Nonparametric Estimation of Conditional Transition Probabilities in a non-Markov Illness-death Model

Luís Meira-Machado · Jacobo de Uña-Álvarez · Somnath Datta

Received: date / Accepted: date

Abstract One important goal in multi-state modeling is the estimation of transition probabilities. In longitudinal medical studies these quantities are particularly of interest since they allow for long-term predictions of the process. In recent years significant contributions have been made regarding this topic. However, most of the approaches assume independent censoring and do not account for the influence of covariates. The goal of the paper is to introduce feasible estimation methods for the transition probabilities in an illness-death model conditionally on current or past covariate measures. All approaches are evaluated through a simulation study, leading to a comparison of two different estimators. The proposed methods are illustrated using real a colon cancer data set.

Keywords Conditional Survival · Dependent Censoring · Kaplan-Meier · Multi-state model · Nonparametric regression

1 Introduction

The so-called "illness-death" model plays a central role in the theory and practice of multi-state models (Andersen et al (1993), Meira-Machado et al (2009)). In the irreversible version of this model, individuals start in the "healthy" state and subsequently move either to the "diseased" state or to the "dead" state. Individuals in the "diseased" state will eventually move to the "dead" state without any possibility of recovery. See Figure 1. Many time-to-event data sets from medical

Luís Meira-Machado
Centre of Mathematics and Department of Mathematics and Applications, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
E-mail: lmachado@math.uminho.pt

Jacobo de Uña-Álvarez
Department of Statistics and O.R., University of Vigo, Spain
E-mail: jacobo@uvigo.es

Somnath Datta
Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, USA
E-mail: somnath.datta@louisville.edu

studies with multiple end points can be reduced to this generic structure. Thus, methods developed for the three-state illness-death model have a wide range of applications. From a theoretical standpoint, this is the simplest multi-state generalization of the survival analysis model that incorporates both branching (as in a multiple decrement/competing risk model) and an intermediate state (as in a progressive tracking model). Thus, unlike the survival or the competing risk model, this model is not necessarily Markovian.

Various aspects of the model dynamics are captured by the transition probabilities. In the presence of right censoring, these can be estimated by the Aalen-Johansen product limit estimator (Aalen and Johansen (1978)) provided the system is Markovian. However, as demonstrated by Meira-Machado et al (2006), the Aalen-Johansen estimator is inconsistent when the Markov assumption does not hold. They also illustrate through a real data example that the Markovianity cannot be taken for granted in practice. Meira-Machado et al (2006) and Amorim et al (2011) provide alternative nonparametric estimators specific to the three-state illness-death model that are consistent even without the Markov assumption.

In this paper, we revisit the problem of estimation of the transition probabilities of an irreversible, possibly non-Markov illness-death model. However, unlike the previous attempts, we are interested in a regression setup where we estimate these probabilities given a continuous covariate that could either be a baseline covariate or a current covariate that is observed for an individual before the individual makes a particular transition of interest. There has been little research on the estimation of conditional transition probabilities. Dabrowska and Lee (1996) introduce an averaged Beran's conditional estimate to yield a consistent estimate of the transition probabilities. The authors consider a vector of sojourn times in past states as the covariate. Dabrowska and Ho (2000) introduce graphical tests based on confidence procedures for the difference between transition probabilities evaluated over distinct covariate values. Arjas and Eerola (1993) considered graphical representations of conditional hazards and conditional survival for prediction purposes. However, none of these approaches lead to flexible estimates for the conditional transition probabilities as those provided in plots shown in Section 4. Our methodology is motivated by the colon cancer data set originally investigated by Moertel et al (1990) and subsequently reanalyzed by Lin et al (1999) to study the joint distribution of gap times between enrolment (curative surgery), the disease recurrence and death. These data can also be viewed as arising from a three-state illness-death model where "recurrence" can be modeled as the intermediate illness state. We are interested in the effect of a covariate (age at surgery, or number of lymph nodes with detectable cancer), on the probabilities of transitions between the states. Standard regression models in this setup (besides imposing Markovianity) usually rely on a parametric specification of the covariate effects on Markovity transition intensity functions; the resulting estimates of the covariate effects on transition probabilities are prone to model misspecifications errors. Therefore, flexible robust effects of the covariates on the transition probabilities as those depicted in Figures 3 and 4 (Section 4) can not be estimated through standard techniques.

In this paper, we provide two competing nonparametric regression estimators of the transition probability matrix of a three-state progressive illness-death model. Both estimators are valid under mild regularity conditions even when the system is non-Markov or conditionally non-Markov. In both estimators, local smoothing

is done by introducing kernel weights that are either based on a local constant (i.e. Nadaraya-Watson) or a local linear regression. Right censoring is handled by appropriate reweighting of the chosen summands and the differences between the two estimators are somewhat subtle in this regard. The first estimator only put mass on observations that are completely uncensored (i.e., fully observed till death) whereas the second estimator jumps on observations that were uncensored till a given time. Extensive simulation studies are provided comparing the two estimators.

The rest of the paper is organized as follows. Section 2 introduces the formal notations and the two estimators. Section 3 describes the simulation setup and the findings of a number of simulation experiments. An illustrative real data application is provided in Section 4. The main body of the paper ends with a discussion section (Section 5). Additional simulation results are presented in the Appendix.

2 Conditional Transition Probabilities

2.1 Notation and Preliminaries

A multi-state model is a stochastic process $(Y(t), t \in \mathcal{T})$ with a finite state space, where $Y(t)$ represents the state occupied by the process at time t. In this paper we consider the progressive illness-death model depicted in Figure 1 and we assume that all subjects are in state 1 at time $t = 0$. This model is encountered in many medical studies (cancer studies, transplantations, etc) where State 1 is some initial stage of the disease (e.g. healthy, disease-free, etc), State 2 is some intermediate stage of the disease (e.g. alive with local recurrence, certain stage of a disease, transplantation, etc) and State 3 is an absorbing state (e.g. dead) which all subjects are expected to reach eventually. For this model the transitions allowed are $1 \to 2$, $1 \to 3$ and $2 \to 3$. This means that an individual may visit State 2 or going directly to State 3 without visiting State 2.

For two states i, j and two time points $s < t$, introduce the so-called transition probabilities

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

In the illness-death model we only need to estimate three different transition probabilities: $p_{11}(s,t)$, $p_{12}(s,t)$ and $p_{22}(s,t)$. The two other transition probabilities ($p_{13}(s,t)$ and $p_{23}(s,t)$) can be obtained from these ones since $p_{13}(s,t) = 1 - p_{11}(s,t) - p_{12}(s,t)$ and $p_{23}(s,t) = 1 - p_{22}(s,t)$.

In the framework of the progressive illness-death model, we may consider three random variables T_{12}, T_{13} and T_{23}, that represent the potential transition times from one state to another one. According to this notation, subjects not visiting state 2 will reach the absorbing state at time T_{13}. This time will be $T_{12} + T_{23}$ if he/she passes through state 2 before, where the variables T_{12} and T_{23} are recorded successively, rather than simultaneously. In this model we have two competing transitions leaving state 1. Therefore, we denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting state 2 at some time, $Z = T_{12} \wedge T_{13}$ the sojourn time in state 1, and $T = Z + \rho T_{23}$ the total survival time of the process.

