RESEARCH ARTICLE | JUNE 07 2024

Presmoothed estimators of the state occupation probabilities in multi-state survival data

Luís Meira-Machado ✉; Gustavo Soutinho

AIP Conf. Proc. 3094, 470003 (2024)
https://doi.org/10.1063/5.0210139

Presmoothed Estimators of the State Occupation Probabilities in Multi-state Survival Data

Luís Meira-Machado[1,a)] and Gustavo Soutinho[2,b)]

[1]*Centre of Mathematics, University of Minho, Campus de Azurém, 4800 - 058 Guimarães, Portugal.*
[2]*EPIUnit - University of Porto, Rua das Taipas 135, 4050-600 Porto, Portugal*

[a]Corresponding author: lmachado@math.uminho.pt
[b]gdsoutinho@gmail.com

Abstract. The progress of a disease can be analyzed using multistate models. These models focus on two key parameters of interest: the transition hazard and the state occupation probabilities. The state occupation probabilities have been consistently estimated by the Aalen-Johansen estimator. This estimator is particularly well-suited for handling censoring and benefits from the Markov assumption in the underlying stochastic process. In some cases, these estimators may lead to estimators with higher variability. To mitigate this issue we propose alternative estimators that incorporate a preliminary estimation approach. We introduce also practical estimation techniques for the state occupation probabilities, considering covariate measures. We explore the finite sample behavior of the estimators through simulations. An application to breast cancer is included.

INTRODUCTION

A multi-state model (Meira-Machado and Sestelo, 2019) is a mathematical framework designed to describe a continuous stochastic process where individuals can transition between a finite set of states. One of the most recognized and commonly used model is the illness-death model. In its irreversible form, this model consists of three states: individuals initially start in the 'alive and disease-free' state and subsequently transition to either the 'diseased' state or the 'dead' state. In this model, once individuals enter the 'diseased' state, there is no possibility of recovery, and they ultimately progress to the 'dead' state.

An essential characteristic of multi-state models lies in their capacity to provide predictions regarding a patient's clinical prognosis at specific stages of the illness. The state occupation probabilities offer a means to capture various facets of the model's dynamics. Traditionally, these probabilities have been estimated using the Aalen-Johansen estimator (Aalen and Johansen, 1978), which is tailored to handle censoring and leverages the Markovian assumption in the underlying stochastic process. In the context of the illness-death model, an alternative estimator introduced by Pepe (1991) involves computing the difference between two Kaplan-Meier estimates. However, when dealing with small sample sizes or heavily censored data, both of these estimators can yield results with considerable variability. To address this challenge, researchers have explored the use of estimators that incorporate a preliminary estimation approach, known as 'presmoothing' for the probability of censoring. Consequently, in this work, we propose a modification of Pepe's estimator that incorporates presmoothing techniques.

ESTIMATORS

A multi-state model is a model for a time continuous stochastic process $(Y(t), t \in \mathcal{T})$ which at any time occupies one of a few possible states. Here, $\mathcal{T} = [0, \tau]$ or $[0, \tau)$ with $\tau \leq +\infty$. In this paper, we focus on the progressive illness-death model comprising States 0, 1, and 2. This model's essence lies in the joint distribution of (Z, T), where Z denotes the time spent in the initial state, and T represents the overall survival time. Under censoring, the information available is limited to the censored versions of both Z and T, complemented by their respective censoring indicators.

International Conference of Numerical Analysis and Applied Mathematics
AIP Conf. Proc. 3094, 470003-1–470003-4; https://doi.org/10.1063/5.0210139
Published under an exclusive license by AIP Publishing. 978-0-7354-4954-1/$30.00

Let's introduce the concept of state occupation probabilities, denoted as $p_k(t)$, which represent the probability that the process $Y(t)$ is in state k at time t, and for which three specific probabilities exist: $p_0(t)$, $p_1(t)$ and $p_2(t)$ with $p_0(t) + p_1(t) + p_2(t) = 1$. In 1978, Aalen and Johansen introduced a nonparametric approach for estimating state occupation probabilities. Their estimation technique extends the widely recognized Kaplan–Meier estimator (Kaplan and Meier, 1958) to the domain of Markov chains. Explicit formulae for these quantities can be found, for example, in Moreira, de Uña-Álvarez and Meira-Machado (2013). It is worth mentioning that Datta and Satten (2001) demonstrated that the estimator of state occupation probabilities, derived from the non-parametric Aalen–Johansen estimator, remains consistent, even in the case of non-Markov multi-state models.

The state occupation probabilities can also be expressed in terms of the pair (Z, T) as follows (Pepe, 1991): $p_0(t) = P(Z > t)$, $p_1(t) = P(Z \leq t, T > t) = P(T > t) - P(Z > t) = S(t) - S_0(t)$ and $p_1(t) = P(T \leq t) = 1 - S(t)$. These quantities can be estimated as follows: $\hat{p}_0(t) = \hat{S}_0(t)$, $\hat{p}_1(t) = \hat{S}(t) - \hat{S}_0(t)$ and $\hat{p}_1(t) = 1 - \hat{S}(t)$, where $\hat{S}_0(t)$ is the Kaplan-meier estimator of the disease-free survival function and $\hat{S}(t)$ is the Kaplan-Meier estimator of survival the total time.

