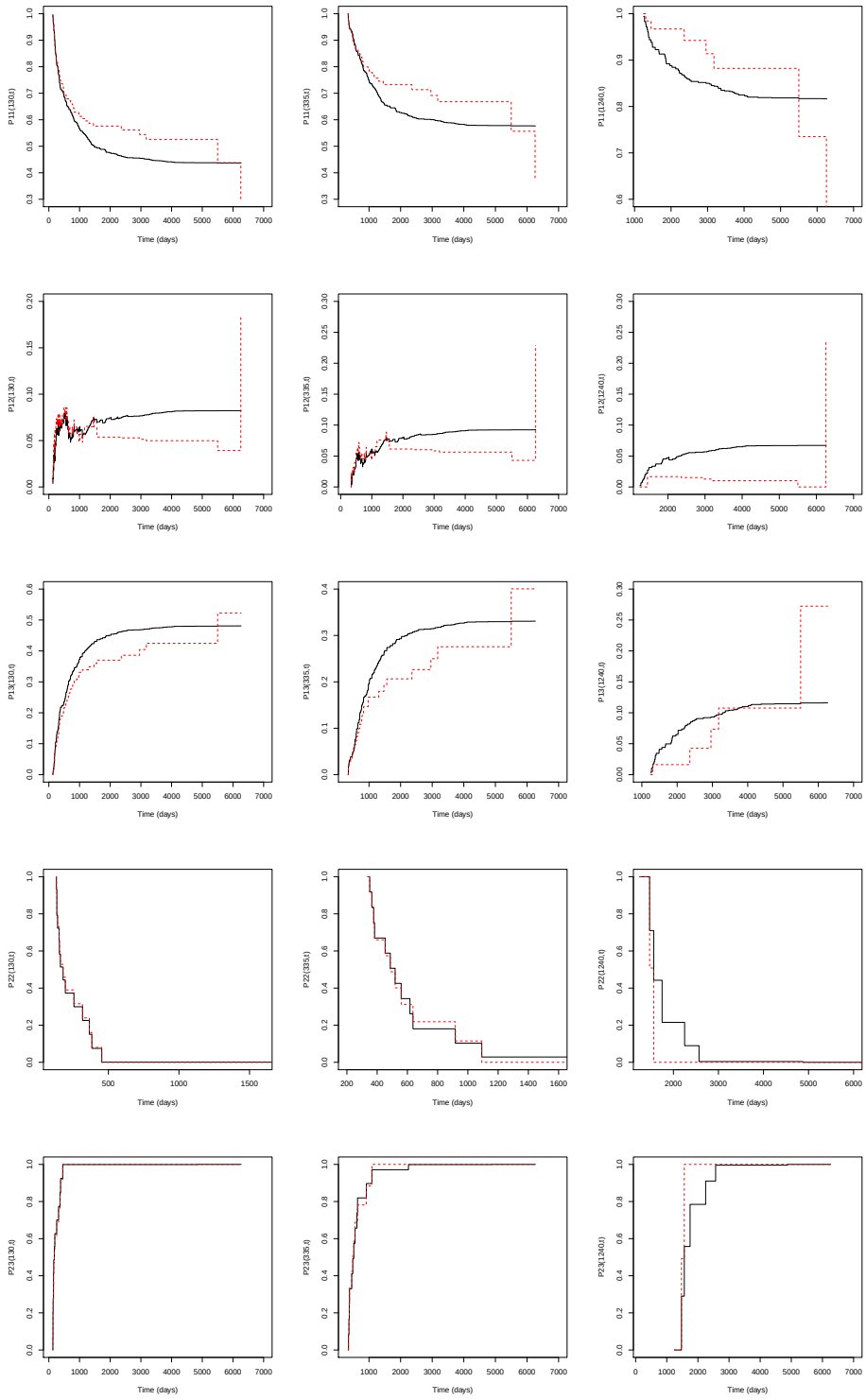

Figure 4.3: Estimated transition probabilities for $p_{ij}(s,t)$ with $s \in \{130, 335, 1240\}$ based on the Kaplan-Meier weights (dashed line) and based on presmoothed Kaplan-Meier weights (solid line). Leukaemia data.

Presmoothing function	Estimated Coefficients	p-value
$m_{0n}(z) = (1 + \exp(\widehat{\eta}_0 + \widehat{\eta}_1 z))^{-1}$	$\widehat{\eta}_0 = 1.1016 \quad (0.1926)$	0.0000
	$\widehat{\eta}_1 = -0.0013 \quad (0.0002)$	0.0000
$m_{1n}(z,t) = (1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 z + \widehat{\beta}_2 t))^{-1}$	$\widehat{\beta}_0 = 2.4920 \quad (0.5102)$	0.0000
	$\widehat{\beta}_1 = 0.0000 \quad (0.0015)$	0.9554
	$\widehat{\beta}_2 = -0.0019 \quad (0.0009)$	0.0391
$m_{2n}(z) = (1 + \exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 z))^{-1}$	$\widehat{\gamma}_0 = 0.2333 \quad (0.2366)$	0.3240
	$\widehat{\gamma}_1 = -0.0012 \quad (0.0003)$	0.0000

Table 4.4: Summary of the three presmoothing functions m_{0n}, m_{1n} and m_{2n} based on logistic models. Leukaemia data.

$(s,t) =$	$(130, 2000)$	$(335, 2000)$	$(1240, 2000)$
$\widehat{p}_{11}^{pkm}(s,t)$	0.4751	0.6273	0.8882
$\widehat{p}_{11}^{km}(s,t)$	0.5760	0.7320	0.9671
$\widehat{p}_{23}^{pkm}(s,t)$	0.9992	0.9714	0.7847
$\widehat{p}_{23}^{km}(s,t)$	1.0000	1.0000	1.0000

Table 4.5: Estimated transition probabilities with and without any presmoothing. Leukaemia data.

Chapter 5

Concluding remarks and future research

Contents

5.1 Concluding remarks 110
5.2 Future research 111

5.1 Concluding remarks

In this thesis we have introduced some semiparametric estimation strategies in the scope of the non-Markov three-state and illness-death progressive models. The proposed estimators make use of 'presmoothing' ideas, and this 'presmoothing' is driven by some specific semiparametric censorship models. Although presmoothed estimators are known in the classical (i.e. univariate) Survival Analysis setup, for the best of our knowledge their application in the complicated field of multi-state models is new. Just to mention a specific problem that need to be addressed, presmoothing functions which are discontinuous will arise when dealing with multivariate survival times.

More explicitly, in Chapter 2 a new semiparametric estimator $\widehat{F}_{12}^{sp}(x,y)$ of the bivariate distribution function of gap times which are observed under censoring is introduced. The semiparametric estimator is based on a parametric specification of the conditional probability of censoring for the second gap time T_2, given the available information. This specification can be tested in practice. We have derived the consistency of the proposed estimator and, more generally, of an empirical functional based on it. We have verified through simulations that the semiparametric estimator may be much more efficient than other available estimators. This will be particularly true at points in which there is a large proportion of censored T_2 among those with first gap time uncensored. Besides, we have seen that the method is robust against miss-specifications of the parametric model. We have also used the simple bootstrap to approximate the standard error of the estimator, and our simulation results suggest that the bootstrap provides an unbiased estimation. A real data illustration has been provided. Finally, an asymptotic representation of the estimator as a sum of i.i.d. random variables has been given, and its asymptotic normality has been consequently established.