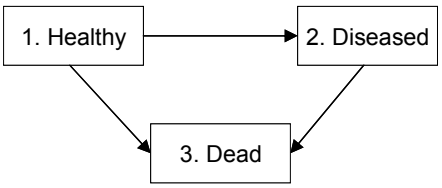

Fig. 1 Illness-death model

Let C be the univariate censoring variable and put $\widetilde{Z} = Z \wedge C$ and $\widetilde{T} = T \wedge C$ for the censored versions of Z and T. Then, let $\Delta_1 = I(Z \leq C)$ and $\Delta = I(T \leq C)$ denote the respective censoring indicators. Note that ρ is observed only when $\Delta_1 = 1$.

In this work we are interested in estimating the conditional transition probabilities: $p_{11}(s,t \mid X)$, $p_{12}(s,t \mid X)$, and $p_{22}(s,t \mid X)$ that can be computed for any times s and t, $s < t$ but conditional on some covariate value which we denote by X. Following the notation introduced above, the conditional transition probabilities are written as

$$\begin{gathered} p_{11}(s,t \mid X) = \frac{1 - P\left(Z \leq t \mid X\right)}{1 - P\left(Z \leq s \mid X\right)}, \quad p_{12}(s,t \mid X) = \frac{P\left(s < Z \leq t, T > t \mid X\right)}{1 - P(Z \leq s \mid X)} \\ \text{and} \quad p_{22}(s,t \mid X) = \frac{P\left(Z \leq s, T > t \mid X\right)}{P\left(Z \leq s, T > s \mid X\right)}. \end{gathered} \tag{1}$$

Now, the conditional transition probability $p_{11}(s,t \mid X)$ and the denominator of $p_{12}(s,t \mid X)$ only involve the conditional distribution of Z given X. This conditional distribution can be estimated nonparametrically following Beran (1981). The remaining quantities involve conditional expectations of particular transformations of the pair (Z,T) given X, $S\left(\varphi \mid X\right) := E\left[\varphi\left(Z,T\right) \mid X\right]$ which can not be estimated so simply. Moreover, the transition probabilities will be hard to estimate in the right tail where censoring effects are stronger. Because of this we also use alternative expressions for the conditional transition probabilities $p_{12}(s,t \mid X)$ and $p_{22}(s,t \mid X)$:

$$\begin{aligned} p_{12}(s,t \mid X) &= \frac{P\left(s < Z \leq t \mid X\right) - P\left(s < Z \leq t, T \leq t \mid X\right)}{1 - P(Z \leq s \mid X)}, \\ p_{22}(s,t \mid X) &= \frac{P\left(Z \leq s \mid X\right) - P\left(Z \leq s, T \leq t \mid X\right)}{P\left(Z \leq s \mid X\right) - P\left(T \leq s \mid X\right)}. \end{aligned} \tag{2}$$

Transition probability estimators, without any covariate, based on equations (2) are implemented in the R based package *p3state.msm* (Meira-Machado and Roca-Pardiñas (2011)).

As mentioned in Section 1 we will provide two competing nonparametric regression estimators for the transition probabilities. The first set of estimators we propose is based on equations (2) while the second set of estimators can only be implemented using the expressions given in equations (1). This will be clarified later while introducing the estimators.

In sum, we need to estimate the following conditional expectations: $S\left(\xi_s \mid X\right)$, $S\left(\psi_{s,t} \mid X\right)$, $S\left(\widetilde{\psi}_{s,t} \mid X\right)$, $S\left(\varphi_{s,t} \mid X\right)$ and $S\left(\widetilde{\varphi}_{s,t} \mid X\right)$, where $\xi_s(u,v) = I(u \leq s)$, $\psi_{s,t}(u,v) = I(s < u \leq t, v \leq t)$, $\widetilde{\psi}_{s,t}(u,v) = I(u \leq s, v \leq t)$, $\varphi_{s,t}(u,v) = I(s < u \leq t, v > t)$ and $\widetilde{\varphi}_{s,t}(u,v) = I(u \leq s, v > t)$.

In the following subsection, we discuss how these conditional expectations can be estimated from the data $\left\{\left(\widetilde{Z}_i, \widetilde{T}_i, \Delta_{1i}, \Delta_i, \Delta_{1i}\rho_i, X_i\right), 1 \leq i \leq n\right\}$, which are assumed to form a random sample of the vector $\left(\widetilde{Z}, \widetilde{T}, \Delta_1, \Delta, \Delta_1\rho, X\right)$. We will estimate these quantities assuming that the censoring variable C is independent of (Z,T) given X. Note that this assumption does not exclude the possibility of induced dependent censoring (i.e., C is unconditionally dependent on (Z,T)). Markovianity will not be assumed.

2.2 The Estimators

In this section, we will introduce two estimators for the conditional transition probabilities, $p_{hj}(s,t \mid X)$, in an illness-death model. As mentioned in Section 2.1, this can be done via estimating the general conditional expectations such as $E\left[\varphi\left(Z,T\right) \mid X = x\right]$. To estimate these quantities we may use kernel smoothing techniques by calculating a local average of the $\varphi(Z,T)$. This can be written as $\sum_{i=1}^{i=n} W_{1i}(x)\varphi(Z_i,T_i)$ where $W_{1i}(x)$ is a weight function which corresponds to either a Nadaraya-Watson (Nadaraya (1965),Watson (1964)) estimator or a local linear estimator. In our case, we have to estimate $f(x;s,t) = E\left[\psi_{s,t}\left(Z,T\right) \mid X = x\right]$, $g(x;s,t) = E\left[\widetilde{\psi}_{s,t}\left(Z,T\right) \mid X = x\right]$, $\tilde{f}(x;s,t) = E\left[\varphi_{s,t}\left(Z,T\right) \mid X = x\right]$, $\tilde{g}(x;s,t) = E\left[\widetilde{\varphi}_{s,t}\left(Z,T\right) \mid X = x\right]$ and $h(x;s) = E[\xi_s\left(Z\right) \mid X = x]$.

To handle right censoring, both estimators employ the inverse probability of censoring weighting (Lin et al (1999); Satten and Datta (2001)). To that end we need to estimate the distribution function of C given X, G_X. The estimation of the conditional distribution function of the response, given the covariate under random censoring has been considered in many papers. This topic was introduced by Beran (1981) and was further studied by several authors (see e.g. papers by Dabrowska (1987), Akritas (1994); Van Keilegom et al (2001) and Van Keilegom (2004)). Recently, Beran's estimator has been extended to regression of state occupation probabilities of a multi-state model by Mostajabi and Datta (2012). Beran's estimator of G_x is given by

$$\widehat{G}_x(t) = 1 - \prod_{T_i \leq t, \Delta_i = 0} \left[1 - \frac{W_{0i}(x,a_n)}{\sum_{j=1}^{n} I(T_j \geq T_i) W_{0j}(x,a_n)}\right] \tag{3}$$

with

$$W_{0i}(x,a_n) = \frac{K_0\left((x - X_i)/a_n\right)}{\sum_{j=1}^{n} K_0\left((x - X_j)/a_n\right)}; \tag{4}$$

here $W_{0i}(x,a_n)$ are the Nadaraya-Watson (NW) weights, K_0 is a known probability density function (the kernel function) and a_n is a sequence of bandwidths. This estimator reduces to the well-known Kaplan-Meier (Kaplan and Meier (1958)) estimator if a constant weight is used instead of the NW weights.