The Aalen-Johansen estimator and Pepe's estimators can lead to estimators with high variability (in particular at the right hand side of the distribution) for small sample sizes or high censoring. Preliminary smoothing, also known as presmoothing, is a good alternative to those situations. The fundamental concept behind presmoothing is to substitute each censoring indicator with a smoothed representation obtained through binary regression analysis with observable variables. The adoption of presmoothed estimators proves to be a valuable alternative in such scenarios, as they give mass to all the event times. The new estimators are thus built using a procedure based on (differences between) presmoothed Kaplan–Meier estimators. One useful parametric candidate for the binary regression function $p(t) = P(\Delta = 1|\widetilde{T} = t)$ belongs to the logit or probit family of binary regression curves. When the parametric model specified for $p(t)$ is accurate, the corresponding semiparametric presmoothed estimator exhibits, at minimum, the same level of efficiency as the original nonparametric unsmoothed estimator. The suitability of a particular model for the presmoothing function can be assessed either graphically or formally, employing goodness-of-fit tests like the one proposed by Hosmer and Lemeshow (1989) for the logistic model, or a Kolmogorov-Smirnov type version of the model-based bootstrap approach. This means that practical control over the risk of using a misspecified model is achievable. Nonparametric presmoothing becomes particularly valuable when there's a distinct concern about the misspecification of the parametric model. In these cases, the Nadaraya-Watson kernel estimator for $p(t)$ is a viable alternative.

Another goal is to estimate these probabilities conditionally on covariate measures. A common approach, especially effective when dealing with multiple covariates, is to utilize estimators based on a Cox's regression model fitted independently to each transition, with the corresponding baseline hazard function estimated using Breslow's method. However, in situations where there is limited knowledge about the data, a nonparametric approach should be taken into account. In such cases, one can adopt the method outlined in Meira-Machado, de Uña-Álvarez, and Datta (2015).

Techniques presented in this article can be readily extended to a broader range of multi-state models. The needed details to explore this possibility are given in Section 5 of the paper by de Uña-Álvarez and Meira-Machado (2015).

SIMULATION STUDY

This section presents the simulation results, which serve to assess the effectiveness of the provided estimators. In particular, we compare the Aalen-Johansen estimator (labeled `AJ`) with nonparametric unsmoothed Pepe's estimator (labeled `PP`), and its presmoothed estimators, the semiparametric estimator using preliminary smoothing based on the logistic model (labeled as `smPP`) and the method based on nonparametric presmoothing (labeled as `npPP`). We conducted simulations using data that mimicked a progressive illness-death model. In this simulated scenario, all individuals initiated in the initial state (State 0) at time $t = 0$. The movement of these individuals was determined following the scenario outlined by Moreira et al. (2013) and Araújo, Roca-Pardiñas, and Meira-Machado (2014), where subjects had the option to pass through the intermediate state (State 1) at some point or proceed directly to the absorbing state (State 2). For those individuals transitioning through the intermediate state, we generated replicates of (Z, T) using Gumbel's bivariate exponential distribution. We generated random censoring times independently from uniform distributions $U \sim U[0, 3]$. State occupation probabilities were computed at specific time points corresponding to the percentiles of 20%, 40%, 60%, and 80% of the exponential marginal distribution functions with a rate parameter of 1. In each simulation, we created 1000 samples, with three sample sizes. For each of these samples, we calculated the mean across all generated datasets. To measure efficiency, we utilized the Mean Squared Error (MSE), but we also computed standard deviations (SD) and bias for each time point t.

Results

Table 1 presents the results of the four estimators obtained from the previously described simulation studies. The results indicate that the accuracy of the estimators diminishes as t increases, primarily due to heightened censoring effects observed at the right tail of the lifetime distribution. As anticipated, the standard deviation (SD) decreases with larger sample sizes and lower levels of censoring. Reported results reveal that the performance of AJ and PP estimators is quite similar for all times and all state occupation probabilities. The semiparametric presmoothed estimator, labeled as smPP, shows lower variability. The good performance exhibited by this estimator is shown in the relative MSE.

TABLE 1. Summary of Bias, Standard Deviation (SD), and relative Mean Squared Errors (MSE) for $p_k(t)$ estimators. The relative MSEs are also given.

