INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1053

ESTIMATION OF TRANSITION PROBABILITIES IN A NON-MARKOV MODEL WITH SUCCESSIVE SURVIVAL TIMES

MEIRA-MACHADO, L., ROCA-PARDINAS, J., VAN KEILEGOM, I. and C. CADARSO-SUAREZ

Estimation of transition probabilities in a non-Markov model with successive survival times

Luís Meira-Machado

Department of Mathematics and Applications, University of Minho, Portugal

Javier Roca-Pardiñas

Department of Statistics and Operation Research, University of Vigo, Spain

Ingrid Van Keilegom

Institut de statistique, Université catholique de Louvain, Belgium

Carmen Cadarso-Suárez

Unit of Biostatistics, Department of Statistics and Operation Research, University of Santiago de Compostela, Spain

Abstract

Times between consecutive events are often of interest in medical studies. Usually the events represent different states of the disease process and are modeled using multi-state models. This paper introduces and studies a feasible estimation method for the transition probabilities in a progressive three-state model. We assume that the vector of gap times (T_1, T_2) satisfies a nonparametric location-scale regression model $T_2 = m(T_1) + \sigma(T_1)\epsilon$, where the functions m and σ are 'smooth', and ϵ is independent of T_1. Under this model, Van Keilegom, de Uña-Álvarez and Meira-Machado (2011) proposed estimators of the transition probabilities. In this paper, we study the performance of their estimator in practice, we propose some modifications and study practical issues related to the implementation of the estimator. In an extensive simulation study the good performance of the method is shown. Simulations also demonstrate that the proposed estimator compares favor-

Email addresses: lmachado@math.uminho.pt (Luís Meira-Machado), roca@uvigo.es (Javier Roca-Pardiñas), ingrid.vankeilegom@uclouvain.be (Ingrid Van Keilegom), carmen.cadarso@usc.es (Carmen Cadarso-Suárez)

ably with alternative estimators. Furthermore, the proposed methodology is illustrated with a real database on breast cancer.

Keywords: Censoring; computational statistics; location-scale model; nonparametric regression; progressive three-state model; transition probabilities.

---

1. Introduction

In many longitudinal studies, patients experience sequentially ordered events. The events may be of the same nature (e.g. cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g. alive and disease-free, alive with recurrence and dead). If the events are of the same nature, they are usually referred to as recurrent events and are modeled as such, whereas if they represent different states, they are usually modeled through their intensity functions via so-called multi-state models. A multi-state model (Andersen et al. 1993; Hougaard 2000) is a model for a stochastic process that at any point in time may occupy one state among a discrete set of states.

The times between two consecutive events (gap times) are often of interest and lead to problems of great interest. These include the estimation of several functions such as the bivariate distribution function, the conditional distribution and the transition probabilities. In this paper we focus on the estimation of the transition probabilities in a progressive three-state model. These quantities provide useful means of characterizing and measuring the disease progression. Traditionally the transition probabilities are estimated via a nonparametric model (using e.g. the Aalen–Johansen estimator; Aalen and Johansen, 1978). However, Meira-Machado et al. (2006) verified that in non-Markov situations, the use of the Aalen-Johansen estimator to empirically estimate the transition probabilities, may be inappropriate. In the latter paper alternative estimators in the scope of the illness-death model are proposed, which do not rely on the Markov assumption. In the presence of heavy censoring, the estimation of these quantities in a completely nonparametric way will be hard, especially in the right tail. In an attempt to solve the inconsistency problems in the right tail, Van Keilegom et al. (2011) propose alternative 'Markov-free' estimators for the transition probabilities. The authors assume a nonparametric location-scale regression model to transfer tail information from light censored regions (small lifetimes) to regions with

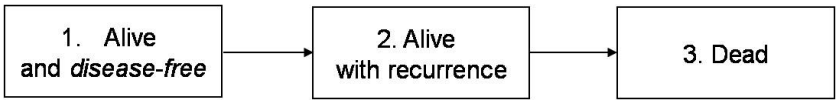

Figure 1: Progressive three-state model for the Galician breast cancer data

heavy censoring. The location-scale model assumes that the relation between the gap times T_1 and T_2 is given by $T_2 = m(T_1) + \sigma(T_1)\epsilon$, where the functions m and σ are 'smooth', and ϵ is independent of T_1. On the basis of the idea of transfer of tail information, the estimator of the error distribution is used to introduce nonparametric estimators of the transition probabilities.

In this paper, we focus on some practical issues related to the computation and the implementation of the proposed estimator, building on the work of Van Keilegom et al. (2011), who laid down the theoretical framework of their estimator. We describe the method in detail, and discuss its implementation in practice, which involves among others the selection of an appropriate bandwidth and the construction of pointwise confidence bands by means of a bootstrap approach.

The present paper is motivated by an application to a prospective study on breast cancer from Galicia, Spain. This study was conducted at the Hospital Clinico Universitario de Santiago (Santiago, Spain) to assess the value of DNA index and S-phase fraction evaluated by flow cytometry as prognostic markers in invasive breast carcinomas. In this study 584 incident cases of breast cancer were diagnosed in the period between April 1991 and December 2000. Out of these 584 patients, 172 relapsed (recurrence) and among these 114 died. Under the gap time's framework, T_1 is the time from randomization to cancer recurrence and T_2 is the time from cancer recurrence to death. These times are both subject to random right-censoring. Therefore, we may use the progressive three-state model with states 'alive and disease-free', 'alive with recurrence' and 'dead' (see Figure 1). Eleven patients died without relapse. These patients are treated as censored on the recurrence transition and they are not considered on the mortality transition from the 'alive with recurrence' state. The rest of the patients remained alive and disease-free up to the end of the follow-up.

Figure 2 displays the scatter plot of the pair of gap times (T_1, T_2) for the breast cancer study. We can see that the time to recurrence, T_1, is more often censored in the time interval between 2 and 6 years. We have a few (four) complete (i.e. uncensored) observations for patients with time to recurrence

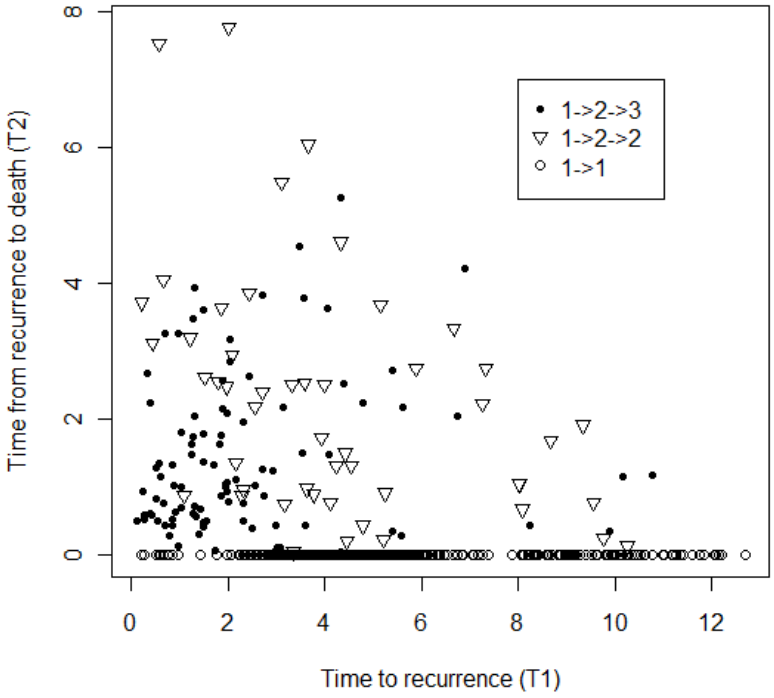

Figure 2: Scatter plot of the pair of successive events for the Galician breast cancer data.

greater than 7 years. Similarly, we have a few complete observations with time to death since recurrence, T_2, greater than 4 years. This means that it is expected that any estimator for the transition probabilities will behave poorly in these areas (the right tail of the distribution).

One main point of interest in breast cancer studies is the diagnosis at a sufficiently early stage of the disease. Thus, it is important to make long-term predictions and to identify possible times of diagnosis (threshold values). Therefore, it is very important to obtain good estimates for the transition probabilities. Multi-state models are usually assumed to be Markovian. This assumption states that the future of the process depends on the history only through the present, thus being independent of the times of previous transitions. However, by ignoring the disease history, Markov models may have severe limitations, and may be unsuitable for a number of applications. This

is the case for the breast cancer dataset for which there is a strong (negative) effect of time since recurrence on the mortality transition (Cadarso et al. 2010). In these cases the use of estimators that do not rely on the Markov assumption is preferable. This is the case for the estimators proposed here. On the other hand, as mentioned above (see Figure 2), the estimator should be efficient in the right tail (for larger values of time to recurrence, T_1, and also for larger values of the time since recurrence, T_2). The idea of transfer of tail information, which is at the basis of the proposed estimator, will help to improve the estimation in the right tail of T_2.

The paper is organized as follows. In the next section, some notations are introduced and the estimators of the transition probabilities are presented. In Section 3, an extensive simulation study is performed to compare our method with alternatives. Some considerations about the choice of the kernel function and about the optimal bandwidth are given in Section 4. The analysis of the Galician breast cancer data is presented in Section 5, where we also discuss some practical computational issues. Finally, we conclude with a discussion section.