In order to introduce our estimators note that, assuming that the support of the conditional distribution of T is contained in that of C given X, we have $E[\psi(Z,T) \mid X] = E[\psi(\widetilde{Z},\widetilde{T})\Delta/(1-G_X(\widetilde{T}^-)) \mid X)]$. We propose to replace G_X by its Beran's estimator $\hat{G}_X$ and use NW or local linear weights to estimate $f(x;s,t)$, by

$$\hat{f}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{\psi_{s,t}(\widetilde{Z}_i,\widetilde{T}_i)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}_i^-)} = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(s<\widetilde{Z}_i\leq t,\widetilde{T}_i\leq t)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}_i^-)}$$

where $W_{1i}(x,b_n)$ are the NW weights as in (4) but with a possibly different bandwidth b_n and kernel K_1, or using local linear weight,

$$W_{1i}(x,b_n) = \frac{K_1\left((x-X_i)/b_n\right)\left[S_{n,2}(x)-(x-X_i)S_{n,1}(x)\right]}{\sum_{j=1}^{n} K_1\left((x-X_j)/b_n\right)\left[S_{n,2}-(x-X_j)S_{n,1}(x)\right]}$$

with $S_{n,l}(x) = \sum_{i=1}^{n} K_1((x-X_i)/b_n)(x-X_i)^l$, $l=0,1,2$ and where b_n is a sequence of bandwidths and K_1 is a known kernel function.

Similarly, we can use Nadaraya-Watson estimators or local linear estimators to estimate $g(x;s,t)$ and $h(x;s)$ i.e.

$$\hat{g}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{\widetilde{\psi}_{s,t}(\widetilde{Z}_i,\widetilde{T}_i)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}_i^-)} = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(\widetilde{Z}_i\leq s,\widetilde{T}_i\leq t)\Delta_i}{1-\hat{G}_{X_i}(\widetilde{T}_i^-)}$$

and

$$\hat{h}(x;s) = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{\xi_s(\widetilde{Z}_i)\Delta_{1i}}{1-\hat{H}_{X_i}(\widetilde{Z}_i^-)} = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{I(\widetilde{Z}_i\leq s)\Delta_{1i}}{1-\hat{H}_{X_i}(\widetilde{Z}_i^-)}$$

where $\hat{H}_X$ stands for the Beran estimator of the conditional distribution of C_1 (the censoring variable of the sojourn time in State 1) given X based on the $(\widetilde{Z}_i, 1-\Delta_{1i})$'s. Finally, we may introduce Inverse Probability Censoring Weighted estimators (IPCW) for the conditional transition probabilities, as follows:

$$\hat{p}_{11}(x;s,t) = \hat{p}_{11}(s,t \mid X=x) = \frac{\hat{h}(x;t)}{\hat{h}(x;s)}, \tag{5}$$

$$\hat{p}_{12}(x;s,t) = \hat{p}_{12}(s,t \mid X=x) = \frac{\hat{h}(x;t)-\hat{h}(x;s)-\hat{f}(x;s,t)}{\hat{h}(x;s)}, \tag{6}$$

$$\hat{p}_{22}(x;s,t) = \hat{p}_{22}(s,t \mid X=x) = \frac{\hat{g}(x;s,t)}{\hat{g}(x;s,s)}. \tag{7}$$

Alternatively, by noting that $E[\varphi_{s,t}(Z,T) \mid X] = E[I(Z\leq s, T>t) \mid X] = E[I(Z\leq s, T>t)I(C>t)/(1-G_X(t^-)) \mid X]$, a different set of estimators may be introduced. This approach has been used previously by Lin et al (1999) to estimate the bivariate distribution for censored gap times. In our setup, this alternative estimators of the transition probabilities are given by

$$\tilde{f}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(s < \widetilde{Z}_i \leq t, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)},$$

$$\tilde{g}(x;s,t) = \sum_{i=1}^{n} W_{1i}(x,b_n)\frac{I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)}{1-\hat{G}_{X_i}(t^-)},$$

and

$$\tilde{h}(x;s) = \sum_{i=1}^{n} W_{1i}(x,c_n)\frac{I(\widetilde{Z}_i \geq s)\Delta_{1i}}{1-\hat{H}_{X_i}(s^-)}.$$

These lead to Lin type estimator of conditional transition probabilities by replacing $\hat{f}$, $\hat{g}$, and $\hat{h}$ in (5)-(7) by $\tilde{f}$, $\tilde{g}$ and $\tilde{h}$, respectively.

Theoretical investigation such as consistency and further asymptotics can be pursued following established paths of studying the bias and the variance terms of various pieces of the estimators. More focused in practical issues, the finite-sample performance of IPCW estimators and the alternative LIN-based estimators is investigated by simulations in the following section.

3 Simulation Study

In this section, we carry out and extensive simulation study to investigate the behavior of the proposed estimators for finite sample sizes. More specifically, the estimators $\hat{p}_{11}(x;s,t)$, $\hat{p}_{12}(x;s,t)$, $\hat{p}_{22}(x;s,t)$, $\tilde{p}_{11}(x;s,t)$, $\tilde{p}_{12}(x;s,t)$ and $\tilde{p}_{22}(x;s,t)$ introduced in Section 2 are considered.

To simulate the data in the illness-death model, we follow closely the work described by Amorim et al (2011), but include covariate effects. 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 1, corresponding to 0.25 correlation between Z and $T-Z$. For the second subgroup of individuals ($\rho = 0$), the value of Z is simulated according to an exponential with rate parameter 1. To include covariate effects, the sojourn time in state 1, Z, was forced to depend on the X. In summary, the simulation procedure is as follows:

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

Step 2. If $\rho = 1$ then:
(2.1) $V_1 \sim U(0,1), V_2 \sim U(0,1)$ and $X \sim U(0,1)$ are independently generated;
(2.2) $U_1 = V_1, A = (2U_1 - 1) - 1, B = (1 - (2U_1 - 1))^2 + 4V_2(2U_1 - 1)$
(2.3) $U_2 = 2V_2/\left(\sqrt{B} - A\right)$
(2.4) $Z = \ln(1/(1 - U_1))$ and $\lambda(X) = 0.6X + 0.4$
(2.5) $Z(X) = Z/\lambda(X), T = \ln(1/(1 - U_2)) + Z(X)$
If $\rho = 0$ then $Z = Z(X)$.

In our simulation we consider $p = 0.7$. To allow for dependent censoring, $C|X = x$ is generated from an exponential distribution with rate $\lambda(x) = 0.15 + 0.35x$. This induces a censoring percentage on T of about 42%. We considered as (s,t) pairs four different points, corresponding to the different combinations of the quantiles 0.2, 0.4, 0.6 and 0.8 of the exponential distribution with rate 1.

In Figure 2, we plot the IPCW and LIN-based conditional transition probabilities, by fixing $s = 0.2231$ and considering two possible values for the covariate information (corresponding to the first and third quartiles, respectively). The results, which are estimators averaged along 1,000 Monte Carlo trials of size $n = 100$, show that (a) IPCW-type and LIN-based estimators are close to each other, and that (b) the transition probabilities greatly depend on covariate information, particularly $p_{11}(x; 0.2231, t)$ and $p_{12}(x; 0.2231, t)$ (not so clear for $p_{22}(x; 0.2231, t)$). This influence of the covariate can be also seen from the simulation steps described above: larger values of X are associated to smaller sojourn times in state 1 and, consequently, to a smaller survival (T).