In Chapter 3 we have introduced new semiparametric estimators for the transition probabilities of a censored, non-Markov illness-death model. The new estimators rely on several parametric models for various 'presmoothing' functions, which vary depending on the involved states. We have derived the consistency of the proposed estimators. The finite sample performance of the introduced estimators was investigated through simulations. As in Chapter 2, the main conclusion of Chapter 3 is that presmoothing leads to improved estimators, even when there is some miss-specification in the parametric family assumed for the presmoothing function. The relative benefits of presmoothing are more clearly seen in the heavily censored case. The new method has been illustrated using data from a colon cancer study, and it has been used to analyze leukaemia data provided by the IPO (Section 4.4).

The new estimators for the transition probabilities are consistent regardless the Markov condition (this is also true for the estimator proposed in Chapter 2). This is interesting because real problems are often far from markovianity and therefore the consistency of the time-honored Aalen-Johansen estimator can not be ensured. To this regard, one may think about the methods

introduced here as a remarkable improvement (in the sense of having less variance) of previous non-Markovian estimators (Meira-Machado et al. 2006).

In practice, the implementation of the proposed methods is far from being straightforward. In Chapter 4, Section 4.3, we have performed an in-detailed example describing the different weights which are needed to compute the presmoothing functions and the several estimators. Also, in Sections 4.2 and 4.3 we have provided our own `R` code, which allows to reproduce the several analysis performed in this thesis and, more importantly, to compute the proposed estimators for new real data sets. We believe that this is a remarkable contribution to practitioners.

5.2 Future research

In Corollary 2.5.2 we have derived the asymptotic normality of the semiparametric estimator $\widehat{F}_{12}^{sp}(x,y)$. It would be interesting to compare the limit variance $\sigma^2(x,y)$ in that Corollary to the asymptotic variance corresponding to the estimator proposed in de Uña-Álvarez and Meira-Machado (2008), which should be larger according to the intuition and the provided evidence (simulations). However, this comparison requires the derivation of an i.i.d. representation (and the asymptotic normality) of the latter estimator, which is missing so far in the literature.

Another point of technical and practical interest is to extend Theorem 2.5.1 to functions $\xi^{\varphi}(x,y)$ which do not satisfy condition (M1). We conjecture that this could be done by following lines similar to those in Dikta et al. (2005), but the complete adaptation of Dikta's theory to the context of censored gap times is still undeveloped.

It is also possible to extend the proposed estimator to the $k-$state progressive model for $k>3$; for this, the censoring indicator for the total survival time $Y=T_1+...+T_k$ could be replaced by a smooth (parametric) fit to the probability of uncensoring given the observed (possibly censored) gap times, and given that the $k-1$ first gap times are uncensored. The details are also left for future research.

In general, it would be interesting to propose semiparametric (non-Markov) estimators in multi-state models other than the $k-$state and the illness-death progressive models as, e.g. the bivariate model. Unfortunately, it seems that every multi-state model requires of specific solutions according to the nature of the sampling information.

Regarding the empirical transition probabilities proposed for the illness-death model, we aim in the future to obtain further asymptotics (e.g. a central limit theorem) as well as to propose a method to estimate the standard errors so confidence limits can be constructed.

Interestingly, nonparametric presmoothing is also possible for the proposed methods, as the main consistency results in Chapters 2 and 3 remain valid. This avoids the problem of choosing a

proper parametric family for the binary regression. However, the gains in efficiency when using a nonparametric binary regression curve should be explored in detail. Typically, this nonparametric presmoothing will involve the selection of several smoothing parameters, which may be a critical point in the final performance of the estimator. In any case, this seems to be another promising field of research.

Chapter 6

Summary in Spanish

Resumen en catellano

El Análisis de Supervivencia se ocupa de los tiempos entre eventos. En un contexto clásico, el centro de atención es el tiempo transcurrido entre dos eventos bien definidos: el evento inicial (o 'nacimiento') y el evento final (o 'muerte'). Este tiempo es, por tanto, llamado 'tiempo de vida' o 'tiempo de supervivencia'. Las aplicaciones del Análisis de Supervivencia incluyen la medicina, la biología, la economía, la astronomía y la ingeniería, entre otros campos. Cuando se analizan datos de supervivencia, uno debe enfrentarse al problema importante de la censura. Un tiempo de vida censurado ocurre cuando la observación del evento final no es posible. Esto puede ser debido a limitaciones de tiempo en el estudio, o porque otro evento relevante ocurre con anterioridad al evento final de interés. En este caso, el tiempo entre eventos registrado es estrictamente menor que el tiempo de interés, y se requieren adecuadas correcciones para realizar una estimación consistente de curvas y parámetros poblacionales.

En este escenario, el estimador límite-producto de Kaplan-Meier se ha convertido en el método estándar para estimar la probabilidad de supervivencia de manera no paramétrica. Las propiedades estadísticas del estimador de Kaplan-Meier han sido investigadas en profundidad; véase por ejemplo Klein y Moeschberger (1997). Además, este estimador ha sido adaptado a distintos problemas tales como la estimación de curvas suaves (como la función de densidad), curvas condicionales (por ejemplo la función de regresión y la función de distribución condicional), distribuciones multivariantes, y parámetros de regresión.

Sin embargo, uno de los principales inconvenientes del estimador de Kaplan-Meier es que presenta una gran varianza cuando la proporción de tiempos de vida censurados es elevada, particularmente en la cola derecha de la distribución. Con el objetivo de reducir la varianza en la estimación, varias alternativas a la curva Kaplan-Meier han sido propuestas. Estos estimadores alternativos hacen uso de alguna información adicional sobre el mecanismo de censura. El ejemplo más famoso es el estimador de Koziol-Green, véase Cheng y Lin (1987), que se basa en el supuesto de que la razón de fallo correspondiente a la variable de censura es proporcional a la razón de fallo de interés. Esta suposición es equivalente a la independencia condicional entre el indicador de censura y el tiempo de vida observable, lo cual no es realista en la práctica. Aún así, assumiendo que la probabilidad condicional de censura es una función suave, quizás no constante, del tiempo de vida observable, uno puede construir estimadores con menor varianza que el estimador de Kaplan-Meier. Este supuesto bastante menos restrictivo fue utilizado por distintos autores, véase por ejemplo Dikta (1998) y Cao et al. (2005), para introducir lo que nosotros en general llamamos 'estimadores presuavizados'.