2. The Method

Let (T_1, T_2) be a pair of gap times corresponding to two consecutive events. As usual, assume that T_1 and T_2 are subject to random right censoring, and that the vector (T_1, T_2) is independent of the censoring time C. Note that both times are subject to right censoring and in particular the second gap time T_2 is subject to right censoring by $C_2 = (C - T_1)I(T_1 \leq C)$. This means that instead of observing (T_1, T_2), we only observe $\left(\tilde{T}_1, \Delta_1, \tilde{T}_2, \Delta_2\right)$, where $\tilde{T}_1 = \min(T_1, C)$, $\Delta_1 = I\left(T_1 \leq C\right)$, $\tilde{T}_2 = \min(T_2, C_2)$ and $\Delta_2 = I\left(T_2 \leq C_2\right)$. The data consist of independently and identically distributed (i.i.d.) copies

$$\left(\tilde{T}_{1i}, \Delta_{1i}, \tilde{T}_{2i}, \Delta_{2i}\right), \quad 1 \leq i \leq n$$

with the same distribution as $\left(\tilde{T}_1, \Delta_1, \tilde{T}_2, \Delta_2\right)$.

Due to the independence assumption between C and (T_1, T_2), the marginal distribution of the first gap time T_1 can be consistently estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958) based on the $\left(\tilde{T}_{1i}, \Delta_{1i}\right)$'s. Similarly, the distribution of the total time may be consistently estimated

by the Kaplan-Meier estimator based on the $\left(\tilde{T}_{1i} + \tilde{T}_{2i}, \Delta_{2i}\right)$'s. Recently, Meira-Machado et al. (2006) introduced nonparametric estimators for the transition probabilities, based on the Kaplan-Meier weights of the distribution of the first gap time and the total time. In the presence of heavy censoring, these estimators are in general inconsistent in the right tail of the distribution. Such problems are very common with survival data and can also appear when estimating e.g. the bivariate distribution of (T_1, T_2) and the conditional distribution of T_2 given T_1. This situation has been discussed by Van Keilegom and Akritas (1999), where T_2 denotes a possible transformation of the variable of interest (subject to censoring) and T_1 is a covariate (not subject to censoring). In an attempt to solve the inconsistency problems in the right tail, the authors assumed a nonparametric location-scale regression model to transfer tail information from light censored regions to heavily censored ones. Recently, Van Keilegom et al. (2011), applied the idea of transferring tail information to a pair of gap times (T_1, T_2) (that are both subject to censoring), showing that the proposed estimators behave well in the right tail even under heavy censoring. Among other quantities the authors propose new estimators of the transition probabilities. To make this transfer possible, it is assumed that (T_1, T_2) follow the heteroscedastic regression model

$$T_2 = m(T_1) + \sigma(T_1)\epsilon, \tag{1}$$

where the error variable ϵ is independent of T_1, $m(\cdot)$ is a location functional and $\sigma(\cdot)$ is a scale functional. These functionals are given by

$$m(x) = \int_0^1 F^{-1}(s|x)J(s)ds \text{ and } \sigma^2(x) = \int_0^1 F^{-1}(s|x)^2 J(s)ds - m(x)^2, \tag{2}$$

where $F(y|x) = P(T_2 \leq y|T_1 = x)$, $F^{-1}(s|x) = \inf\{t : F(t|x) \geq s\}$ is the quantile function of T_2 given $T_1 = x$ and J is a score function satisfying $\int_0^1 J(s)ds = 1$. Model (1) implies that

$$F(y|x) = F_e\left(\frac{y - m(x)}{\sigma(x)}\right), \tag{3}$$

where F_e denotes the distribution of the error variable ϵ.

To estimate the functionals given in (2) we use an extension of the Beran

(1981) estimator, which copes with censoring in the first gap time:

$$\tilde{F}(y|x) = 1 - \prod_{T_{2i} \leq y, \Delta_{2i}=1} \left[1 - \frac{B_{ni}(x; a_n)}{\sum_{j=1}^{n} B_{nj}(x; a_n) I(T_{2j} \geq T_{2i})} \right] \tag{4}$$

with

$$B_{ni}(x; a_n) = \frac{\Delta_{1i} K\left((x - T_{1i})/a_n\right)}{\sum_{j=1}^{n} \Delta_{1j} K\left((x - T_{1j})/a_n\right)}$$

and K is a known probability density function (kernel) and a_n a sequence of bandwidths. Then, we define

$$\hat{m}(x) = \int_0^1 \tilde{F}^{-1}(s|x) J(s) ds \text{ and } \hat{\sigma}^2(x) = \int_0^1 \tilde{F}^{-1}(s|x)^2 J(s) ds - \hat{m}(x)^2.$$

Now, let

$$\hat{E}_i = \frac{\tilde{T}_{2i} - \hat{m}(\tilde{T}_{1i})}{\hat{\sigma}(\tilde{T}_{1i})}$$

and let $\hat{F}_e$ be the Kaplan-Meier estimator of F_e based on the $(\hat{E}_i, \Delta_{2i})$'s:

$$\hat{F}_e(y) = 1 - \prod_{\hat{E}_i \leq y, \Delta_{2i}=1} \left[1 - \frac{1}{n - i + 1} \right].$$

The estimator $\hat{F}_e$, together with relation (3), is the key for the construction of an estimator of $F(y|x)$:

$$\hat{F}(y|x) = \hat{F}_e \left(\frac{y - \hat{m}(x)}{\hat{\sigma}(x)} \right).$$

Consider now the problem of estimating the transition probabilities $p_{hj}(s, t)$. These quantities represent the probability of an individual being in state j at time t conditional on being in state h at time s. The transition probability is defined for a stochastic process that at any point in time may occupy one state among a discrete set of states. This includes, recurrent events data (or gap times), which may be seen as arising from a progressive k-state model (a special case of multi-state model). Under this framework, for the pair of gap times (T_1, T_2), we have three transition probabilities to estimate:

$$p_{11}(s,t) = (1 - F_1(t))/(1 - F_1(s)),$$

$$p_{12}(s,t) = \frac{1}{1 - F_1(s)} \int_s^t [1 - F(t-u|u)]\, F_1(du),$$

$$p_{22}(s,t) = \frac{\int_0^s [1 - F(t-u|u)]\, F_1(du)}{\int_0^s [1 - F(s-u|u)]\, F_1(du)},$$

where $F_1(\cdot)$ is the marginal distribution of the first gap time, which we may estimate by the Kaplan-Meier estimator based on the $(\tilde{T}_{1i}, \Delta_{1i})$'s:

$$\hat{F}_1(t) = 1 - \prod_{\tilde{T}_{1i} \le t, \Delta_{1i}=1} \left[1 - \frac{1}{n-i+1}\right].$$

Now, replace F_1 by $\hat{F}_1$ and $F(y|x)$ by $\hat{F}(y|x)$ to get the following estimators:

$$\hat{p}_{11}(s,t) = \Big(1 - \hat{F}_1(t)\Big)/\Big(1 - \hat{F}_1(s)\Big),$$

$$\hat{p}_{12}(s,t) = \frac{1}{1 - \hat{F}_1(s)} \int_s^t \Big[1 - \hat{F}(t-u|u)\Big]\, \hat{F}_1(du),$$

$$\hat{p}_{22}(s,t) = \frac{\int_0^s \Big[1 - \hat{F}(t-u|u)\Big]\, \hat{F}_1(du)}{\int_0^s \Big[1 - \hat{F}(s-u|u)\Big]\, \hat{F}_1(du)}.$$

We note that $\hat{p}_{11}(s,t)$ is equivalent to the estimator proposed by Meira-Machado et al. (2006) (denoted by M-M hereafter), and also equivalent to the Aalen-Johansen estimator of $p_{11}(s,t)$. Furthermore, in the framework of the progressive three-state model (Figure 1) the transition probabilities to be estimated reduce to $p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{22}(s,t)$, since $p_{13}(s,t) = 1 - p_{11}(s,t) - p_{12}(s,t)$ and $p_{23}(s,t) = 1 - p_{22}(s,t)$.

3. Simulation Study

In this section the results of a simulation study are presented, to assess the behavior of the proposed estimators for finite sample sizes.

Given the first gap time T_1 the second gap time T_2 was generated according to model (1) with $m(T_1) = 0.5 + 4\exp(-0.1T_1)$, $\sigma(T_1) = 1.5 - 0.5T_1/6$ and $\epsilon \sim N(0,1)$. The censoring variable C was drawn independently from a Uniform$[0, a]$. Note that the constant a determines the expected proportion of censored responses. We have chosen several values for a, in such a way that the censoring percentages of the total time $Y = T_1 + T_2$ equal 0%, 10.8% (4.9% and 5.9% for the first and second gap times respectively), 21.6% (10% and 11.6 %), 43.3% (20% and 23.4%), 64.9% (29.8% and 35.0%). In all cases, T_1 was chosen to be a uniform random variable on the interval $[0, 6]$, being independent of the error ϵ and the censoring time C. The model behavior was evaluated using one thousand independent samples $\left\{(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}); i = 1, \ldots, n\right\}$ generated from model (1), with different sample sizes: $n = 50$, $n = 100$, $n = 150$ and $n = 300$. The model was evaluated in the following points (s, t): (1,2), (1,3.5), (1,5), (2.5,3), (2.5,4), (2.5,5), (4,4.5), (4,4.75) and (4,5). The true values of $p_{hj}(s, t)$ for all considered values of h, j, s and t are reported in Table 1. In order to evaluate the performance of the transition probability functions p_{11}, p_{12} and p_{22} we examined respectively the mean squared errors MSE_{11}, MSE_{12} and MSE_{22}, defined as follows:

$$MSE_{hj} = (1/1000) \sum\nolimits_{i=1}^{1000} \left(\hat{p}_{hj}^{(i)}(s, t) - p_{hj}(s, t)\right)^2,$$

where $\hat{p}_{hj}^{(i)}(s, t)$ is the estimator obtained from sample i.