A main aim of this simulation study is to investigate the comparative performance of the two proposed estimators (IPCW and LIN-based). For measuring the estimators' performance, we computed the integrated mean square error (IMSE) of the estimators. For each simulated setting we derived the analytic expression of $p_{ij}(x; s, t)$ so the MSE of the estimator could be computed. K = 1000 Monte Carlo trials were generated, with two different sample sizes $n = 100$ and $n = 200$. Let $\hat{p}_{ij}^k(x; s, t)$ denote the estimated conditional transition probability based on the kth generated data set. For each fixed (x, s, t) we computed the pointwise estimates of the MSE as:

$$\widehat{MSE}(\hat{p}_{ij}(x; s, t)) = \frac{1}{K} \sum_{k=1}^{K} [\hat{p}_{ij}^k(x; s, t) - p_{ij}(x; s, t)]^2. \tag{8}$$

To summarize the results we fixed the values of (s,t) using several quantiles (the same pairs as those used in the paper by Lin et al (1999)) and we calculated the IMSE as

$$\widehat{IMSE} = \sum_{x_l} \widehat{MSE}(\hat{p}_{ij}(x_l; s, t)) \times \delta \tag{9}$$

where x_l denotes a set of grid points for the covariate, going from 0 to 1 with step $\delta = 0.025$. The results are displayed in Tables 1 to 3. To compute the conditional transition probabilities $\hat{p}_{ij}(x; s, t)$ and $\tilde{p}_{ij}(x; s, t)$ we have used a common bandwidth selector and Gaussian kernels. To this end we have used the `dpik` function which is available from the `R KernSmooth` package. This is the data based

bandwidth selector of Wand and Jones (1997). We also performed additional simulations using other bandwidth selectors; for example, the plug-in bandwidth of Altman and Leger (1995), `ALbw`, available from the `R kerdiest` package, was also used. This alternative bandwidth did not provide better results (not shown). In addition, we have seen that, by choosing the 'optimal' bandwidth from a sequence of fixed bandwidths the results would not change much when compared with those attained with the `dpik` function. For the computation of $W_1(x; b_n)$ we have used Nadaraya-Watson (NW) and local linear weights (for the weights W_0 of the Beran's estimator we simply used NW). Since the results for NW weights were always superior (results not shown) to those based on local linear weights, we only provide here the results corresponding to the former. Additional simulation results are provided in the Appendix.

When using NW weights the two estimators (IPCW and LIN-based) for $p_{11}(x; s, t)$ are equal and, therefore, in Table 1 we only give one set of results. In general, both methods provide good results with IMSE values which decrease with an increasing sample size. It is also seen that the estimation of the transition probabilities is performed with less accuracy as s and t grow but for $p_{22}(x; s, t)$, for which the smallest values of IMSE are obtained for large s and t. Results shown in Table 2 suggest that the IPCW method leads to better results for $p_{12}(x; s, t)$ while neither one seems to be uniformly the best for estimating $p_{22}(x; s, t)$ (Table 3). The IPCW method obtains better results for estimating $p_{22}(x; s, t)$ for all pairs (s, t) but for $t = 1.6094$ and $s < 0.9163$ where censoring effects are stronger. LIN-based method deals more efficiently at those points.

		t	0.5108	0.9163	1.6094
	s				
n=100	0.2231		4.1895	7.1066	8.1987
	0.5108		—	6.8274	10.7996
	0.9163		—	—	15.1051
n=200	0.2231		2.8102	4.7099	5.5501
	0.5108		—	4.8029	7.3873
	0.9163		—	—	9.9158

Table 1 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{11}(x; s, t)$ along 1,000 trials for different sample sizes

4 Example of an Application

To illustrate our estimators we consider a well known data set from a colon cancer study which is freely available as part of the `R survival` package (Moertel et al (1990)).

These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. From the total of 929 patients, affected by colon cancer, that underwent a curative surgery for colorectal cancer, 468 developed a recurrence and among these 414 died. 38 patients died without recurrence. The remaining 423 patients contributed with censored survival times. For each individual, an indicator of his/her final vital status (censored or not), the survival times (time to recurrence, time to death) from the entry of the patient in the study (in days), and

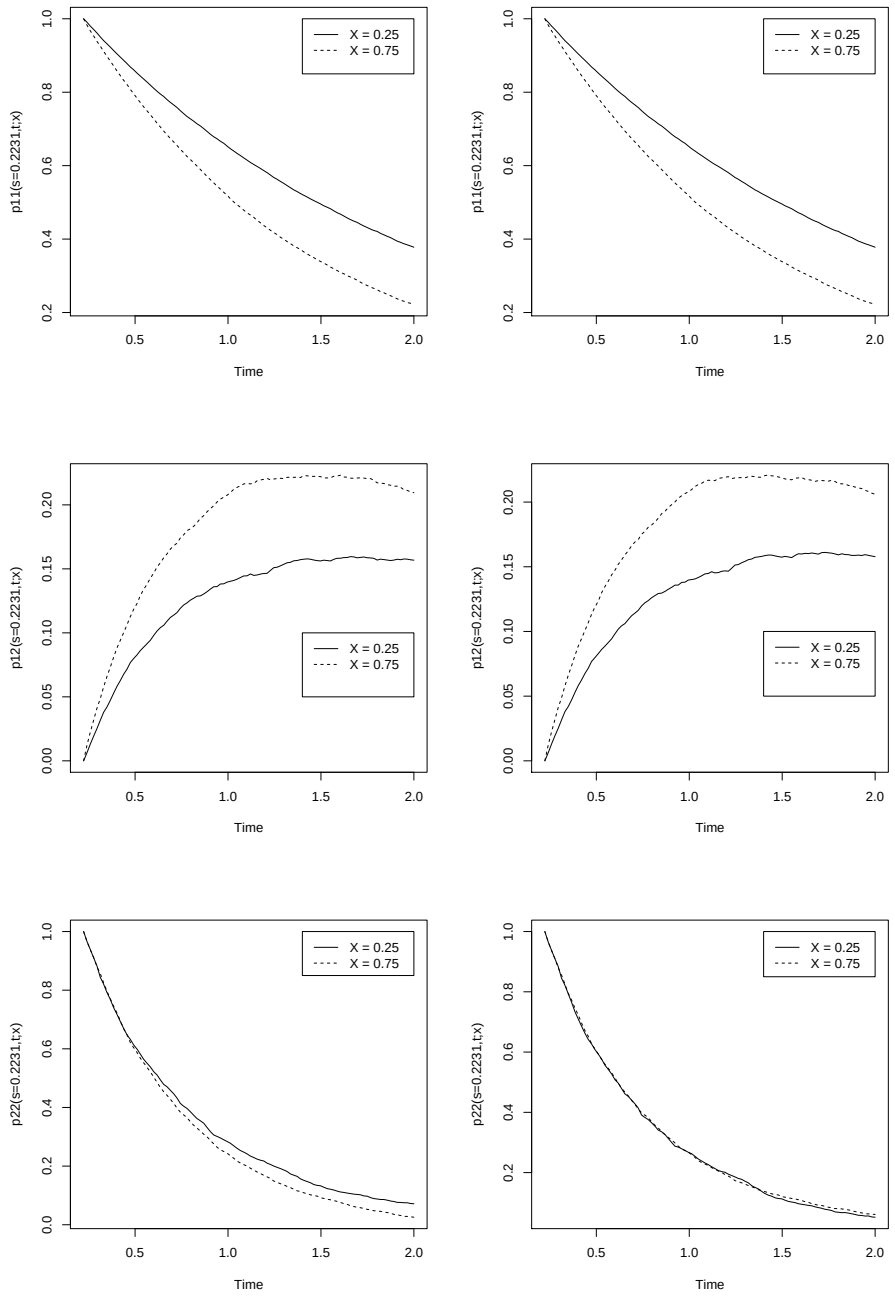

Fig. 2 Conditional transition probabilities $P_{hj}(s,t;X)$ based on simulated data. IPCW method (left hand-side) and LIN-based method (right hand-side)