En este contexto, 'presuavizar' significa reemplazar los indicadores de no censura por algún ajuste suave a la probabilidad condicional de no censura dado el tiempo de vida observable. Esto ha permitido reducir la varianza de los estimadores basados en la curva Kaplan-Meier en

diferentes problemas, incluyendo la estimación no paramétrica de curvas (Cao y Jácome (2004); Cao et al. (2005)) o el análisis de regresión (de Uña-Álvarez y Rodríguez-Campos (2004); Yuan (2005); Iglesias-Pérez y de Uña-Álvarez (2008)). Cuando la 'presuavización' se realiza en base a un modelo paramétrico, uno obtiene un modelo semiparamétrico de censura y, consecuentemente, un sustituto de tipo semiparamétrico para el estimador de Kaplan-Meier. Esta filosofía ha sido investigada con mucho detalle por Dikta (1998, 2000, 2001), véase también Dikta et al. (2005). Uno de los principales resultados derivados de tal investigación es que el estimador semiparamétrico tiene menor varianza (cuando se compara con el Kaplan-Meier), siendo por otra parte robusto a malas especificaciones del modelo paramétrico. El objetivo del presente trabajo es utilizar estas ideas en el contexto específico del modelo de tres estados progresivo, y en el modelo 'illness-death. Estos dos importante modelos multi-estado se discuten brevemente en los dos puntos siguientes.

Datos tipo 'gap times'

El análisis estadístico de 'gap times' consecutivos es un problema de mucha importancia en un número de campos, incluyendo la ingeniería, la economía, la epidemiología, y el análisis de supervivencia. En la mayor parte de los casos, uno estará interesado no sólo en describir la distribución marginal de los 'gap times' sino también en estudiar su estructura de correlación. Esto ocurre, por ejemplo, cuando se analizan datos sobre eventos recurrentes, que surgen cuando cada individuo puede sufrir un evento bien definido varias veces a lo largo de su historia. Entonces, los tiempos entre eventos se denominan 'gap times', y están obviamente determinados por los instantes en los que la recurrencia tiene lugar (es decir, por los tiempos de recurrencia). Véase Cook and Lawless (2007) para una revisión actualizada de los métodos estadísticos para datos sobre eventos recurrentes.

De manera alternativa, podemos pensar en los 'gap times' como tiempos que surgen de un modelo multi-estado particular. Los modelos multi-estado (Andersen et al. (1993); Meira-Machado et al. (2009)) son los modelos más habitualmente utilizados para describir datos de supervivencia longitudinales. Un modelo multi-estado es un modelo para un proceso estocástico que está caracterizado por un conjunto de estados y las posibles transiciones entre ellos. Los estados representan diferentes situaciones del individuo (sano, enfermo, etc) a lo largo de un seguimiento. Modelos multi-estado particulares que han sido ampliamente utilizados en aplicaciones biomédicas son el modelo de tres estados progresivo, el modelo 'illness-death', o el modelo bivariante (Hougaard (2000)). Como revisiones recientes sobre modelos multi-estado citamos Commenges (1999), Hougaard (1999), Andersen y Keiding (2002), y Meira-Machado et al. (2009).

El modelo de tres estados progresivo está formado por tres estados y dos posibles transiciones: del estado 1 al estado 2, y del estado 2 al estado 3. Consecuentemente, la observación de un proceso de este tipo proporciona información sobre dos 'gap times' consecutivos (que son los tiempos de transición entre los tres estados). En la práctica, al igual que en el contexto clásico del Análisis de Supervivencia que se describió arriba, la presencia de información censurada provoca compli-

caciones a la hora de hacer estimaciones. Véase por ejemplo Lin et al. (1999) y las referencias que cita. Éste es el contexto en el que se sitúa el Capítulo 2.

Modelo 'illness-death'

El modelo 'illness-death' es una generalización del modelo de tres estados progresivo en la cual se permite una transición directa del estado 1 al estado absorbente final (estado 3). Este modelo es muy importante en aplicaciones. Uno de los principales objetivos en este modelo es la estimación de las llamadas probabilidades de transición (véase el Capítulo 3 para una definición formal). Tradicionalmente, esta estimación se realiza bajo el supuesto de Markov, lo cual lleva al famoso estimador Aalen-Johansen (Aalen y Johansen, 1978). Sin embargo, en algunas aplicaciones la condición de Markov no se satisface (por ejemplo Andersen et al. (2000)), y el estimador Aalen-Johansen puede ser inconsistente. Para superar este problema, Meira-machado et al. (2006) propusieron un sustituto para el estimador de Aalen-Johansen que no depende de la condición de Markov. Desafortunadamente, la varianza de este estimador alternativo puede ser muy grande en escenarios fuertemente censurados. La posibilidad de mejorar el estimador de Meira-Machado et al. (2006) a través de la presuavización se explora en el Capítulo 3.

Datos reales

En esta tesis serán utilizados algunos conjuntos de datos a modo de ilustración. Uno de estos conjuntos de datos (los datos de cáncer de vejiga) se ajusta al modelo de tres estados progresivo, mientras que los datos sobre cáncer de colon se adaptan al modelo 'illness-death'. Además, utilizamos varios estimadores para analizar datos clínicos recientes proporcionados por el IPO (el Instituto Oncológico de Portugal, en Oporto) sobre transplantes de médula ósea para pacientes con leucemia aguda; este conjunto de datos se analiza también bajo la perspectiva del modelo 'illness-death'. Comentamos ahora brevemente cada uno de estos conjuntos de datos reales.

El Veterans Administration Cooperative Urological Research Group desarrolló un estudio sobre el cáncer de vejiga (Byar, 1980). En este estudio, los pacientes tenían tumores superficiales en la vejiga, que fueron eliminados de forma transuretral. Muchos pacientes tuvieron múltiples recurrencias de tumores, y los nuevos tumores fueron eliminados en cada visita. Aquí consideramos los 85 individuos en los grupos placebo y tratamiento 'thiotepa'; estos datos se encuentran listados en Wei et al. (1989). Están también disponibles en el paquete `survival` del software `R` (R-Development-Core-Team (2009)). Estos datos se utilizan en la Sección 2.4 del Capítulo 2 para ilustrar el comportamiento del estimador semiparamétrico de la función de distribución conjunta de los 'gap times'. Para ello consideramos únicamente las dos primera recurrencias recogidas en la base de datos.

Por su parte, los datos de cáncer de colon están también disponibles en el paquete `survival` de `R`. Estos datos vienen de un ensayo clínico a gran escala sobre pacientes en el nivel III de

Duke, afectados de cáncer de colon, que recibieron una cirugía curativa para el cáncer colo-rectal (Moertel et al. (1990)). En este estudio, del total de 929 pacientes, 468 tuvieron una recurrencia y, de entre éstos, 414 murieron. 38 pacientes murieron sin tener una recurrencia. El resto de los pacientes (423) permanecieron vivos y libres de la enfermedad hasta el fin del seguimiento. Debido a que la recurrencia puede verse como un evento intermedio, utilizamos el modelo 'illness-death' para representar estos datos. En la Sección 3.4 del Capítulo 3 utilizamos estos datos para ilustrar los estimadores semiparamétricos propuestos para las probabilidades de transición.