	(s,t)								
	(1,2)	(1,3.5)	(1,5)	(2.5,3)	(2.5,4)	(2.5,5)	(4,4.5)	(4,4.75)	(4,5)
$p_{11}(s,t)$	0.8000	0.500	0.2000	0.8571	0.5714	0.2857	0.7500	0.6250	0.5000
$p_{12}(s,t)$	0.1966	0.4736	0.6614	0.1407	0.4159	0.6611	0.2459	0.3677	0.4882
$p_{22}(s,t)$	0.9763	0.8167	0.4498	0.9678	0.8349	0.5993	0.9138	0.8609	0.8017

Table 1: True values of the transition probabilities $p_{hj}(s, t)$ $(1 \leq h \leq j \leq 2)$ for different values of (s, t) under the simulated model.

The performance of the location-scale estimator (LS; Van Keilegom et al. 2011) is compared with the estimator proposed by Meira-Machado et al. (2006).

n	$P(\Delta_2 = 0)$	(s,t)								
		(1,2)	(1,3.5)	(1,5)	(2.5,3)	(2.5,4)	(2.5,5)	(4,4.5)	(4,4.75)	(4,5)
50	64.9	7.2	21.6	41.4	19.1	51.3	79.8	110.7	146.9	185.6
	43.3	4.6	14.4	25.6	10.5	34.6	49.5	50.3	73.8	97.3
	21.6	1.6	6.4	11.1	5.0	12.4	21.1	18.4	29.7	41.9
	10.8	0.7	3.3	6.5	2.2	6.7	11.8	9.0	15.6	21.4
	0	0.2	1.5	3.8	0.2	2.2	6.1	2.2	4.9	8.7
100	64.9	4.0	12.7	13.8	9.9	21.5	26.9	37.6	53.8	68.1
	43.3	2.7	7.8	7.7	5.6	11.7	15.1	19.8	31.3	35.3
	21.6	1.5	3.1	3.2	2.5	4.9	6.7	9.9	14.9	16.6
	10.8	0.8	1.5	1.6	1.0	2.4	3.3	4.4	6.4	8.6
	0	0.2	0.4	0.2	0.2	0.1	0.5	0.6	1.3	2.3
150	64.9	1.9	7.4	10.3	5.9	12.7	20.5	36.3	49.1	53.4
	43.3	1.2	4.8	5.2	3.2	8.4	10.2	17.5	24.8	26.0
	21.6	0.5	2.0	2.0	1.6	3.6	4.1	7.7	10.8	10.4
	10.8	0.3	1.0	0.9	0.8	1.6	1.9	3.9	6.2	4.6
	0	0.0	0.2	0.0	0.2	0.1	0.0	1.0	2.3	0.0
300	64.9	1.2	3.3	4.6	2.7	6.3	9.2	14.2	21.2	24.8
	43.3	0.8	2.1	2.7	1.6	3.6	5.3	8.6	11.5	14.3
	21.6	0.3	0.9	1.1	0.8	1.8	2.2	3.6	4.8	5.4
	10.8	0.2	0.4	0.5	0.4	0.8	1.0	1.7	2.6	2.6
	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0

Table 2: Simulation-based averages of the MSE's ($\times$10000) of the estimated transition probabilities $\hat{p}_{11}(s,t)$ along 1000 replications for different sample sizes, and different censoring percentages.

Tables 2, 3 and 4 summarize, for the sample sizes considered, numerical average results of the considered mean squared errors (MSE) over 1000 replicated samples according to each model of the corresponding scenario. The results are presented in terms of the average of the MSE over the 1000 simulated samples. From these tables, we obtain the following conclusions. As one would expect the simulations show a good behavior of the proposed estimator (LS) when compared to the M-M estimator. In both cases the MSE's increase with increasing probability of censoring and decrease with increasing sample size.

Graphical average results for the LS method are displayed in Figures 3 and 4. These figures plot the data generating functions and pointwise 95% oscillation limits of the estimates $\hat{p}_{12}$ and $\hat{p}_{22}$ along the simulations, for percentages of censored data of 0%, 21.6% and 64.9%. The good performance of the resulting estimates of p_{12} is evident, recovering the functional forms of the corresponding true curves very successfully. The use of transfer of tail information for the estimation of p_{22} reveals to be somewhat more difficult. However, in both cases our approach yields estimators with small variability.

In Figures 5 and 6 the M-M estimators and pointwise 95% oscillation limits of the estimates $\hat{p}_{12}$ and $\hat{p}_{22}$ are plotted, respectively. Note that these figures can be compared with Figures 3 and 4, respectively. From these figures we observe that the variability of the M-M estimator is clearly larger than that of the LS estimator.

n	$P(\Delta_2 = 0)$		(s,t) (1,2)	(1,3.5)	(1,5)	(2.5,3)	(2.5,4)	(2.5,5)	(4,4.5)	(4,4.75)	(4,5)
50	64.9	M-M	68.8	163.2	246.0	153.9	381.9	515.1	352.8	608.5	860.8
		LS	7.2	23.5	47.1	18.4	53.0	88.6	106.1	144.5	187.9
	43.3	M-M	28.3	57.4	99.2	46.5	116.5	171.0	136.1	195.7	265.3
		LS	4.7	15.8	29.2	10.4	34.7	50.8	48.4	72.2	97.0
	21.6	M-M	9.3	23.4	50.4	13.3	43.8	68.0	48.2	73.9	100.4
		LS	1.6	6.8	16.4	4.9	11.7	21.3	17.8	28.5	40.5
	10.8	M-M	4.2	12.9	34.7	6.6	17.8	34.5	19.2	27.5	38.8
		LS	0.7	3.5	12.3	2.1	5.8	11.4	8.2	14.4	19.9
	0	M-M	0.2	4.8	23.8	0.6	3.7	16.1	2.3	4.3	8.9
		LS	0.1	1.1	11.0	0.1	1.1	5.5	1.5	3.2	5.3
100	64.9	M-M	37.6	94.9	133.4	58.7	148.8	227.3	214.2	363.8	496.9
		LS	4.0	13.7	16.5	9.6	20.1	29.3	49.4	60.4	72.0
	43.3	M-M	16.4	29.5	39.5	18.8	52.0	69.8	61.4	87.2	113.8
		LS	2.7	9.5	11.8	5.4	10.9	18.0	29.9	38.6	39.4
	21.6	M-M	5.3	12.0	19.4	6.4	16.2	23.4	20.6	30.3	38.5
		LS	1.8	3.9	7.3	2.4	5.1	8.3	17.6	19.2	17.6
	10.8	M-M	3.0	6.7	13.7	3.2	8.0	14.6	9.2	13.9	17.9
		LS	1.1	1.7	6.2	0.9	2.6	4.8	10.5	9.5	8.9
	0	M-M	0.6	2.2	9.0	0.2	2.0	5.4	0.5	1.4	3.0
		LS	0.6	0.7	5.7	0.1	0.9	2.1	6.2	5.0	4.0
150	64.9	M-M	23.0	57.5	83.2	28.0	78.4	136.7	141.8	199.2	265.1
		LS	1.9	10.3	16.4	5.9	13.7	26.9	35.2	49.6	52.2
	43.3	M-M	8.1	17.3	31.7	11.4	30.9	50.2	45.4	63.8	81.8
		LS	1.2	7.0	8.8	3.1	8.0	12.3	17.1	24.4	24.5
	21.6	M-M	2.6	8.1	14.1	4.1	10.9	19.1	13.6	19.6	25.1
		LS	0.6	2.4	4.8	1.5	3.6	5.2	7.5	10.8	9.6
	10.8	M-M	1.1	4.3	10.3	1.8	5.4	11.5	7.0	10.7	12.5
		LS	0.3	1.3	3.4	0.7	1.4	3.0	3.5	5.4	4.9
	0	M-M	0.1	1.7	7.9	0.1	0.8	4.7	0.7	1.8	1.5
		LS	0.1	0.6	3.7	0.1	0.2	1.5	0.7	1.6	0.3
300	64.9	M-M	9.2	28.9	48.2	18.0	46.8	99.0	90.9	143.6	190.2
		LS	1.2	5.4	8.4	2.6	7.4	15.8	14.0	21.5	25.6
	43.3	M-M	3.9	9.6	12.9	6.1	14.7	19.4	26.2	38.3	41.7
		LS	0.8	3.4	4.7	1.6	3.8	6.9	8.4	11.3	13.6
	21.6	M-M	1.5	4.2	6.6	1.9	5.7	7.7	8.5	11.6	17.1
		LS	0.4	1.3	2.5	0.8	1.8	2.8	3.6	4.6	5.3
	10.8	M-M	0.8	2.3	5.2	1.0	2.7	5.1	3.6	5.1	7.0
		LS	0.2	0.7	1.9	0.4	0.9	1.5	1.6	2.3	2.5
	0	M-M	0.1	1.1	3.4	0.1	0.7	2.3	0.2	0.5	0.9
		MLS	0.1	0.6	1.7	0.0	0.4	1.4	0.1	0.0	0.4

Table 3: Simulation-based averages of the MSE's ($\times 10000$) of the estimated transition probabilities $\hat{p}_{12}(s,t)$ along 1000 replications for different sample sizes and different censoring percentages. Both the location-scale (LS) estimator and the estimator of Meira-Machado et al (M-M) are given.