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		2.6486	4.3377	5.8970
	LIN-based			2.7429	4.6313	6.2518
	IPCW	0.5108		—	4.4952	7.7858
	LIN-based			—	4.7285	8.2572
	IPCW	0.9163		—	—	10.1581
	LIN-based			—	—	10.6292
n=200	IPCW	0.2231		1.6855	2.8087	3.8764
	LIN-based			1.7269	2.9663	4.0430
	IPCW	0.5108		—	3.0378	5.1289
	LIN-based			—	3.2022	5.4715
	IPCW	0.9163		—	—	6.7937
	LIN-based			—	—	7.0697

Table 2 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{12}(x; s, t)$ along 1,000 trials for different sample sizes

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		80.6841	75.8523	53.3070
	LIN-based			85.6907	76.9551	32.9987
	IPCW	0.5108		—	58.0651	47.9647
	LIN-based			—	62.4197	47.3690
	IPCW	0.9163		—	—	58.4868
	LIN-based			—	—	59.7247
n=200	IPCW	0.2231		56.5257	54.6998	36.4748
	LIN-based			60.3715	57.4587	28.3083
	IPCW	0.5108		—	40.2206	33.0083
	LIN-based			—	42.5663	31.7394
	IPCW	0.9163		—	—	41.6085
	LIN-based			—	—	42.5566

Table 3 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{22}(x; s, t)$ along 1,000 trials for different sample sizes

a vector of covariates including *age* (in years), *nodes* (number of lymph nodes with detectable cancer) and *recurrence* (coded as 1 = yes; 0 = no) were recorded. The covariate *recurrence* is a time-dependent covariate which can be used to identify an intermediate event in an illness-death model with states "Alive and disease-free", "Alive with recurrence" and "dead".

Using a 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 (p-value < 0.001). This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set and that, consequently, Aalen-Johansen type estimators should not be used. In this section we will present estimated transition probabilities conditionally on current or past covariate measures such as *age* or *nodes* (minimum = 0 and maximum = 13). These estimators were calculated using the IPCW method and/or LIN-based procedures as explained above. Both approaches do not assume the process to be Markovian, allowing for dependent censoring and flexible (i.e. nonparametric) covariate effects otherwise.

Figures 3 and 4 depict respectively the IPCW estimates of $p_{11}(x; 379, 1000)$ and $p_{12}(x; 379, 1000)$ as functions of the covariate *age* together with a 95% pointwise confidence bands based on simple bootstrap which resamples each datum with

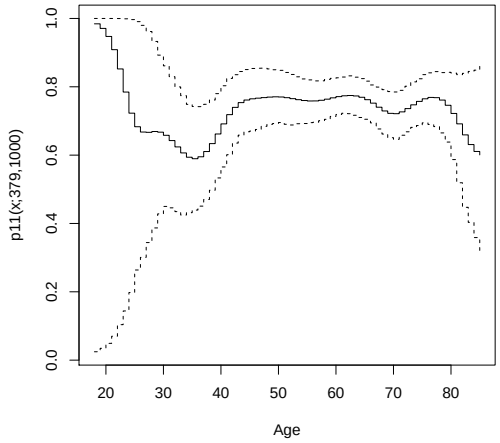

Fig. 3 Evolution of the transition probability $p_{11}(379, 1000)$ along the covariate age with 95% bootstrap confidence bands. Colon cancer data.

probability $1/n$. In both plots it is seen that these curves are not constant; the effects of *age* depicted in these plots, which are purely nonparametric, indicate the real influence of this covariate in the survival prognosis. In fact, it would not be possible to include an horizontal line within the confidence bands of Figure 4, suggesting a significative influence of age on survival. More specifically, patients near forties have a larger probability of recurrence than older patients. This is in agreement with Figure 5 where it is shown, among other things, that 40 years old patients have a higher probability of recurrence than patients with 68 years (bottom-left plot). In Figure 6 we present similar plots for the covariate *nodes*, revealing that this covariate has also a real impact on the conditional transition probabilities.

Figures 7 and 8 report the results corresponding to the LIN-based estimator. Roughly speaking, conclusions from these plots are similar to those obtained from Figures 5 and 6, but with more jump points at large values of t. However, a particular problem of LIN-based estimator is appreciated at the bottom-left plots of Figures 7 and 8, because the displayed curves for $p_{22}(x; s, t)$ are not monotone decreasing in t and, therefore, they are not admissible. This is a consequence of the specific reweighting of the data which is used in this approach, which may lead to problems of interpretation at the right tail of the distribution. Similar problems were found in Lin-estimator of the bivariate distribution function. The problem with Lin's estimator (Lin et al (1999)) is that isn't a proper bivariate distribution, in the sense that it doesn't attaches positive mass to each pair of recorded gap times. A proper estimator for the conditional transition probability $p_{22}(x; s, t)$ can be obtained when s is fixed. This is obtained by keeping the estimator constant until it starts decreasing again. However, it is difficult to deal with the general situation for any times s and t, $s \leq t$. Such issues do not arise with the IPCW method.

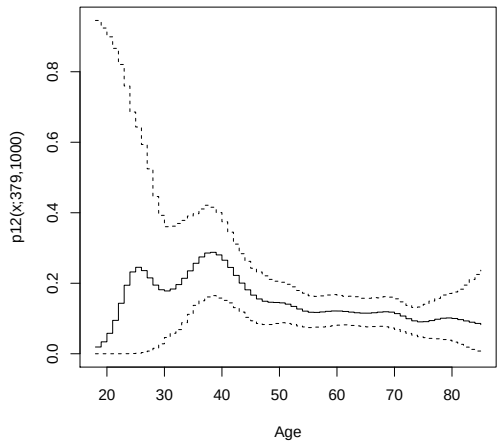

Fig. 4 Evolution of the transition probability $p_{12}(379, 1000)$ along the covariate age with 95% bootstrap confidence bands (IPCW method). Colon cancer data.

5 Conclusions and final remarks

There have been several recent contributions for the estimation of the transition probabilities in the context of multi-state models. However, most of the approaches assume independent censoring and do not account for the influence of covariates. In this paper we have proposed two estimation methods for the transition probabilities given a continuous covariate. Both methods are based on local smoothing which is introduced using regression weights. Two different schemes of inverse censoring probability reweighting have been used to deal with right censoring. In one approach, the corresponding estimator (reweighting) is based on observations that are fully observed till death (IPCW estimator), whereas the other estimator is based on observations that were uncensored till a given time (LIN-based estimator).

The methods implemented in this paper can be computationally demanding. In particular, the use of bootstrap resampling techniques are time-consuming processes because it is necessary to estimate the model a great number of times. From the point of view of computational time cost, the LIN-based estimator is the best among the two estimators without any covariate while both have similar computational time cost when accounting the influence of covariates. To obtain the point estimation and the bootstrap confidence bands, we developed an R based package called TPmsm which among other estimators implements the two methods (with and without covariates) implemented in this paper. This package is available from (http://cran.r-project.org/web/packages/TPmsm/) and provides both numerical and graphical output for all methods with considerably fast computing, even with large sample sizes.