Finalmente, los datos de leucemia consisten en todos los individuos diagnosticados de leucemia aguda (linfocítica o mielocítica) entre Junio de 1989 y Abril de 2009 en el IPO (Instituto Oncológico de Portugal en Oporto). El número de individuos fue 251. El tratamiento estándar para la leucemia aguda es un transplante de médula ósea. Después del transplante puede existir una recaída. La recaída se definió en base a la evidencia morfológica de la leucemia en la médula ósea o en otros lugares. En caso de recaída, el paciente sufrió inmediatamente un segundo transplante, y así sucesivamente. Aquí consideramos únicamente los dos primeros transplantes, e investigamos el tiempo transcurrido entre los dos transplantes sucesivos así como el tiempo hasta la muerte (por cualquier causa). Estas variables temporales están disponibles (aunque quizás de manera censurada) porque la base de datos contiene información sobre la fecha del primer transplante de médula ósea, la fecha del segundo transplante, y la fecha del último contacto o muerte. Al igual que en el caso de los datos de cáncer de colon, un modelo 'illness-death' es adecuado en este caso. Estos datos se utilizan en el Capítulo 4, donde las distintas probabilidades de transición se estiman y se muestran gráficamente.

Guión de la tesis

La tesis se organiza como sigue. En el Capítulo 2 introducimos un estimador semiparamétrico de la función de distribución conjunta de un par de 'gap times' posiblemente censurados. Se establece la consistencia de un funcional general basado en tal estimador (Sección 2.2). Se realiza un estudio de simulación (Sección 2.3) para investigar las propiedades del estimador propuesto en muestras finitas, cuando se compara con un estimador puramente no paramétrico. El estudio de simulación incluye el comportamiento de un estimador bootstrap del error estándar. La ilustración con el ejemplo de datos de cáncer de vejiga se da en la Sección 2.4. En la Sección 2.5 se establece una representación del estimador como suma de variables independientes e idénticamente distribuidas (i.i.d.) y, como consecuencia, se obtiene la normalidad asintótica del estimador. La prueba del resultado de consistencia se recoge en la Sección 2.6.

En el Capítulo 3 se propone un estimador semiparamétric de las probabilidades de transición en el modelo 'illness-death'. Al igual que en el Capítulo 2, se investigan las propiedades del estimador tanto teóricamente (consistencia, Sección 3.2) como a través de simulaciones (Sección 3.3). La Sección 3.4 está dedicada a la ilustración con los datos reales de cáncer de colon.

En el Capítulo 4 datos parte del código `R` que hemos desarrollado para implementar los métodos propuestos. Más específicamente, en la Sección 4.2 se proporciona el código `R` utilizado para obtener los resultados de las simulaciones de la Sección 2.3. En la Sección 4.3 se da un ejemplo simple (basado en un conjunto de datos simulado) del cómputo de los estimadores semiparamétricos de las probabilidades de transición en el modelo 'illness-death'. También damos el correspondiente código `R` aquí. Finalmente, en la Sección 4.4 estimamos las probabilidades de transición para los datos de leucemia, comparando los distintos estimadores no-markovianos alternativos.

El Capítulo 5 contiene las principales conclusiones de los distintos Capítulos de la tesis (Sección 5.1). También damos aquí algunos problemas abiertos que son interesantes para nuestra investigación futura (Sección 5.2).

Los resultados del Capítulo 2 (excepto por lo que se refiere a la Sección 2.5) están contenidos en la publicación de Uña-Álvarez and Amorim (2011), mientras que el Capítulo 3 es casi en su totalidad reproducido en Amorim et al. (2011).

A continuación damos, a modo de resumen, algunos de los contenidos de los Capítulos 2, 3 y 4.

Un estimador semiparamétrico de la distribución conjunta de dos 'gap times'

Como ya se ha dicho, el análisis estadístico de dos 'gap times' sucesivos es un problema de mucha importancia en un número de campos, incluyendo la ingeniería, la economía, la epidemiología, y el análisis de supervivencia. En la mayor parte de los casos, uno estará interesado en describir no sólo la distribución marginal de los 'gap times' sino también la estructura de correlación entre ellos. Esto ocurre, por ejemplo, cuando se analizan tiempos de recurrencia, que afloran cuando cada individuo puede experimentar un evendo bien definido varias veces a lo largo de su historia. Entonces, los tiempos entre eventos son referidos como los 'gap times', y están determinados por supuesto por los tiempos en los que la recurrencia tiene lugar (es decir, los tiempos de recurrencia). Véase Cook and Lawless (2007) para una revisión hasta la fecha de métodos estadísticos para datos de eventos recurrentes. En este Capítulo, el interés se centra en un par de 'gap times' sucesivos. En nuestro ejemplo con datos reales de la Sección 2.4, estos 'gap times' son el tiempo hasta la primera recurrencia y el tiempo desde la primera hasta la segunda recurrencia para pacientes con cáncer de vejiga. Para formalizar la discusión, introducimos ahora nuestra notación.

Sea (T_1, T_2) un par de 'gap times' de eventos sucesivos, que son observados sujetos a censura aleatoria por la derecha. Sea C la variable de censura por la derecha, que se asume independiente de (T_1, T_2), y sea $Y = T_1 + T_2$ el tiempo total. Debido a la censura, en lugar de (T_1, T_2) observamos $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$, donde $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$ y $\widetilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$, donde $C_2 = (C - T_1)\, I\,(T_1 \leq C)$ es la variable de censura para el segundo 'gap time'. Nótese que $\Delta_2 = 1$

implica $\Delta_1 = 1$. Por tanto, $\Delta_2 = \Delta_1\Delta_2 = I\left(Y \leq C\right)$ es el indicador de censura perteneciente al tiempo total. Ponemos $\widetilde{Y} = Y \wedge C$. Sea $\left(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right)$, $1 \leq i \leq n$, datos iid data con la misma distribución que $\left(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2\right)$. Debido a que el tiempo de censura se asume independiente del proceso, la distribución marginal del primer 'gap time' T_1 puede ser consistentemente estimada por el estimador de Kaplan-Meier basado en los $\left(\widetilde{T}_{1i}, \Delta_{1i}\right)$'s. Similarmente, la distribución del tiempo total puede ser consistentemente estimada por el estimador de Kaplan-Meier basado en los $\left(\widetilde{T}_{1i} + \widetilde{T}_{2i}, \Delta_{2i}\right)$'s. Sin embargo, T_2 y C_2 serán en general dependientes (a causa de la esperadas correlasción entre los 'gap times'), y por tanto la estimación de la distribución marginal del segundo 'gap time' no es un problema tan sencillo. También, no está claro en principio cómo la función de distribución bivariante $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$ puede ser eficientemente estimada. Este problema fue investigado, entre otros, por Wang and Wells (1998), Lin et al. (1999), Wang and Chang (1999), Peña et al. (2001), van der Laan et al. (2002), Schaubel and Cai (2004), Van Keilegom (2004), o de Uña-Álvarez and Meira-Machado (2008).