n	$P(\Delta_2 = 0)$		(s,t) (1,2)	(1,3.5)	(1,5)	(2.5,3)	(2.5,4)	(2.5,5)	(4,4.5)	(4,4.75)	(4,5)
50	64.9	M-M	218.9	487.2	872.9	37.2	135.9	272.4	51.7	85.4	123.9
		LS	6.8	81.7	206.3	4.6	41.9	97.5	8.0	16.2	26.9
	43.3	M-M	19.4	238.9	453.4	25.9	105.6	184.5	38.1	60.7	80.1
		LS	6.4	75.1	132.5	3.5	32.4	68.9	5.4	11.0	18.5
	21.6	M-M	19.2	191.8	315.2	21.0	85.2	128.5	28.5	44.1	57.3
		LS	6.5	71.3	150.9	3.0	31.5	66.8	4.0	7.8	11.8
	10.8	M-M	19.6	175.2	293.3	19.5	84.6	120.4	27.2	41.8	55.5
		LS	4.9	72.9	178.6	2.8	33.5	78.3	4.5	8.9	13.3
	0	M-M	21.1	175.6	252.0	20.0	79.0	103.6	25.5	37.8	52.5
		LS	5.4	89.8	132.5	3.5	35.3	69.6	5.0	10.2	16.0
100	64.9	M-M	15.9	168.7	390.3	12.0	64.5	124.6	25.9	38.9	46.1
		LS	8.9	37.8	96.7	2.6	16.3	49.7	3.0	7.0	13.7
	43.3	M-M	14.0	117.4	230.5	9.2	49.7	83.0	18.5	28.7	32.3
		LS	8.5	39.5	57.6	2.1	14.5	28.4	1.9	4.0	6.9
	21.6	M-M	13.0	102.8	191.3	7.9	39.4	66.6	13.6	21.1	25.8
		LS	4.9	39.8	69.3	1.6	14.5	28.9	1.8	3.7	5.8
	10.8	M-M	13.5	94.9	171.5	7.4	34.5	59.0	13.5	19.4	25.4
		LS	2.4	33.0	88.4	1.2	15.4	34.3	2.1	4.2	6.7
	0	M-M	13.3	91.0	161.1	7.0	32.2	55.8	13.0	18.4	24.3
		LS	2.5	40.2	88.7	1.4	15.8	35.8	2.1	4.4	7.0
150	64.9	M-M	11.0	93.0	206.6	8.3	39.9	76.6	14.8	26.8	36.8
		LS	9.4	27.2	64.9	2.4	10.9	34.6	2.4	5.5	10.7
	43.3	M-M	10.1	79.3	156.4	6.8	28.0	57.0	14.5	21.9	30.0
		LS	9.9	38.0	42.2	2.3	12.5	23.0	1.5	3.3	6.1
	21.6	M-M	9.3	69.4	112.7	5.8	25.3	43.1	11.8	15.3	20.5
		LS	2.8	26.2	56.7	1.1	10.7	21.7	1.4	2.8	4.4
	10.8	M-M	8.6	60.1	96.9	5.4	22.7	38.4	10.6	13.8	18.7
		LS	3.5	31.8	60.6	1.1	9.8	21.2	1.2	2.4	3.9
	0	M-M	8.4	60.0	94.5	5.5	21.5	34.0	9.4	12.5	16.8
		LS	1.4	23.6	98.0	0.8	10.6	31.3	1.8	3.9	6.5
300	64.9	M-M	6.2	42.6	98.3	3.6	16.7	36.2	8.9	13.4	18.7
		LS	8.3	15.5	43.6	1.7	5.2	25.6	1.3	3.7	8.3
	43.3	M-M	5.3	32.5	68.3	3.1	12.0	28.3	6.3	10.0	14.4
		LS	8.8	32.4	17.3	2.1	8.3	10.3	0.7	1.6	3.4
	21.6	M-M	4.7	29.9	50.7	2.7	11.5	20.6	5.3	7.7	10.5
		LS	5.4	30.6	26.1	1.4	7.8	9.6	0.6	1.2	2.0
	10.8	M-M	4.6	25.6	47.7	2.4	10.3	18.3	4.9	7.1	9.2
		LS	0.9	12.6	54.0	0.4	5.5	15.9	1.0	2.2	3.5
	0	M-M	4.4	22.9	44.5	2.4	9.9	17.1	4.6	6.4	8.4
		LS	1.0	20.7	60.2	0.5	7.0	17.6	0.8	1.6	2.6

Table 4: Simulation-based averages of the MSE's ($\times$10000) of the estimated transition probabilities $\hat{p}_{22}(s,t)$ along 1000 replications for different sample sizes and different censoring percentages.

4. About the choice of the kernel and the optimal bandwidth

For the kernel weights in the Beran-type estimator (4) we use the biquadratic kernel $K(u) = (15/16)\left(1-u^2\right)^2 I\left(|u| \leq 1\right)$, but we believe that this choice has relatively little impact on the mean squared error. However, the use of different bandwidths in the calculation of the Beran-type estimator, $\tilde{F}(y|x)$, may have a substantial effect on the performance of the estimators. In the paper by Van Keilegom et al. (2011), the bandwidth sequence a_n is chosen by minimizing the asymptotic mean-squared error (AMSE):

$$\text{AMSE} = \text{AsVar}\left(\tilde{F}(y|x)\right) + \left(\text{AsBias}\left(\tilde{F}(y|x)\right)\right)^2 = (na_n)^{-1}s^2(y|x) + a_n^4 b^2(y|x),$$

where the precise formulas for $s(y|x)$ and $b(y|x)$ are derived from the formulas provided in Van Keilegom et al. (2001). The adaptation to the context of censored T_1 is obtained by calculating Beran's estimator by using only the observations for which $\Delta_{1i} = 1$. Hence the optimal choice for the bandwidth sequence is given by

$$a_n = a_n(x,y) = \left(\frac{s^2(y|x)}{4b^2(y|x)}\right)^{1/5} n^{-1/5}.$$

However, this method cannot be directly applied for real data. To overcome this problem, Dabrowska (1992) replaces all the unknown quantities in the expressions of $s(y|x)$ and $b(y|x)$ ($H\left(y|x\right)$, $H^u\left(y|x\right)$ and their derivatives) by consistent estimators. Another way to select the bandwidth is to use bootstrap techniques (see, for example, Li and Datta, 2001). We believe that the performance of the location-scale regression model depends on how successful the location and scale functionals $m(x)$ and $\sigma^2(x)$ are estimated. Therefore, in this paper we propose to use two bandwidths, one for each functional. More precisely, we have

$$\hat{m}_{h_1}(x) = \int_0^1 \tilde{F}_{h_1}^{-1}(s|x)J(s)ds \text{ and } \hat{\sigma}^2_{h_2}(x) = \int_0^1 \tilde{F}_{h_2}^{-1}(s|x)^2 J(s)ds - \hat{m}_{h_1}(x)^2,$$

where $\tilde{F}_{h_j}(y|x)$ stands for the Beran-type estimator (4) computed using the optimal bandwidth $h_j, j = 1, 2$.

In this paper we propose the following procedure to obtain the bandwidth h_1 used to obtain the estimator $\hat{m}$ and the bandwidth h_2 used to obtain $\hat{\sigma}$:

Step 1. First for $b = 1$ to B (e.g. B=1000) simulate two independent random samples

$$S_1^b = \left\{\tilde{T}_{1i}^{\bullet 1b}, \tilde{T}_{2i}^{\bullet 1b}, \Delta_{1i}^{\bullet 1b}, \Delta_{2i}^{\bullet 1b}\right\}_{i=1}^{n} \quad \text{and} \quad S_2^b = \left\{\tilde{T}_{1i}^{\bullet 2b}, \tilde{T}_{2i}^{\bullet 2b}, \Delta_{1i}^{\bullet 2b}, \Delta_{2i}^{\bullet 2b}\right\}_{i=1}^{n}$$

by randomly sampling the n items from the original data set $\left\{\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right\}_{i=1}^{n}$ with replacement.

Step 2. Then, the bandwidth h_1 is automatically selected by minimizing the following weighted cross-validation error criterion:

$$CV_1 = \sum\nolimits_{b=1}^{B} \sum\nolimits_{i=1}^{n} W_i \left(\tilde{T}_{2i}^{\bullet 2b} - \hat{m}^{\bullet b}(\tilde{T}_{1i}^{\bullet 2b})\right)^2,$$

where $\hat{m}^{\bullet b}$ is the estimate obtained from the sample S_1^b and W_i is the Kaplan-Meier weight based on $\left(\tilde{T}_{1i}^{\bullet 2b} + \tilde{T}_{2i}^{\bullet 2b}, \Delta_{2i}^{\bullet 2b}\right)$.

Step 3. Using the h_1 obtained in the previous step we compute the residuals $\tilde{E}_i^{\bullet 1b} = \tilde{T}_{2i}^{\bullet 1b} - \hat{m}^{\bullet b}(\tilde{T}_{1i}^{\bullet 1b})$ and $\tilde{E}_i^{\bullet 2b} = \tilde{T}_{2i}^{\bullet 2b} - \hat{m}^{\bullet b}(\tilde{T}_{1i}^{\bullet 2b})$, and then h_2 is selected in a similar way as in **Step 1** by minimizing

$$CV_2 = \sum\nolimits_{b=1}^{B} \sum\nolimits_{i=1}^{n} W_i \left(\tilde{E}_i^{\bullet 2b} - \hat{\sigma}^{2\bullet b}(\tilde{T}_{1i}^{\bullet 2b})\right)^2,$$

where $\hat{\sigma}^{2\bullet b}$ is the estimate obtained from the sample $\left\{\tilde{T}_{1i}^{\bullet 1b}, \tilde{E}_i^{\bullet 1b}, \Delta_{1i}^{\bullet 1b}, \Delta_{2i}^{\bullet 1b}\right\}_{i=1}^{n}$.