We have investigated the performance of the estimators through simulations, showing that they are valid even when the system is non-Markov or condition-

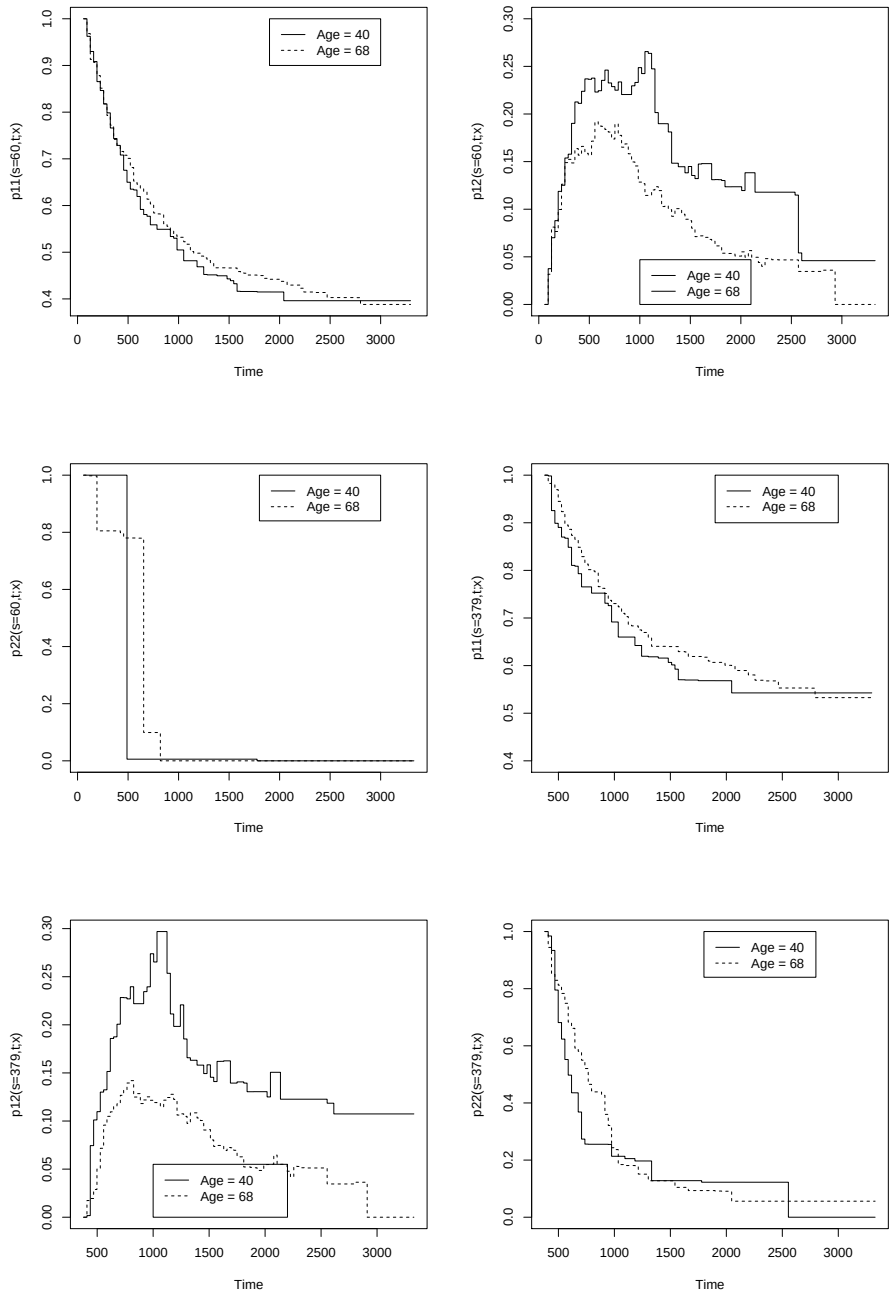

Fig. 5 Conditional transition probabilities for the colon cancer data (IPCW method) for $age = 40$ and $age = 68$.

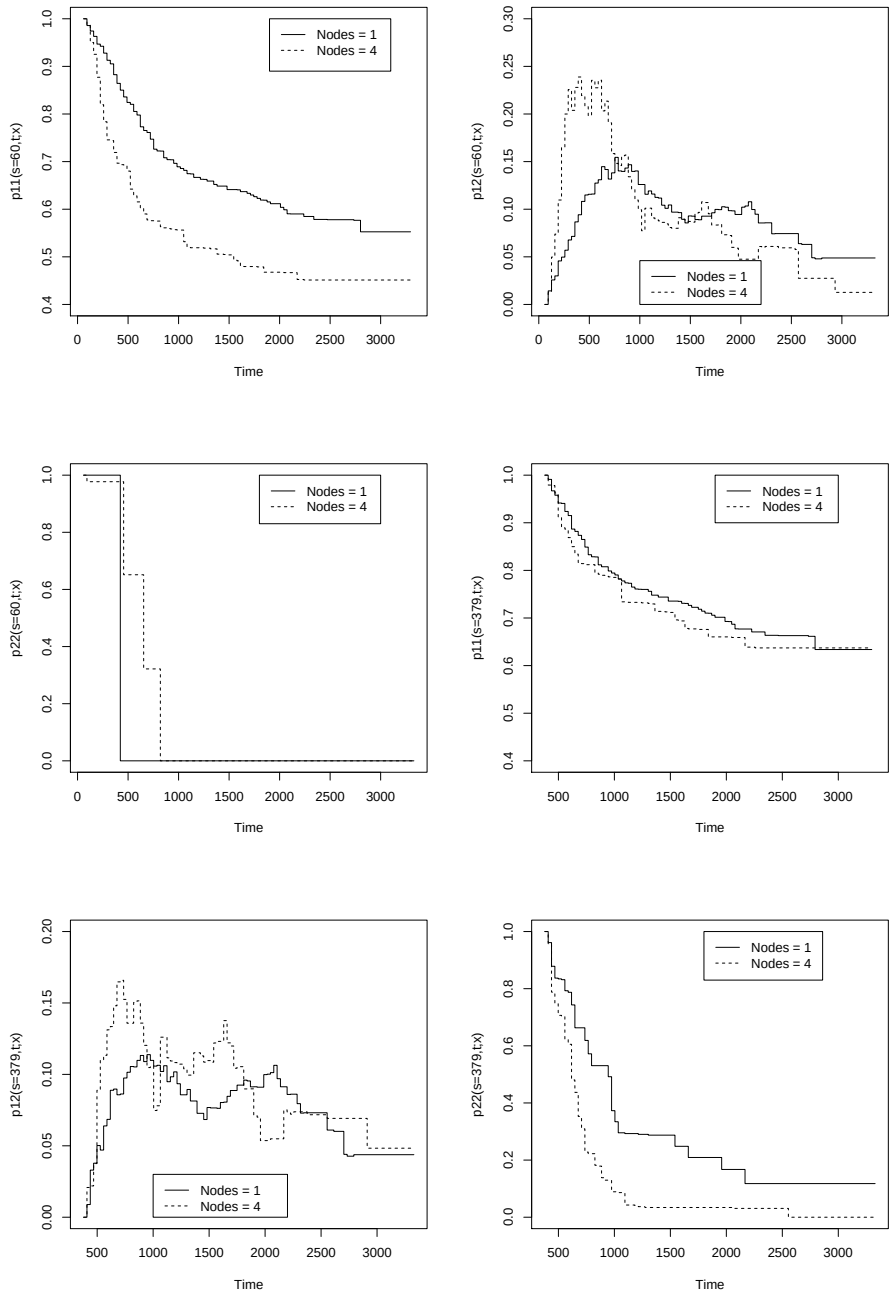

Fig. 6 Conditional transition probabilities for the colon cancer data (IPCW method) for $nodes = 1$ and $nodes = 4$.