En este Capítulo proponemos un estimador semiparamétrico de la función de distribución bivariante de los 'gap times', $F_{12}(x, y)$. Para esto, asumimos que la probabilidad de censura para T_2 dados los (posiblemente censurados) 'gap times' pertenece a una familia paramétrica de curvas de regresión binarias. Es decir, siendo $m(x, y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y)$, se asume que $m(x, y)$ sigue algún modelo paramétrico. En la Sección 2.2 vemos que, en esencia, esto implica asumir un modelo paramétrico (suave) para $m_1(x, y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1)$. Nótese que, debido a que $\widetilde{T}_1$, $\widetilde{Y}$, Δ_1, y Δ_2 son observadas, esta suposición es contrastable en la práctica, véase e.g. Hosmer and Lemeshow (1989). En base a esta suposición paramétrica, somos capaces de introducir un nuevo estimador. Básicamente, el nuevo método usa una versión presuavizada del estimador Kaplan-Meier (véase e.g. Dikta (1998)) perteneciente a la distribución del tiempo total (la Y) para ponderar los datos bivariantes. En el caso límite de no presuavización, el estimador que proponemos se reduces al de de Uña-Álvarez and Meira-Machado (2008), el cual se mostró tener buenas propiedades. Sin embargo, la introducción de presuavizado paramétrico puede reducir grandemente la varianza en la estimación, particularmente en la cola derecha de la distribución (bivariante) o para censura pesada sobre T_2.

En la Sección 2.2, se establece la consistencia del estimador. El comportamiento del estimador en muestras finitas se investiga a través de simulaciones en la Sección 2.3. Los resultados de simulación también se utilizan para evaluar el comportamiento de un estimador bootstrap del error estándar. Una ilustración con datos reales se proporciona en la Sección 2.4, mientras que en la Sección 2.5 derivamos una representación asintótica del estimador útil para establecer un Teorema central del Límite. La prueba del resultado de consistencia se da en la Sección 2.6.

La idea de presuavizar el estimador de Kaplan-Meier fue introducida por Dikta (1998), quien denominó este método como 'modelo de censura semiparamétrico'. Véase también Dikta (2000,

2001) y Dikta et al. (2005). El presuavizado paramétrico con covariables fue considerado por de Uña-Álvarez and Rodríguez-Campos (2004), Yuan (2005), o Iglesias-Pérez and de Uña-Álvarez (2008). Todas estas referencias concluyen que los estimadores (semiparamétricos) presuavizados tienen varianza mejorada cuando se comparan con estimadores puramente no paramétricos. Aquí mostramos que el presuavizado es también útil para mejorar la eficiencia en el contexto multivariante de los 'gasp times'.

Probabilidades de transición presuavizadas en el modelo 'illness-death'

Los modelos multi-estado (Andersen et al. (1993); Meira-Machado et al. (2009)) son los modelos más comunmente utilizados para la descripción de datos longitudinales de supervivencia. Un modelo multi-estado es un modelo para un proceso estocástico, el cual está caracterizado por un conjunto de estados y las posibles transiciones entre ellos. Los estados representan diferentes situaciones del individuo (sano, enfermo, etc) a lo largo de un seguimiento. Modelos multi-estado particulares que han sido ampliamente utilizados en aplicaciones biomédicas son el modelo de tres estados progresivos, el modelo 'illness-death', o el modelo bivariante (Hougaard (2000)).

Sea $X(t)$ el estado ocupado por el processo en tiempo $t \geq 0$. Para dos estados i,j y $s < t$, introducimos la probabilidad de transición

$$p_{ij}(s,t) = P\left(X(t) = j | X(s) = i\right).$$

Ha habido mucho interés en la estimación de $p_{ij}(s,t)$ ya que permite hacer predicciones a largo plazo del proceso. Aalen and Johansen (1978) introdujo un estimador no paramétrico de $p_{ij}(s,t)$ para modelos markovianos. La condición de Markov establece que la evolución futura del proceso es independiente de los estados previamente visitados y de los tiempos de transición entre ellos, dado el estado presente del proceso. Esta suposición simplificadora permite la construcción de estimadores simples, ya que individups con diferentes historias pasadas se convierten en comparables. Sin embargo, se ha referenciado que la condición de Markov es violada en algunas aplicaciones (e.g. Andersen et al., 2000). Ésta es una anotación relevante, ya que el estimador Aalen-Johansen puede ser inconsistente si el proceso no es markoviano. Estimadores de $p_{ij}(s,t)$ que son consistentes en situaciones no markovianas escasean en la literatura.

Meira-Machado et al. (2006) introdujeron un sustituto para el estimador Aalen-Johansen en el caso de un modelo 'illness-death' no markoviano. Ellos mostraron que cuando la condición de Markov no es válida, el nuevo estimador puede comportarse mucho mejor que el Aalen-Johansen, que puede estar sistemáticamente sesgado. Sin embargo, al eliminar la condición de Markov, el sustituto del Aalen-Johansen propuesto proporciona errores estándar que son indeseablemente grandes. Este problema empeora cuando hay una proporción elevada de datos censurados. Para superar este problema, proponemos aquí una modificación del estimador de Meira-Machado et al. (2006) basado en presuavizado, lo cual permite una reducción de la varianza en presencia de censura.

Para ilustrar nuestros estimadores usando datos reales, consideramos datos de uno de los primeros ensayos clínicos exitoso en quimioterapia adyuvante para el cáncer de colon, los cuales están libremente disponibles en el paquete `survival` de `R`. En este estudio, 929 pacientes afectados de cáncer de colon se sometieron a una cirugía potencialmente curativa. Desafortunadamente, alguno de estos pacientes tuvieron cáncer residual, lo cual llevó a la recurrencia de la enfermedad y a la muerte (en algunos casos). Por lo tanto, podemos considerar la recurrencia como un estado asociado de riesgo, y utilizar el llamado modelo 'illness-death' con estados "vivo y libre de enfermedad", "vivo con recurrencia" (local-regional o metástasis) y "muerto". Véase la Sección 3.2 para una descripción más formal del modelo.

Código R y más ejemplos

En el Capítulo 4 proporcionamos el código `R` utilizado en el estudio de simulación del Capítulo 2. Esto incluye código `R` para el cómputo de errores cuadráticos medios de varios estimadores de la función de distribución conjunta de 'gap times', a lo largo de un número de ensayos de Monte Carlo. La lista de estimadores incluye el nuevo estimador semiparamétrico, su 'versión óptima' (basada en la verdadera función de presuavización), y el estimador basado en el Kaplan-Meier , cfr. de Uña-Álvarez and Meira-Machado (2008). Además, también se proporciona el código `R` necesario para el estudio del comportamiento del estimador bootstrap del error estándar. Toda esta información está contenida en la Sección 4.2.

En la Sección 4.3 también proporcionamos un ejemplo simple del cómputo de las probabilidades de transición presuavizadas en el modelo 'illness-death', tal y como se definen en el Capítulo 3. A este fin, simulamos una muestra de datos hipotéticos a partir de uno de los modelos descritos en la Sección 3.3, y computamos los estimadores presuavizados de las probabilidades de transición $p_{11}(s,t)$ y $p_{23}(s,t)$ para pares específicos (s,t). A efectos comparativos, los estimadores basados en el Kaplan-Meier en Meira-Machado et al (2006) también se evalúan.

Finalmente, en la Sección 4.4 analizamos los datos de leucemia introducidos en la Sección 1.2. Para estos datos, damos un número de gráficos representando las probabilidades de transición cuando se estiman con los pesos Kaplan-Meier ordinarios y también via presuavizado logístico. También proporcionamos un número de comentarios. Esta Sección 4.4 se recoge en el punto siguiente.