For the location and scale functionals $\hat{m}(x)$ and $\hat{\sigma}(x)$ we need to choose a proper score function J. Since $\tilde{F}^{-1}(s|x)$ is only defined for s less than $\tilde{F}(\infty|x)$, the score function must be 0 from that point on (otherwise $\hat{m}(x)$ and $\hat{\sigma}(x)$ are not defined). Therefore we choose $J(s) = I(a \leq s \leq b)/(b-a)$, where $a = 0$ and $b = \min_i \tilde{F}(+\infty|X_i)$. This choice of b ensures that $\hat{m}(x)$ and $\hat{\sigma}(x)$ are always defined.

5. Breast Cancer Study

To illustrate the methods discussed in Section 2, we use survival data on breast cancer, diagnosed at the Santiago University Teaching Hospital, Spain. In the period between 1991 and 2000, 584 incident cases of breast cancer were diagnosed. The main goal of this study was to assess the prognostic value of flow cytometry-based proliferation markers, DNA index and S-phase fraction,

in breast cancer. Other goals included the study of the relationship between the different covariates and the disease evolution (Chavez-Uribe et al., 2007; Cadarso et al. 2010).

For each patient the vital status and date of relapse or death were obtained from their physicians, until the end of follow-up on December 31, 2004. In the period between 1991 and 2004, out of the 584 women, 402 (69%) were alive and disease-free, 167 (29%) experienced a recurrence (local-regional or metastases), 117 patients (20%) died due to cancer, and 11 due to other causes. In the analysis of the breast cancer data set, recurrence may be regarded as an associated state of risk, and the progressive three-state model depicted in Figure 1 can thus be used. The 11 patients dying due to other causes are included in the study but in the multi-state model (of Figure 1) they are treated as censored on the recurrence transition and they are not considered on the mortality transition from the 'alive with recurrence' state.

As mentioned in the introduction section, in these longitudinal cancer studies, the amount of time spent in the healthy state (sojourn time) is often of interest. By including covariates depending on the history (Kay, 1986), we verified the assumption that the transition rate from state 2 to state 3 is affected by the time spent in the previous state (p-value <0.05). This allowed us to conclude that the Markov model was unsatisfactory for the Galician breast cancer dataset.

Cancer patients who have experienced a recurrence are known to be at a substantially higher risk of death, thus rendering it essential to make diagnosis at sufficiently early stages. A balance between the number of points of diagnosis and the risk of recurrence should exist, making it essential to obtain good estimates for the transition probabilities. Since the process for the breast cancer data is not Markovian it is preferable to use estimators which do not rely on the Markov assumption. Both estimators considered in this section are free of the Markov assumption and therefore are suitable for the breast cancer data.

In this section we present results (plots) for the transition probabilities using the location-scale regression model (LS) and the estimator by Meira-Machado et al (2006). Recall that the results obtained in the simulation section suggest that the LS estimator is more efficient than the M-M estimator.

To illustrate the difference between the estimator proposed by Meira-Machado et al. (2006) and the new estimator (LS), we present several plots. Figures 7, 8 and 9 depict respectively the LS nonparametric estimates of

$p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{22}(s,t)$ together with pointwise confidence bands based on the bootstrap. Likewise, in Figures 9 and 10 the M-M estimators are plotted. Given s and t, the steps for construction of the confidence interval for the true transition probabilities $p_{hj}(s,t)$ are as follows:

Step 1. Obtain the estimated $\hat{p}_{hj}(s,t)$ from the sample $\left\{\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right\}_{i=1}^{n}$ as explained above.

Step 2. For $b = 1$ to B (e.g. $B = 1000$), simulate a random sample $\left\{\tilde{T}_{1i}^{\bullet b}, \tilde{T}_{2i}^{\bullet b}, \Delta_{1i}^{\bullet b}, \Delta_{2i}^{\bullet b}\right\}_{i=1}^{n}$ by randomly sampling the n items from the original data set $\left\{\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i}\right\}_{i=1}^{n}$ with replacement (that is, each individual value $(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$ has a probability n^{-1} of occurring), and obtain the bootstrap estimates $\hat{p}_{hj}^{b}(s,t)$.

Finally, the $(1-\alpha)\,100\%$ limits for the confidence interval of $p_{hj}(s,t)$ are given by

$$\left(2\hat{p}_{hj}(s,t) - \hat{p}_{hj}^{(1-\alpha/2)}(s,t), 2\hat{p}_{hj}(s,t) - \hat{p}_{hj}^{(\alpha/2)}(s,t)\right),$$

where $\hat{p}_{hj}^{(\alpha)}(s,t)$ represents the $\alpha \times 100\%$-percentile of the bootstrapped estimates $\hat{p}_{hj}^{b}(s,t)$ for $b = 1, \ldots, B$.

As expected, the LS estimator has less variability than the M-M estimator, which has fewer jump points as s increases. Since few events ('death') are observed at higher time values, consistency problems are expected at the right tail of the distribution when using the M-M estimator. These features can be seen in all plots but especially in the figures of the transition probability $p_{22}(s,t)$. While the LS estimator decreases smoothly with time (showing a good behavior in the right tail), the M-M estimator shows a sharp decrease to zero.

We note however the strange shape of the wiggly curve of the transition probability p_{12} when using the LS estimator. Though this can be explained by the location-scale model, we believe that a smoothed version of this curve could be more plausible. This could be obtained by smoothing the functions F_1 and F_e. Further details about this topic are out of the scope of the present work and will be reported elsewhere.

6. Discussion

In this paper we proposed and investigated nonparametric estimators for the transition probabilities in a progressive three-state model. The estimators are based on a location-scale model which is typically used to transfer tail information from lightly censored areas to heavily censored ones. We verified through simulations that the method based on the location-scale model may be much more efficient than other available estimators. We considered some modifications and practical issues while illustrating the proposed methodology using a real database on breast cancer from Galicia in Spain.

The estimates obtained from the location-scale regression method revealed to be quite wiggly (in particular the estimated curves for the transition probability p_{12}). We therefore propose to use smoothed versions of these curves (obtained by smoothing the functions F_1 and F_e). The smoothed version of the location-scale estimator provides smoothed curves for the transition probabilities which we believe to be more plausible for explaining real problems. The asymptotic properties of this smoothed estimator are not yet studied and such a theory is left for future research. Simulation results (not reported here) suggest a good behavior of these smoothed estimators.

We plan to provide users a R based package which, among others, will enable the estimation of the transition probabilities using the location-scale model. This package will be submitted to CRAN.

Acknowledgments

The authors acknowledge financial support from the Spanish Ministry of Education & Science in the form of grants MTM2008-01603. Luís F. Meira-Machado also acknowledges financial support by Grant PTDC/MAT/104879/2008 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. The research was also partially funded by CMAT and FCT under the POCI 2010 program. Ingrid Van Keilegom's research was financially supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

References

Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Statist.*, **5**, 141-150.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

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

Cadarso-Suárez, C., Meira-Machado, L., Kneib, T. and Gude, F. (2010). Flexible hazard ratio curves for continuous predictors in multi-state models: a P-spline approach. *Statist. Modelling*, **10**, 291-314.

Chavez-Uribe, E., Cameselle-Teijeiro, J., Vinuela, J.E., et al. (2007) Hypoploidy defines patients with poor prognosis in breast cancer. *Oncology Reports*, **17**, 1109-14.

Dabrowska, D.M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. *Scand. J. Statist.*, **19**, 351-361.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New-York.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Amer. Statist. Assoc.*, **53**, 457-481.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, **42**, 855-865.

Li, G. and Datta, G. (2001). A bootstrap approach to nonparametric regression for right censored data. *Ann. Inst. Statist. Math.*, **53**, 708-729.

Meira-Machado, L.F., de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifet. Data Anal.*, **12**, 325-344.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. *Ann. Statist.*, **27**, 1745-1784.

Van Keilegom, I., Akritas, M. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Comput. Statist. Data Anal.*, **35**, 487-500.

Van Keilegon, I., de Uña-Álvarez, J. and Meira-Machado, L. (2011). Non-parametric location-scale models for successive survival times under dependent censoring. *J. Statist. Plann. Infer.*, **141**, 1118-1131.