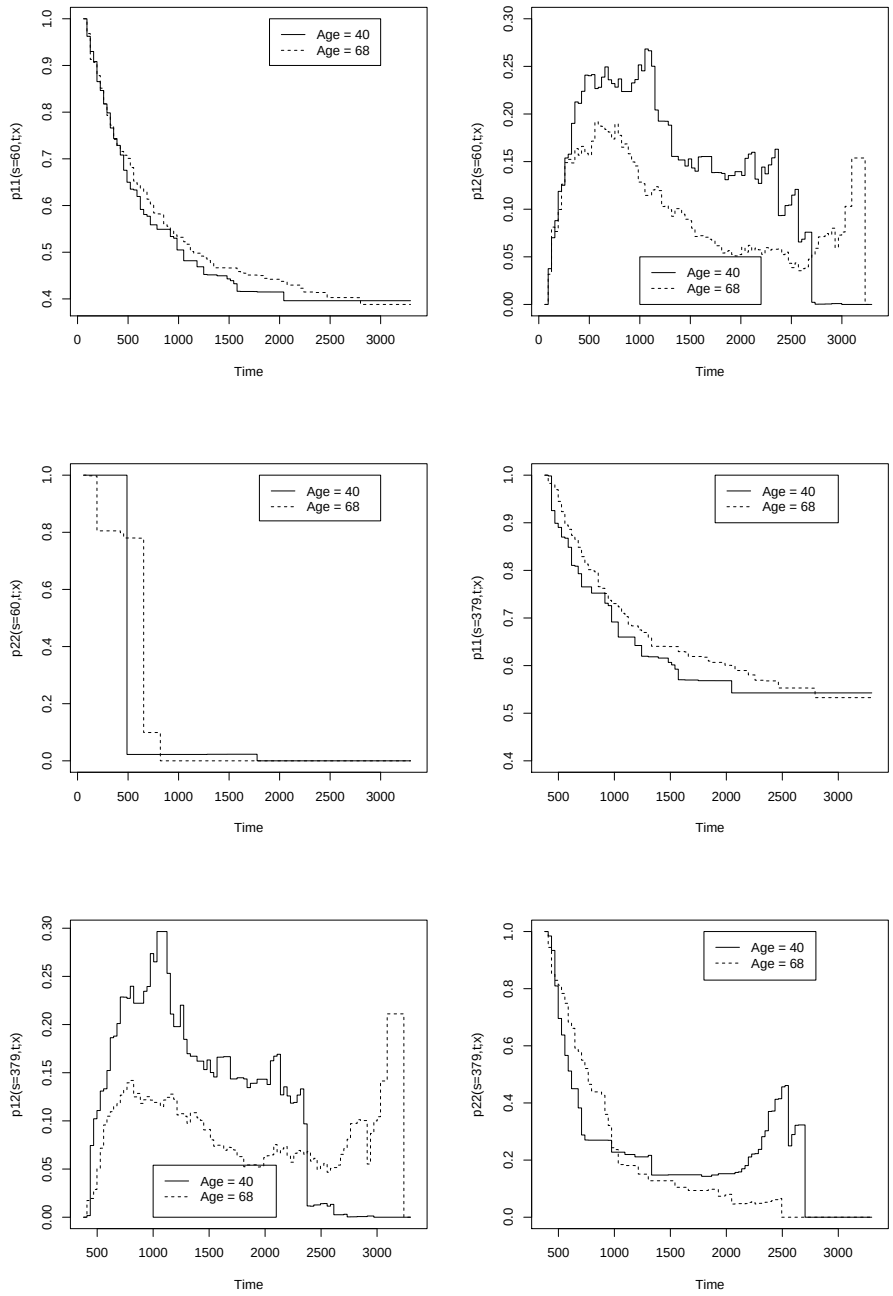

Fig. 7 Conditional transition probabilities for the colon cancer data (LIN-based method) for $age = 40$ and $age = 68$.

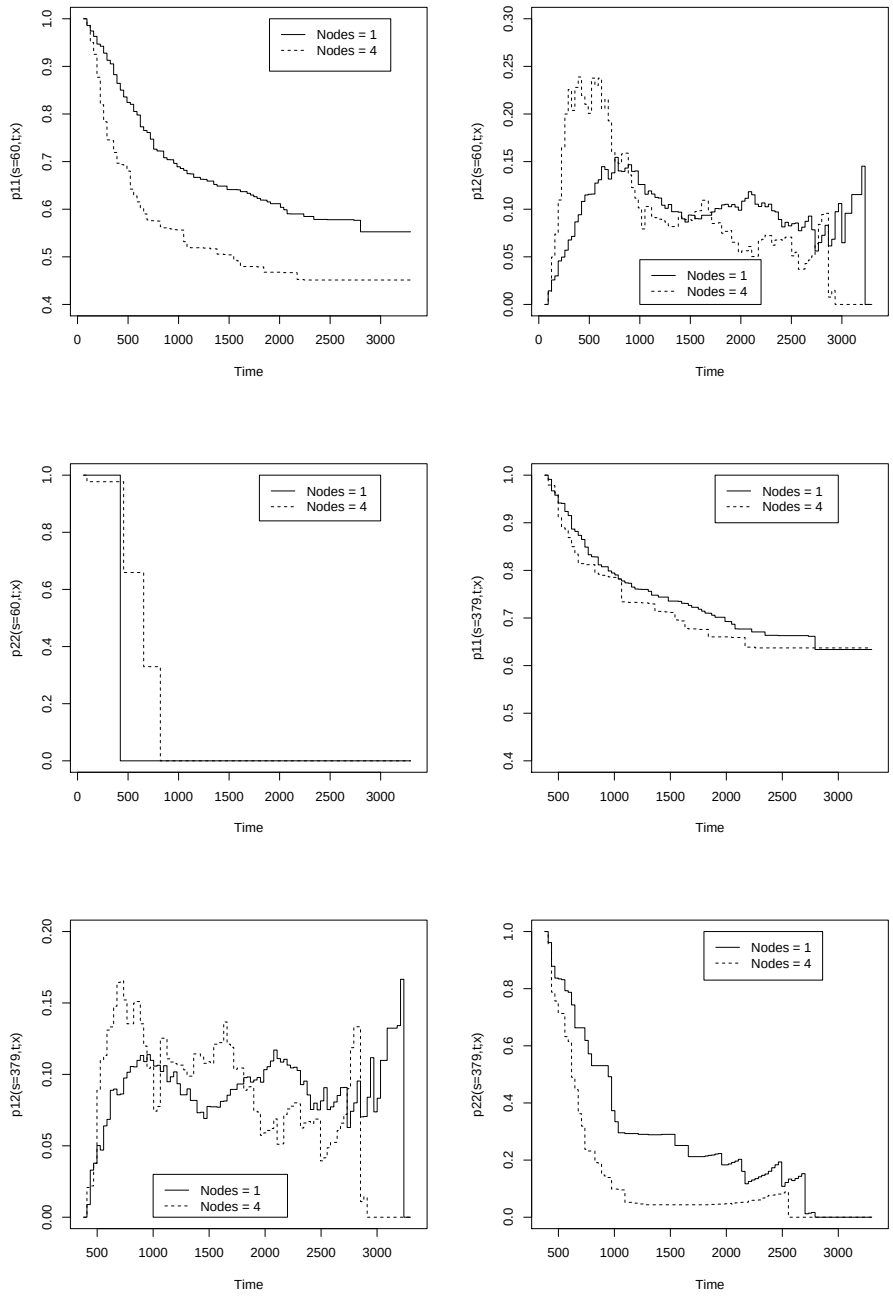

Fig. 8 Conditional transition probabilities for the colon cancer data (LIN-based method) for $nodes = 1$ and $nodes = 4$.

ally non-Markov. Simulation results show that the general performance difference between the two methods is quite small, and both methods perform quite well. Results also show that the IPCW method leads to better results for the conditional transition probability $p_{12}(s,t \mid X)$ while neither one seems to be uniformly the best for estimating $p_{22}(s,t \mid X)$. We have also illustrated the proposed methodology using real data. The analysis of the real data revealed that one of the two approaches (LIN-based one) has the drawback of occasionally providing non-monotone curves for transition probabilities which are indeed monotone and, therefore, its practical use could be less recommended. Some modification of the LIN-based estimator can be implemented to produce proper estimators of the conditional transition probabilities, specially, $p_{22}(s,t|X)$. This can be obtained for a fixed s by keeping the estimator constant until it starts decreasing again. However, it is difficult to deal with the general situation for any times s and t, $s \leq t$.