Datos de leucemia

En esta Sección proporcionamos algunos resultados de nuestro análisis de los datos de leucemia cedidos por el IPO, los cuales fueron brevemente descritos en la Sección 1.2. Recuérdese que usamos un modelo 'illness-death' para este conjunto de datos, donde el estado 1 representa el primer transplante, el estado 2 se reserva para el segundo transplante, mientras que el estado 3 representa la muerte del paciente. Los 251 datos se presentan en la Figura 4.1, donde se utilizan símbolos distintos de acuerdo con el estado de censura de cada individuo.

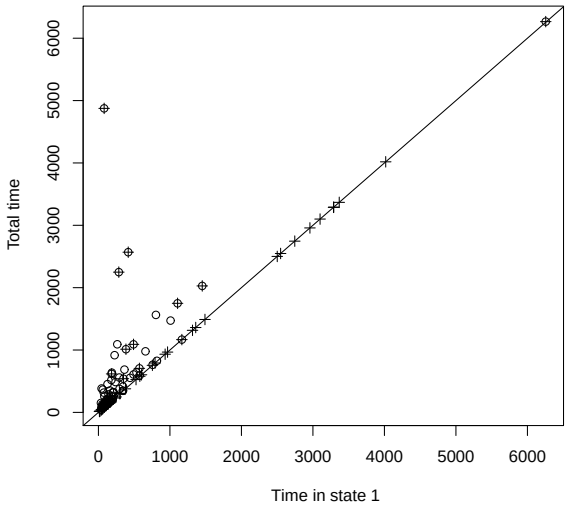

Figure 6.1: Datos de leucemia: pares no censurados ($\circ$), segundo 'gap time' censurado ($\oplus$), y ambos tiempos censurados ($+$).

El cómputo de las probabilidades de trasnsición semiparamétricas requiere la estimación preliminar de tres modelos paramétricos, uno para cada funciónde presuavizado involucrada en el problema. A este fin utilizamos un modelo logístico en todos los casos. Los resultados correspondientes al ajuste de estos modelos logísticos se dan en la Tabla 6.1. A partir de esta Tabla vemos que el impacto del tiempo total de supervivencia es estadísticamente significativo en los tres casos (tiempos de supervivencia observados mayores se corresponden con probabilidades más grandes de censura), mientras que el tiempo de permanencia en el estado 1 no es significativo para la función de presuavizado m_1. Los modelos ajustados se muestran en la Figura 6.2 junto con la información muestral en la cual se basan. En la Figura 6.2, derecha, los valores de la función de presuavización estimada m_1 parecen ser rugosos, como resultado de la influencia oculta (no significativa) del tiempo de permanencia en el estado 1.

La Figura 6.3 muestra las probabilidades de transición $p_{i,j}(s,t)$ cuando se estiman por los estimadores semiparamétricos propuestos en el Capítulo 3 o por el estimador no presuavizado propuesto por Meira-Machado et al. (2006). Como valores de s tomamos los tres cuartiles muestrales pertenecientes a $\widetilde{Z}$: 130, 335 y 1240 días. Cuando se comparan los dos estimadores, vemos que son casi idénticos para t próximo a s; sin embargo, comienzan a ser más diferentes a medida que t crece. Esto es debido a la redistribución de la masa asociada a los tiempos de transición censurados que se logra a través del estimador semiparamétrico.

Interesantemente, la primera fila en la Figura 6.3 sugiere que la probabilidad de tener una

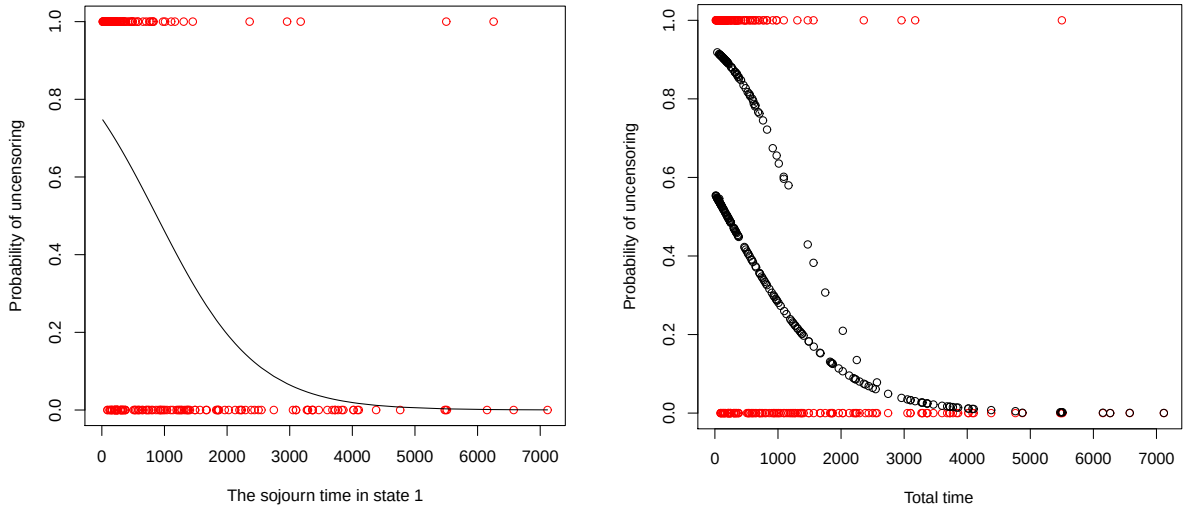

Figure 6.2: Funciones de presuavizado m_0 (izquierda), m_1 (arriba) y m_2 (abajo) estimadas por modelos logísticos. Datos de leucemia.

recaída decrece a medida que el tiempo pasa; lo mismo es cierto para la probabilidad de muerte (tercera fila). Similarmente, las dos últimas filas de la Figura 6.3 indican que el riesgo de muerte es mayor justo después de tener el segundo transplante, y que luego decrece con el tiempo. Los resultados numéricos aportados en la Tabla 6.2 también apoyan estos comentarios. Específicamente, de acuerdo con el estimador semiparamétrico, la probabilidad de permanecer en el estado inicial ("sano") en tiempo t=2000 días se incrementa un 87% cuando el tiempo que ha pasado desde el primer transplante (s) se incrementa de 130 a 1240 días. Respectivamente, la probabilidad de haber muerto en tiempo t=2000 días decrece un 21% 1240 días después del segundo transplante cuando se compara con 130 días después de esta segunda cirugía. Interesantemente, nótese que esta última información no puede obtenerse a partir del estimador basado en el Kaplan-Meier, que concentra su masa en los tiempos de transición no censurados.