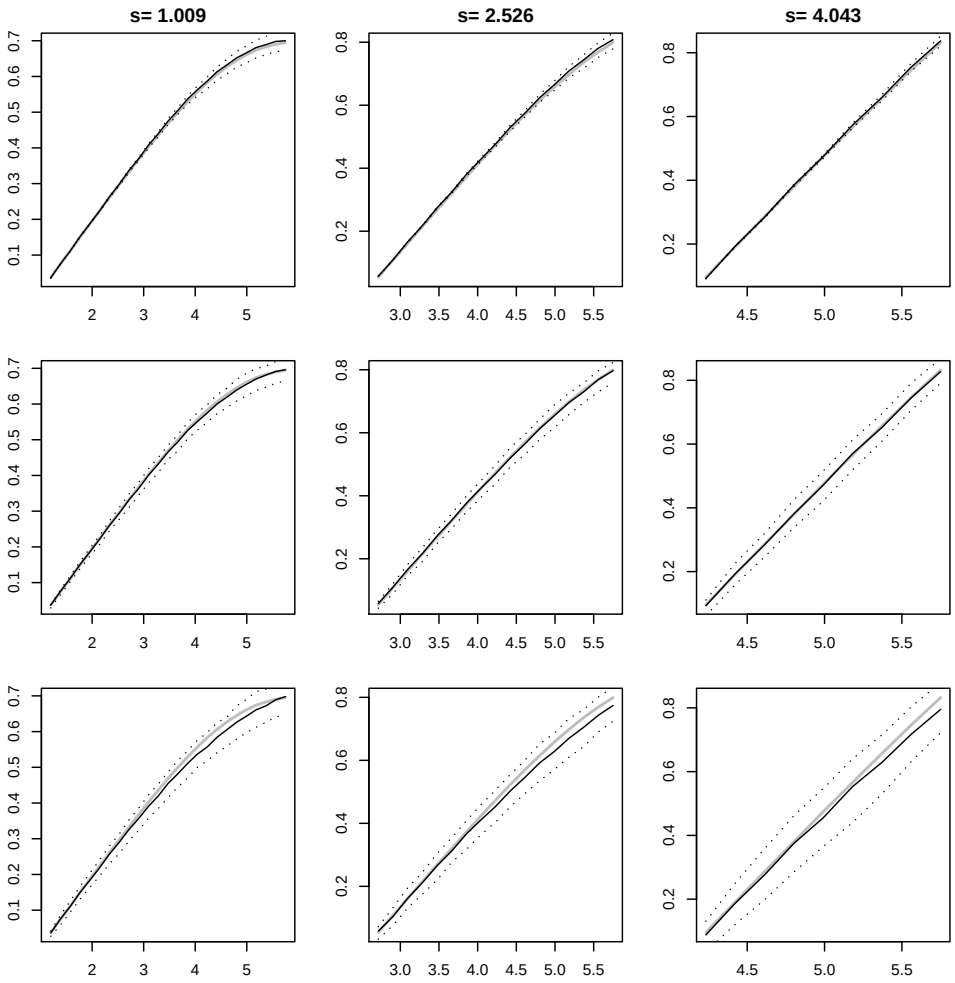

Figure 3: True transition probability functions $p_{12}(s,t)$ and 95% oscillation limits of the LS estimates $\hat{p}_{12}(s,t)$ for $s = 1.009$, $s = 2.526$ and $s = 4.043$. Estimates with $n = 300$ for percentage of censored data equal to 0 % (first row), 21.6 % (second row) and 64.9 % (third row).

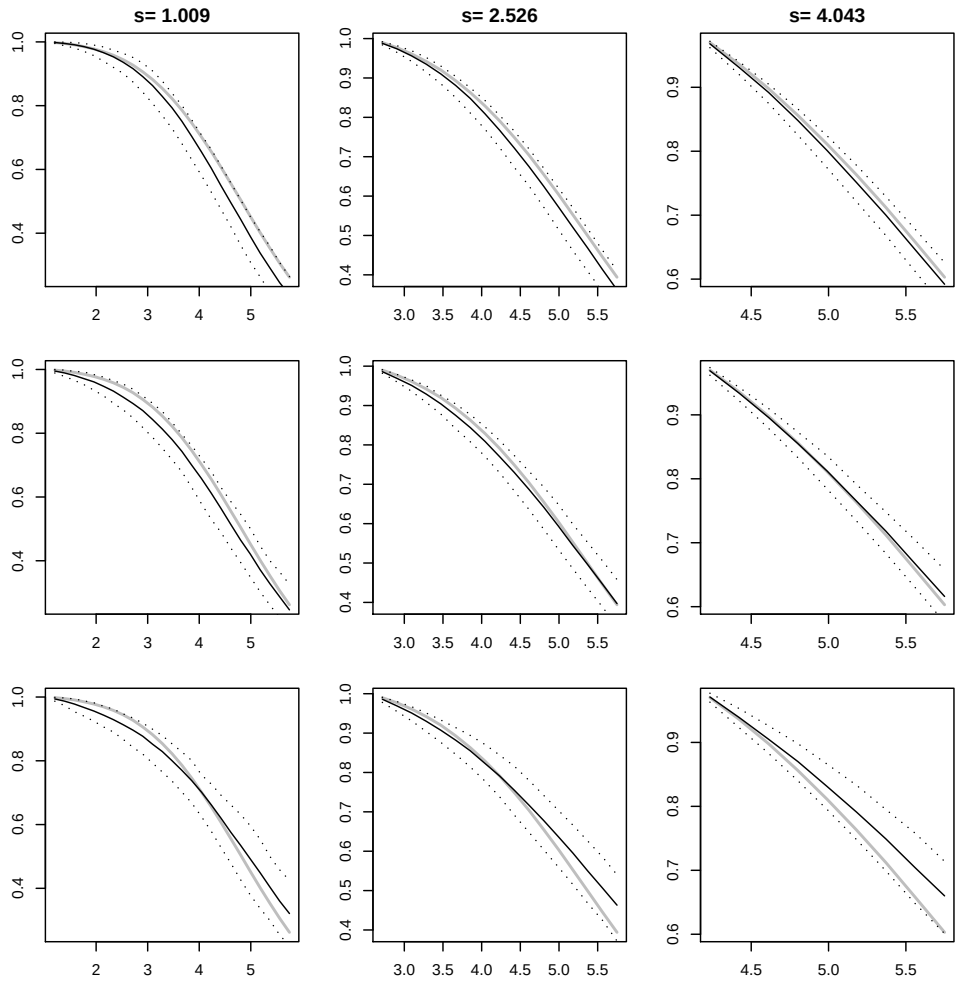

Figure 4: True transition probability functions $p_{22}(s,t)$ and 95% oscillation limits of the LS estimates $\hat{p}_{12}(s,t)$ for $s = 1.009$, $s = 2.526$ and $s = 4.043$. Estimates with $n = 300$ for percentage of censored data equal to 0 % (first row), 21.6 % (second row) and 64.9 % (third row).

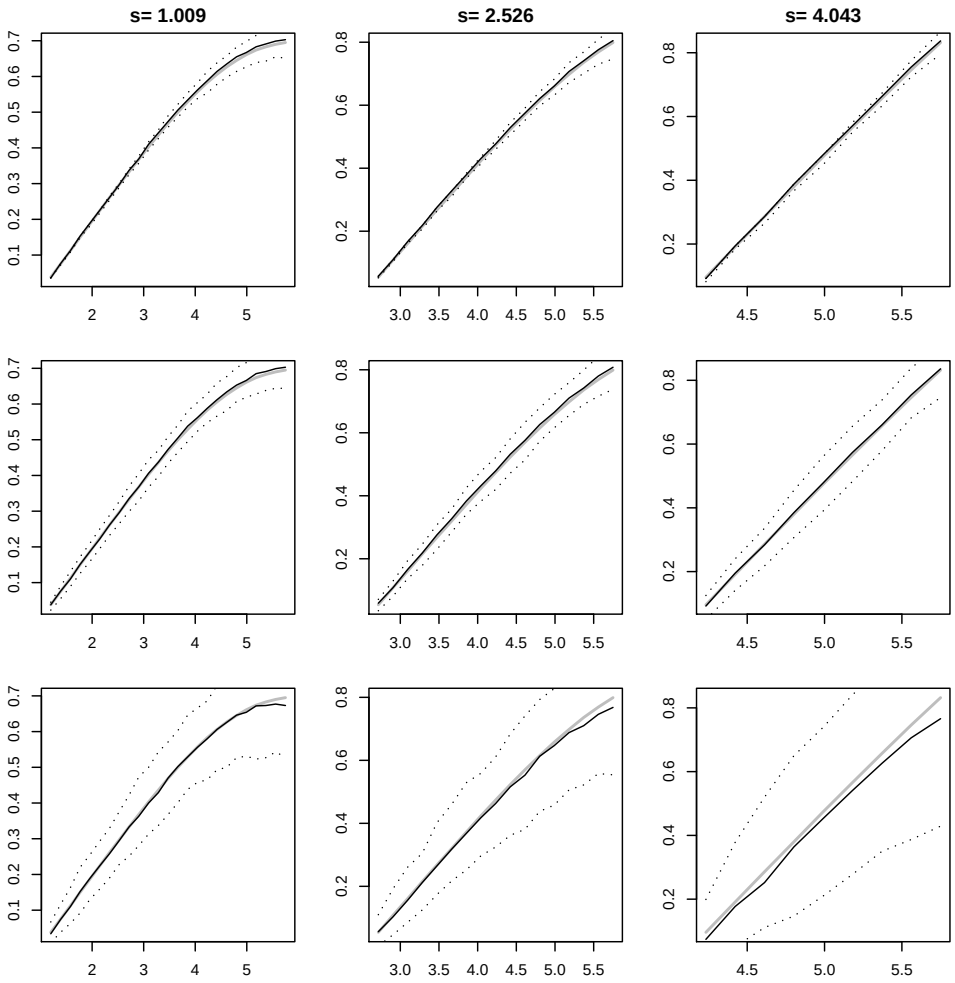

Figure 5: True transition probability functions $p_{12}(s,t)$ and 95% oscillation limits of the M-M estimates $\hat{p}_{12}(s,t)$ for $s = 1.009$, $s = 2.526$ and $s = 4.043$. Estimates with $n = 300$ for percentage of censored data equal to 0 % (first row), 21.6 % (second row) and 64.9 % (third row).