As we demonstrate, these estimates provide useful data summaries and will typically be evaluated at values of the covariate that are well inside the range of the covariate distribution. If for some applications it becomes necessary to evaluate such conditional estimators at a covariate value that is on the boundary of the covariate distribution, one could easily modify them by using a modified Beran's estimator either using a one sided kernel or by some other boundary kernel regression method (Kyung-Joon and Schucany (1998)) to obtain a better estimate.

An interesting open question is if this idea can be generalized (and how) to more complex multi-state models; this is left to future research. Another issue is the application of the proposed methods to multiple covariates. Although this could be formally done, the practical performance of the estimators heavily depend on the dimensionality. We note however that the proposed methods can accommodate discrete covariates in addition to a continuous one by splitting the sample for each level of the covariate and repeating the described procedures for each subsample. The presence of a moderate or large set of factors could recommend the application of some semiparametric technique to avoid the curse of dimensionality. Feasible solutions to this problem will be explored in the future.

Acknowledgements This research was financed by FEDER Funds through Programa Operacional Factores de Competitividade COMPETE and by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia, within Projects Est-C/MAT/UI0013/2011 and PTDC/MAT/104879/2008. We also acknowledge financial support from the project Grants MTM2008-03129 and MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación and 10PXIB300068PR of the Xunta de Galicia. Partial support from a grant from the US National Security Agency (H98230-11-1-0168) is greatly appreciated. We thank the reviewers and the AE for their constructive comments.

Appendix: Additional simulation results

In this section we give the additional simulation results for the two estimators (IPCW and LIN-based) using local linear weights instead of NW weights. The results were obtained using the `dpik` function which is available from the `R KernSmooth` package. See Tables 4 to 6 below. Results for independent censoring were also obtained (not shown), leading to similar conclusions to those shown in Section 3 and in this Appendix.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		4.6502	7.9205	9.2364
	LIN-based			4.7995	8.0749	9.3405
	IPCW	0.5108		—	7.7655	12.2793
	LIN-based			—	8.6041	12.3998
	IPCW	0.9163		—	—	17.5301
	LIN-based			—	—	17.6816
n=200	IPCW	0.2231		3.3500	5.6123	6.8104
	LIN-based			3.5076	5.7783	6.9609
	IPCW	0.5108		—	5.7970	9.1641
	LIN-based			—	5.9530	9.3326
	IPCW	0.9163		—	—	12.5984
	LIN-based			—	—	12.9697

Table 4 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{11}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		2.9452	4.8950	6.6629
	LIN-based			3.0569	5.2002	7.0611
	IPCW	0.5108		—	5.0906	8.7010
	LIN-based			—	5.5323	10.8614
	IPCW	0.9163		—	—	11.7334
	LIN-based			—	—	12.2587
n=200	IPCW	0.2231		2.0465	3.3456	4.7253
	LIN-based			2.0914	3.5387	4.9551
	IPCW	0.5108		—	3.6667	6.1847
	LIN-based			—	3.8777	6.6768
	IPCW	0.9163		—	—	8.3759
	LIN-based			—	—	9.1652

Table 5 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{12}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

			t	0.5108	0.9163	1.6094
		s				
n=100	IPCW	0.2231		85.9702	81.9559	59.0274
	LIN-based			91.5632	81.1756	34.1622
	IPCW	0.5108		—	63.7464	53.9567
	LIN-based			—	68.2274	50.5756
	IPCW	0.9163		—	—	64.2243
	LIN-based			—	—	65.0858
n=200	IPCW	0.2231		64.5395	63.1126	44.4921
	LIN-based			68.7074	64.0163	30.5495
	IPCW	0.5108		—	47.7855	41.1531
	LIN-based			—	50.0397	36.5208
	IPCW	0.9163		—	—	49.3859
	LIN-based			—	—	49.8701

Table 6 IMSE (×1000) of the estimated transition probabilities $\hat{p}_{22}(x; s, t)$ along 1,000 trials for different sample sizes. Estimates based on the local linear estimators.

References

Aalen O, Johansen S (1978) An empirical transition matrix for non homogeneous markov and chains based on censored observations. Scandinavian Journal of

Statistics 5:141–150
Akritas M (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. Annals of Statistics 22:1299–1327
Altman N, Leger C (1995) Bandwidth selection for kernel distribution function estimation. Journal of Statistical Planning and Inference 46:195–214
Amorim A, de Uña-Álvarez J, Meira-Machado L (2011) Presmoothing the transition probabilities in the illness-death model. Statistics & Probability Letters 81(7):797–806
Andersen P, Borgan Ø, Gill R, Keiding N (1993) Statistical Models Based on Counting Processes. Springer-Verlag, New York
Arjas E, Eerola M (1993) On predictive causality in longitudinal studies. Journal of Statistical Planning and Inference 34:361–384
Beran R (1981) Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley
Dabrowska D (1987) Non-parametric regression with censored survival data. Scandinavian Journal of Statistics 14:181–197
Dabrowska D, Ho W (2000) Confidence bands for comparison of transition probabilities in a markov chain model. Lifetime Data Analysis 6(1):5–21
Dabrowska D, Lee W (1996) Nonparametric estimation of transition probabilities in a two-stage duration model. Journal of Nonparametric Statistics 7:75–103
Kaplan E, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481
Kyung-Joon C, Schucany W (1998) Nonparametric kernel regression estimation near endpoints. Journal of Statistical Planning and Inference 66:289–304
Lin D, Sun W, Ying Z (1999) Nonparametric estimation of the gap time distributions for serial events with censored data. Biometrika 86:59–70
Meira-Machado L, Roca-Pardiñas J (2011) p3state.msm: Analyzing survival data from an illness-death model. Journal of Statistical Software 38:1–18
Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C (2006) Nonparametric estimation of transition probabilities in a non-markov illness-death model. Lifetime Data Analysis 12:325–344
Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen P (2009) Multistate models for the analysis of time to event data. Statistical Methods in Medical Research 18:195–222
Moertel C, Fleming T, McDonald ea JS (1990) Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. New England Journal of Medicine 322:352–358
Mostajabi F, Datta S (2012) Nonparametric regression of state occupation, entry, exit and waiting times with multistate right censored data, statistics in Medicine, to appear
Nadaraya E (1965) On nonparametric estimates of density functions and regression curves. Theory of Applied Probability 10:186–190
Satten G, Datta S (2001) The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. The American Statistician 55:3:207–210
Van Keilegom I (2004) A note on the nonparametric estimation of the bivariate distribution under dependent censoring. Journal of Nonparametric Statistics 16:659–670
Van Keilegom I, Akritas M, Veraverbeke N (2001) Estimation of the conditional distribution in regression with censored data: a comparative study. Computa-

tional Statistics and Data Analysis 35:487–500
Wand M, Jones M (1997) Kernel Smoothing. Chapman & Hall, London
Watson G (1964) Smooth regression analysis. Sankhya 26:15:175–184