Conclusiones

En esta tesis hemos introducido algunas estrategias de estimación semiparamétricas en el ámbito del modelo de tres estados progresivo no markoviano, y del modelo 'illness-death' no markoviano. Los estimadores propuestos hacen uso de ideas de 'presuavización', y este 'presuavizado' está dirigido por ciertos modelos semiparamétricos de censura específicos. Aunque los estimadores presuavizados son conocidos en el contexto clásico (es decir, univariante) del Análisis de Super-

Presmoothing function	Estimated Coefficients	p-value
$m_{0n}(z) = (1+\exp(\widehat{\eta}_0 + \widehat{\eta}_1 z))^{-1}$	$\widehat{\eta}_0 = 1.1016 \quad (0.1926)$	0.0000
	$\widehat{\eta}_1 = -0.0013 \quad (0.0002)$	0.0000
$m_{1n}(z,t) = (1+\exp(\widehat{\beta}_0 + \widehat{\beta}_1 z + \widehat{\beta}_2 t))^{-1}$	$\widehat{\beta}_0 = 2.4920 \quad (0.5102)$	0.0000
	$\widehat{\beta}_1 = 0.0000 \quad (0.0015)$	0.9554
	$\widehat{\beta}_2 = -0.0019 \quad (0.0009)$	0.0391
$m_{2n}(z) = (1+\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 z))^{-1}$	$\widehat{\gamma}_0 = 0.2333 \quad (0.2366)$	0.3240
	$\widehat{\gamma}_1 = -0.0012 \quad (0.0003)$	0.0000

Table 6.1: Resumen de las tres funciones de presuavización m_{0n}, m_{1n} y m_{2n} basadas en modelos logísticos. Datos de leucemia.

$(s,t) =$	$(130, 2000)$	$(335, 2000)$	$(1240, 2000)$
$\widehat{p}_{11}^{pkm}(s,t)$	0.4751	0.6273	0.8882
$\widehat{p}_{11}^{km}(s,t)$	0.5760	0.7320	0.9671
$\widehat{p}_{23}^{pkm}(s,t)$	0.9992	0.9714	0.7847
$\widehat{p}_{23}^{km}(s,t)$	1.0000	1.0000	1.0000

Table 6.2: Probabilidades de transición estimadas con y sin presuavizado. Datos de leucemia.

vivencia, por lo que a nuestro conocimientor respecta su aplicación en el complicado contexto de los modelos multi-estado es nueva. Sólo por citar un problema específico que necesita ser solucionado, cuando se manejan tiempos de supervivencia multivariantes aparecerán funciones de presuavización de carácter discontinuo.

Más explícitamente, en el Capítulo 2 se introduce un nuevo estimador semiparamétrico $\widehat{F}_{12}^{sp}(x,y)$ de la función de distribución bivariante de 'gap times' que son observados bajo censura. El estimador semiparamétrico se basa en una especificación paramétrica de la probabilidad condicional de censura para el segundo 'gap time' T_2, dada la información disponible. Esta especificación puede ser contrastada en la práctica. Hemos derivado la consistencia del estimador propuesto y, con más generalidad, de un funcional empírico basado en él. Hemos verificado a través de simulaciones que el estimador semiparamétrico puede ser mucho más eficiente que otros estimadores disponibles. Esto será particularmente cierto en puntos en los cuales existe una proporción elevada de valores de T_2 censurados entre los casos con primer 'gap time' no censurado. Además, hemos visto que el método es robusto ante malas especificaciones del modelo paramétrico. Hemos también utilizado el bootstrap simple para aproximar el error estándar del estimador, y nuestros resultados de simulación sugieren que el bootstrap proporciona una estimación insesgada. Se ha proporcionado una ilustración con datos reales. Finalmente, se ha dado una representación del estimador como suma

de variables aleatorias i.i.d., y consecuentemente se ha establecido su normalidad asintótica.

En el Capítulo 3 hemos introducido nuevos estimadores semiparamétricos para las probabilidades de transición de un modelo 'illness-death' no-markoviano censurado. Los nuevos estimadores se basan en varios modelos paramétricos para distintas funciones de 'presuavización', las cuales varían dependiendo de los estados implicados. El comportamiento no asintótico de los estimadores propuestos ha sido investigado a través de simulaciones. Como en el Capítulo 2, la principal conclusión del Capítulo 3 es que la presuavización lleva a estimadores más precisos, incluso cuando existe cierta mala especificación en la familia paramétrica asumida para la función de presuavización. Los beneficios relativos de la presuavización se ven más claramente en casos fuertemente censurados. El nuevo método ha sido ilustrado utilizando datos de un estudio sobre cáncer de colon, y ha sido utilizado para analizar los datos de leucemia proporcionados por el IPO (Sección 4.4).

Los nuevos estimadores para las probabilidades de transición son consistentes independientemente de la condición de Markov (esto es también cierto para el estimador propuesto en el Capítulo 2). Esto es interesante ya que los problemas reales se encuentran frecuentemente alejados de la markovianidad y, por tanto, la consistencia del estimador de Aalen-Johansen no puede ser garantizada. A este respecto, uno puede pensar que los métodos introducidos aquí son una mejora notable (en el sentido de tener menos varianza) de estimadores no-markovianos previamente existentes (Meira-Machado et al. (2006)).

En la práctica, la implementación de los métodos propuestos está lejos de ser sencilla. En el Capítulo 4 hemos realizado en detalle un ejemplo describiendo los diferentes pesos que se necesitan para computar las funciones de presuavización y los distintos estimadores. También, en las Secciones 4.2 y 4.3 hemos proporcionado nuestro propio código `R`, que permite reproducir los distintos análisis realizados en esta tesis y, más importante, computar el estimador para nuevos conjuntos de datos reales. Creemos que ésta es una importante contribución para los investigadores aplicados.