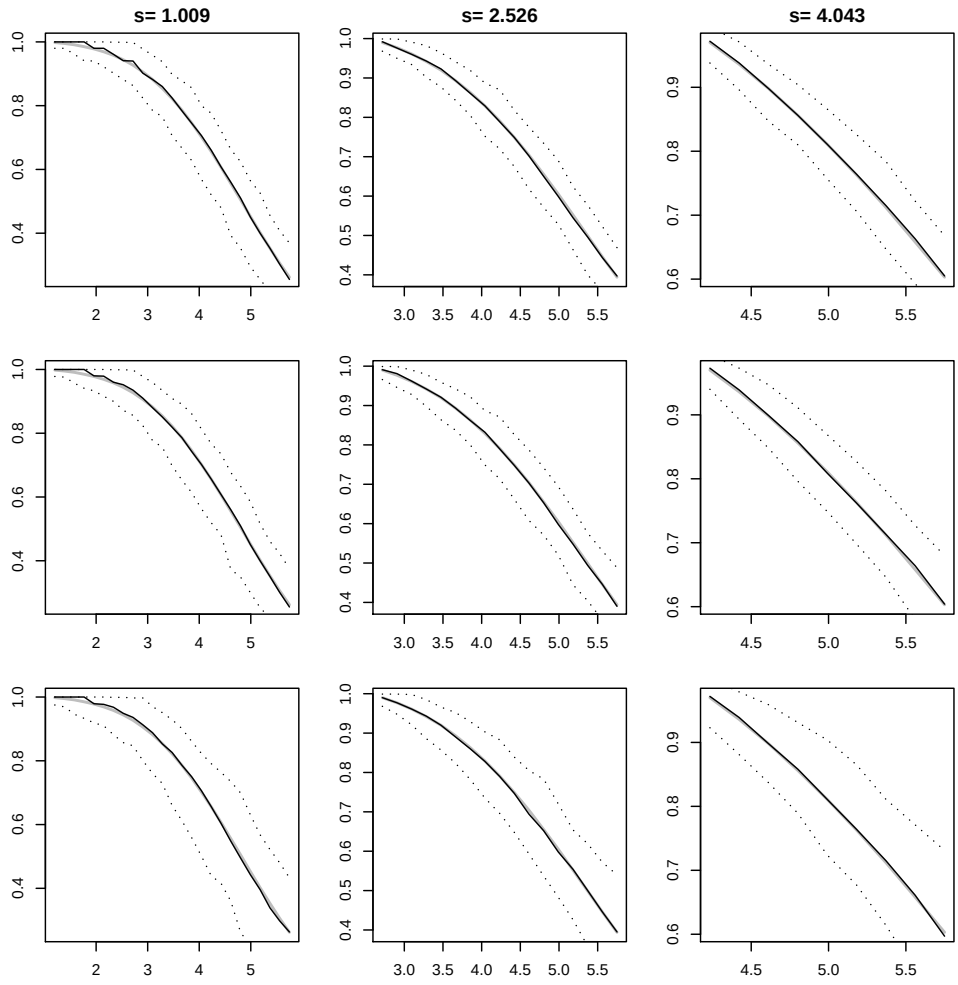

Figure 6: True transition probability functions $p_{22}(s,t)$ and 95% oscillation limits of the M-M estimates $\hat{p}_{12}(s,t)$ for $s = 1.009$, $s = 2.526$ and $s = 4.043$. Estimates with $n = 300$ for percentage of censored data equal to 0 % (first row), 21.6 % (second row) and 64.9 % (third row).

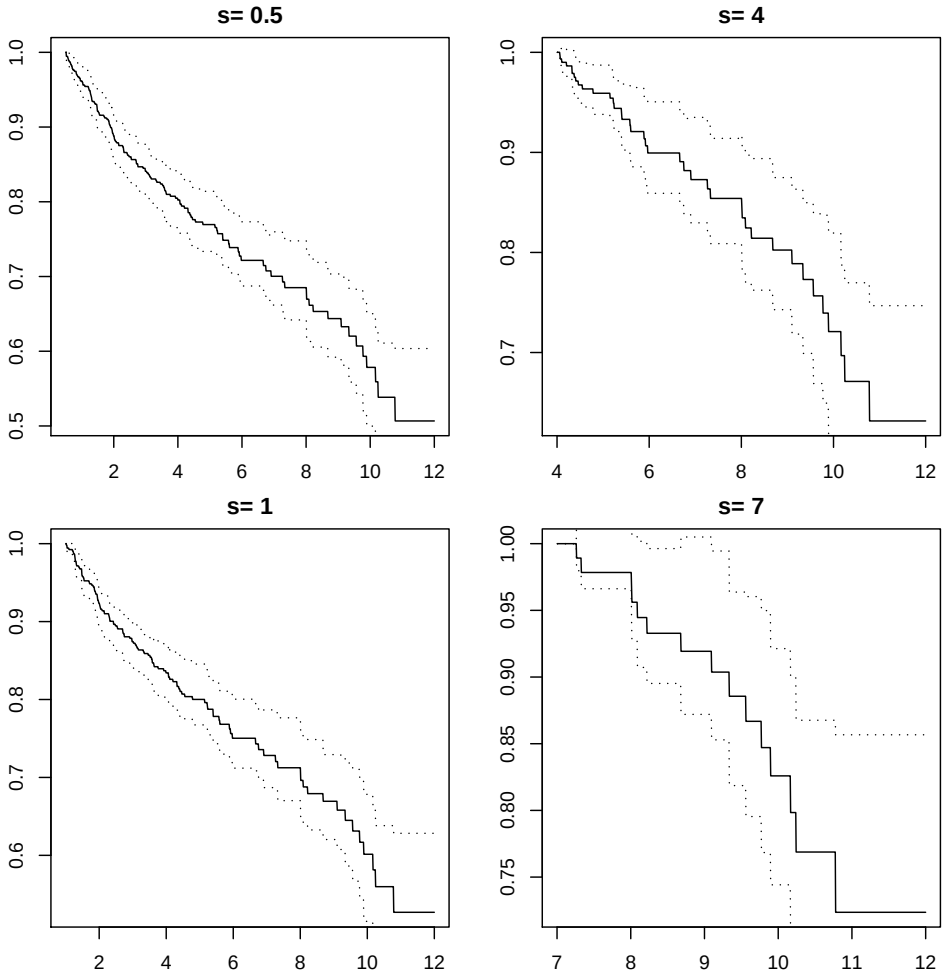

Figure 7: Estimated transition probabilities $\hat{p}_{11}(s,t)$, for some fixed values of s, using the LS estimator together with the corresponding 95 % pointwise confidence bands.

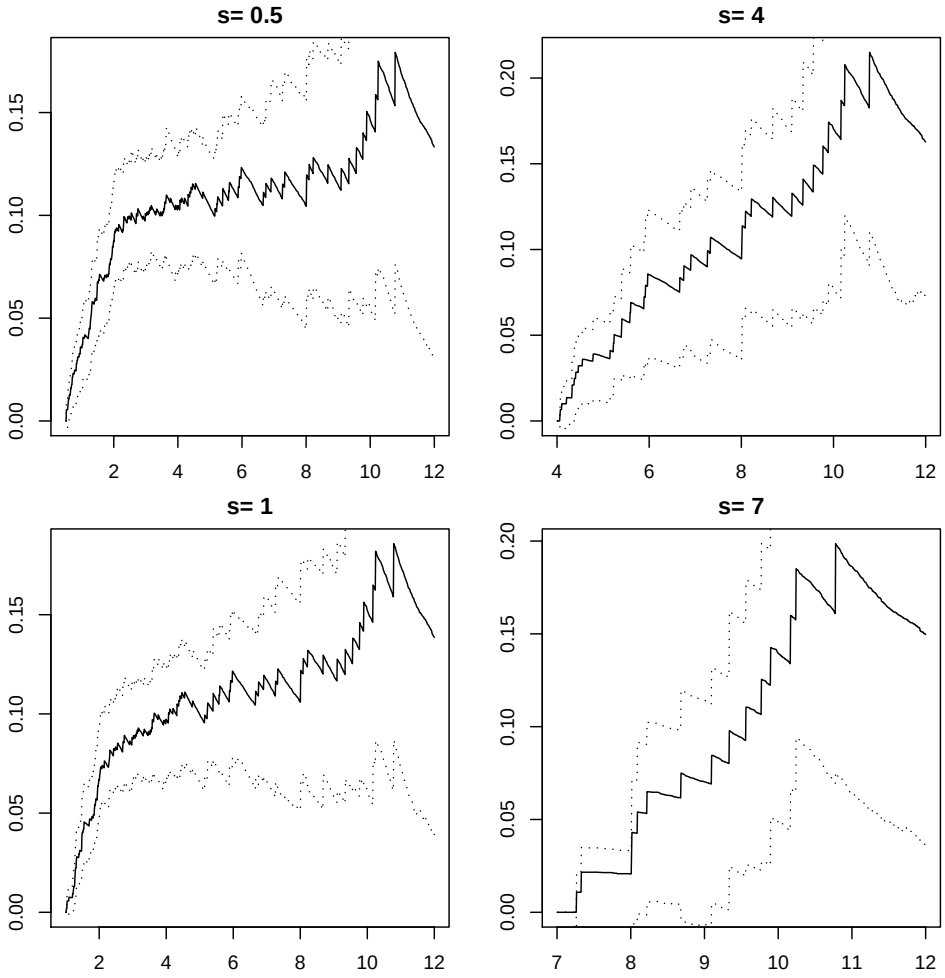

Figure 8: Estimated transition probabilities $\hat{p}_{12}(s,t)$, for some fixed values of s, using the LS estimator together with the corresponding 95 % pointwise confidence bands.

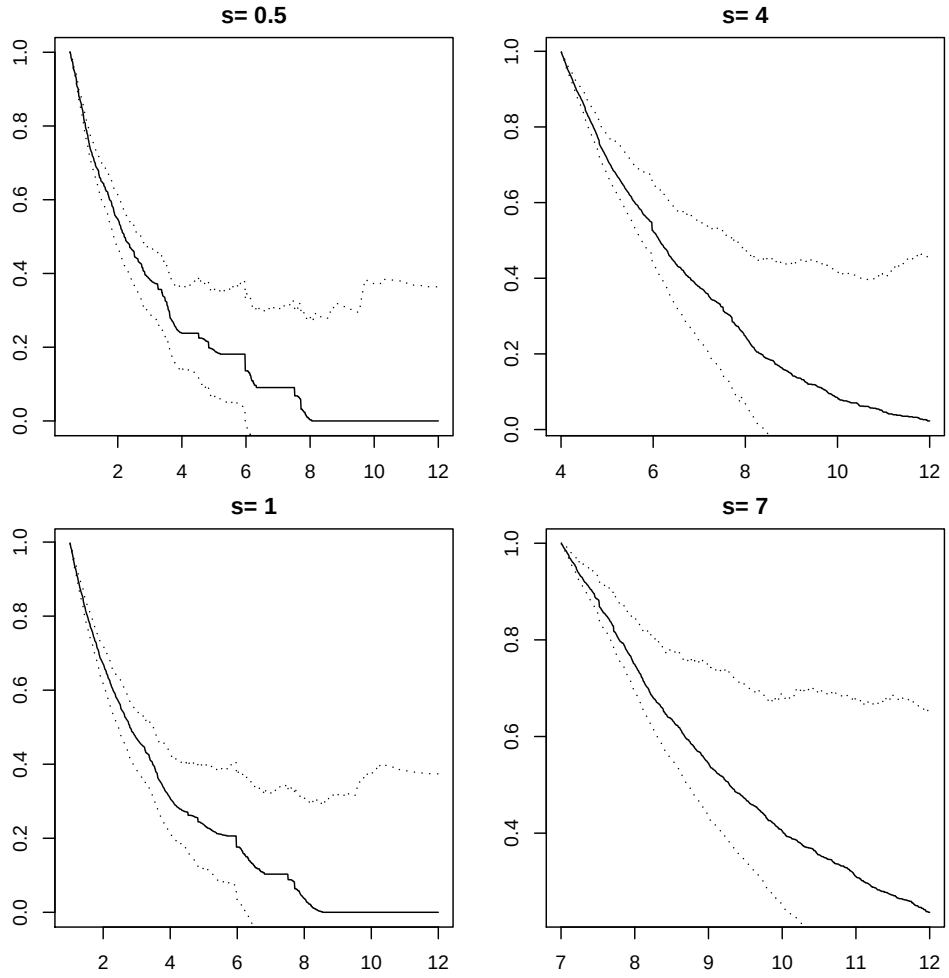

Figure 9: Estimated transition probabilities $\hat{p}_{22}(s,t)$, for some fixed values of s, using the LS estimator together with the corresponding 95 % pointwise confidence bands.

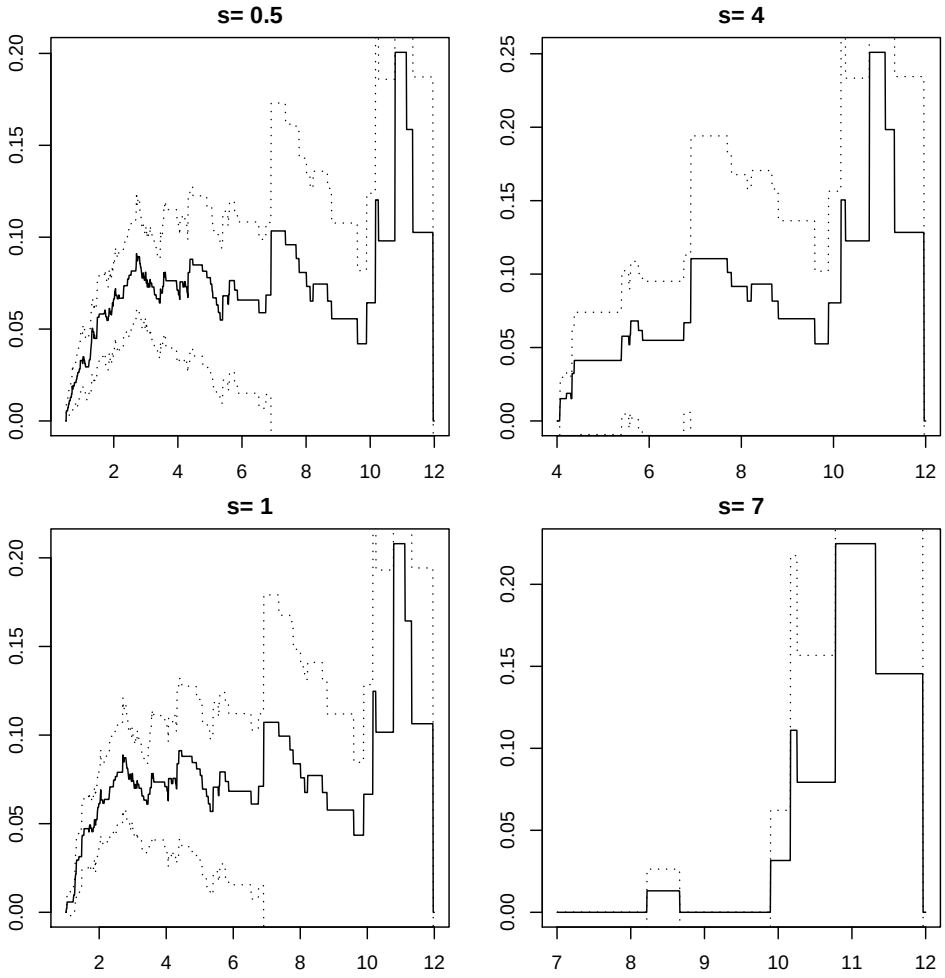

Figure 10: Estimated transition probabilities $\hat{p}_{12}(s,t)$, for some fixed values of s, using the M-M estimator together with the corresponding 95 % pointwise confidence bands.

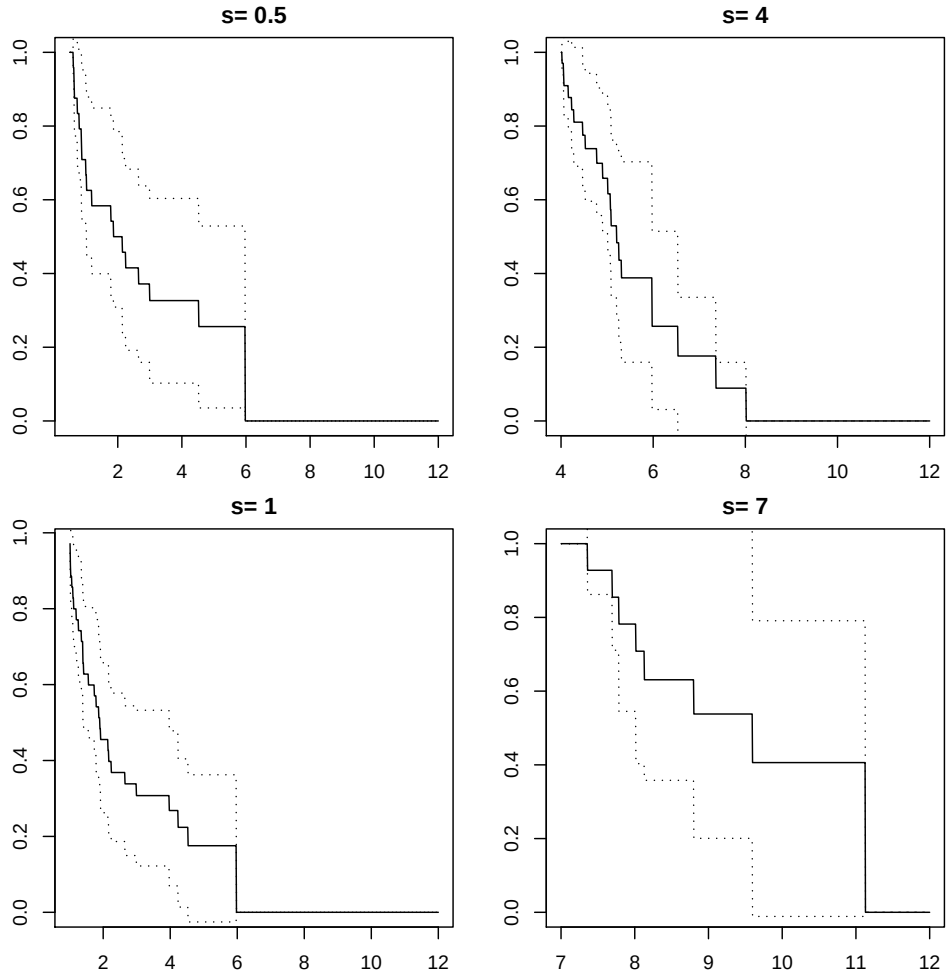

Figure 11: Estimated transition probabilities $\hat{p}_{22}(s,t)$, for some fixed values of s, using the M-M estimator together with the corresponding 95 % pointwise confidence bands.