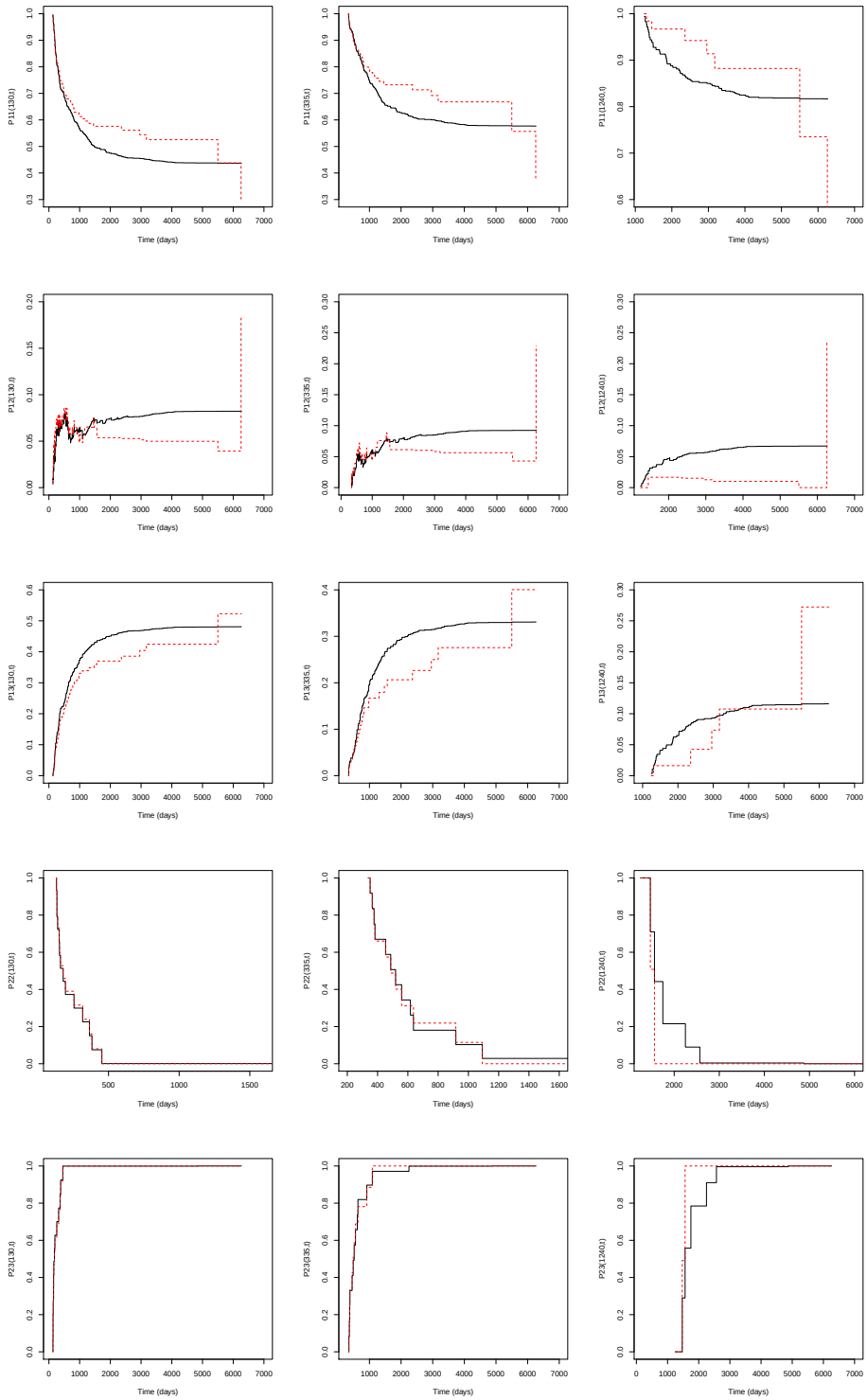

Figure 6.3: Probabilidades de transición estimadas $p_{ij}(s,t)$ con $s \in \{130, 335, 1240\}$ basadas en los pesos Kaplan-Meier (línea discontinua) y basadas en los pesos Kaplan-Meier presuavizados (línea continua). Datos de leucemia.

Chapter 7

Bibliography

Bibliography

Aalen, O. and S. Johansen (1978). An empirical transition matrix for nonhomogeneous markov and chains based on censored observations. *Scandinavian Journal of Statistics 5*, 141–150.

Akritas, M. (1986). Bootstrapping the kaplan-meier estimator. *Journal of the American Statistical Association 81*, 1032–1038.

Amorim, A. P., J. de Uña-Álvarez, and L. Meira-Machado (2011). Presmoothing the transition probabilities in the illness-death model. *Statistics & Probability Letters 81*, 797–806.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Andersen, P. K., S. Esbjerq, and T. I. Sorensen (2000). Multistate models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine 19*, 587–599.

Andersen, P. K. and N. Keiding (2002). Multi-state models for event history analysis. *Statistical Methods Medical Research 11*, 91–115.

Byar, D. P. (1980). Veterans administration study of chemoprophylaxis for recurrent stage i bladder tumors: Comparisons of placebo, pyridoxine and topical thiotepa. *In:M. Pavone-Macaluso, P.H. Smith, and Edsmyn, F. (Eds.), Bladder Tumors and Other Topics in Urological Oncology.Plenum, New York 36*, 363–370.

Cao, R. and M. A. Jácome (2004). Presmoothed kernel density estimator for censored data. *Journal of Nonparametric Statistics 16*, 289–309.

Cao, R., I. López de Ullibarri, P. Janssen, and N. Veraverbeke (2005). Presmoothed kaplan-meier and nelson-aalen estimators. *Journal of Nonparametric Statistics 17*, 31–56.

Cheng, P. E. and G. D. Lin (1987). Maximum lilkelihood estimation of a survival function under the koziol-green proportional hazards model. *Statistics & Probability Letters 5*, 75–80.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis 5*, 315–327.

Cook, R. J. and J. F. Lawless (2007). *The Analysis of Recurrent Event Data*. New York: Springer.

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

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

de Uña-Álvarez, J. and C. Rodríguez-Campos (2004). Strong consistency of presmoothed kaplan-meier integrals when covariables are present. *Statistics 38*, 483–496.

Devroye, L. P. (1978a). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory 24*, 142–151.

Devroye, L. P. (1978b). The uniform convergence of the nadaraya-watson regression function estimate. *Canadian Journal of Statistics 6*, 179–191.

Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference 66*, 253–279.

Dikta, G. (2000). The strong law under semiparametric random censorship models. *Journal of Statistical Planning and Inference 83*, 1–10.

Dikta, G. (2001). Weak representation of the cumulative hazard function under semiparametric random censorship models. *Statsitics 35*, 395–409.

Dikta, G., J. Ghorai, and C. Schmidt (2005). The central limit theorem under semiparametric random censorship models. *Journal of Statistical Planning and Inference 127*, 23–51.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics 9*, 139–172.

Hosmer, D. W. and S. Lemeshow (1989). *Applied Logistic Regression*. NewYork: Wiley.

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

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer.

Härdle, W. and S. Luckhaus (1984). Uniform consistency of a class of regression function estimators. *Annals of Statistics 12*, 612–623.

Iglesias-Pérez, M. C. and J. de Uña-Álvarez (2008). Nonparametric estimation of the conditional distribution function in a semiparametric censorship model. *Journal of Statistical Planning and Inference 138*, 3044–3058.

Kay, R. (1986). A markov model for analysing cancer makers and desease states in survival studies. *Biometrics 42*, 855–865.

Klein, J. P. and M. L. Moeschberger (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

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

Mack, Y. P. and B. W. Silverman (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 61*, 405–415.

Meira-Machado, L., J. de Uña-Álvarez, and C. Cadarso-Suárez (2006). Nonparametric estimation of transition probabilities in a non-markov illness-death model. *Lifetime Data Analysis 12*, 325–344.

Meira-Machado, L., J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research 18*, 195–222.

Moertel, C. G., T. R. Fleming, J. S. McDonald, and et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New Engl. Journal of Medicine 322*, 352–358.

Neveu, J. (1975). *Discrete-parameter Martingales*. Amsterdam/Oxford: North-Holland.

Peña, E. A., R. L. Strawderman, and M. Hollander (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association 96*, 1299–1315.

R-Development-Core-Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schaubel, D. E. and J. Cai (2004). Non-parametric estimation of gap-time survival functions for ordered multivariate failure time data. *Statistics in Medicine 23*, 1885–1900.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis 45*, 89–103.

Stute, W. (1994). U-statistic processes: a martingale approach. *Annals of Probability 22*, 1725–1744.

Stute, W. (1995). The central limit theorem under random censorship. *Scandinavian Journal of Statistics 23*, 422–439.

Stute, W. and J. L. Wang (1993). The strong law under random censorship. *Annals of Statistics 21*, 1591–1607.

van der Laan, M. J., A. E. Hubbard, and J. M. Robins (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *Journal of the American Statistical Association 97*, 494–507.

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

Wang, M. C. and S. H. Chang (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association 94*, 146–153.

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

Wei, L. J., D. Y. Lin, and L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association 84*, 1065–1073.

Yuan, M. (2005). Semiparametric censorship model with covariates. *Test 14*, 489–514.