		$\widehat{p}_0^{\mathtt{AJ}}(t)$	$\widehat{p}_0^{\mathtt{PP}}(t)$	$\widehat{p}_0^{\mathtt{smPP}}(t)$	$\widehat{p}_0^{\mathtt{npPP}}(t)$	Relative MSE		
		bias (SD)	bias (SD)	bias (SD)	bias (SD)	AJ/PP	AJ/smPP	AJ/npPP
t	n							
0.2231	50	-0.0020 (0.058)	0.0021 (0.058)	-0.0061 (0.054)	0.0277 (0.053)	1.0000	1.1140	0.9322
	100	0.0010 (0.040)	0.0007 (0.040)	-0.0026 (0.037)	0.0316 (0.036)	1.0000	1.1679	0.7189
	250	0.0010 (0.026)	0.0011 (0.026)	-0.0048 (0.024)	0.0303 (0.022)	1.0000	1.1226	0.4651
		$\widehat{p}_1^{\mathtt{AJ}}(t)$	$\widehat{p}_1^{\mathtt{PP}}(t)$	$\widehat{p}_1^{\mathtt{smPP}}(t)$	$\widehat{p}_1^{\mathtt{npPP}}(t)$	Relative MSE		
		bias (SD)	bias (SD)	bias (SD)	bias (SD)	AJ/PP	AJ/smPP	AJ/npPP
t	n							
0.5108	50	-0.0011 (0.061)	-0.0012 (0.061)	-0.0072 (0.053)	-0.0031 (0.055)	1.0016	1.3000	1.2158
	100	0.0010 (0.040)	0.0010 (0.040)	-0.0059 (0.036)	-0.0009 (0.037)	1.0036	1.2434	1.1778
	250	0.0008 (0.027)	0.0008 (0.027)	-0.0060 (0.023)	0.0002 (0.025)	1.0011	1.2831	1.1596
		$\widehat{p}_2^{\mathtt{AJ}}(t)$	$\widehat{p}_2^{\mathtt{PP}}(t)$	$\widehat{p}_2^{\mathtt{smPP}}(t)$	$\widehat{p}_2^{\mathtt{npPP}}(t)$	Relative MSE		
		bias (SD)	bias (SD)	bias (SD)	bias (SD)	AJ/PP	AJ/smPP	AJ/npPP
t	n							
0.9163	50	-0.0010 (0.076)	-0.0015 (0.076)	-0.0044 (0.071)	-0.0068 (0.074)	0.9914	1.1316	1.0525
	100	0.0010 (0.053)	0.0012 (0.054)	-0.0011 (0.050)	-0.0014 (0.052)	0.9806	1.1480	1.0315
	250	-0.0014 (0.035)	-0.0014 (0.035)	-0.0025 (0.031)	-0.0024 (0.034)	0.9816	1.2114	1.0531

For the sake of comprehensiveness, Figure 1 displays boxplots illustrating the estimates of state occupation probabilities derived from 1000 Monte Carlo replicates using the four sets of estimators. These boxplots corroborate our findings and are consistent with the results in Table 1. From these plots, it can be seen the lower variability of the presmoothed estimators for all state occupation probabilities. However, a large bias can be observed for the npPP estimator when estimating the $p_0(t)$. This poor performance occurred at all three sample sizes. In contrast, the presmoothed estimator based on a parametric logistic regression model had a small bias while revealing low variability in all cases. The good performance of the smPP seems to be independent of the state occupation probability, which one aims to estimate.

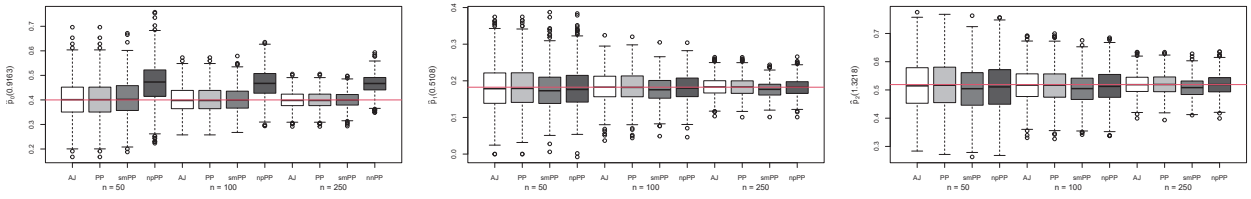

FIGURE 1. Boxplots summarizing 1000 estimates of the transition probabilities. The solid red horizontal line represents the true transition probability.

A more detailed comparison of the proposed estimators and additional simulation results will be published elsewhere.

APPLICATION TO BREAST CANCER DATA

Our methodology is motivated by the analysis of German breast cancer data, which includes 686 women tracked from breast cancer diagnosis to censoring or breast cancer-related death. Among them, 299 had recurrences, and 171 died.

These data fit an illness-death model with three states: 'Alive and disease-free,' 'Alive with Recurrence,' and 'Dead'. In this section, we present plots using all proposed state occupation probability estimation methods.

Figure 2 displays estimates of state occupation probabilities for three states, offering insights into the state occupied by the process after surgery. The first column presents recurrence-free survival fractions, while the second column illustrates the probability of remaining alive with a recurrence over time. As the recurrence state is transient, this curve initially rises and then may fall. The third column depicts one minus the survival fraction over time. In Figure 2, the presmooth estimators exhibit expected behavior similar to nonparametric unsmoothed estimators but with reduced variability in the right tail of the distribution.

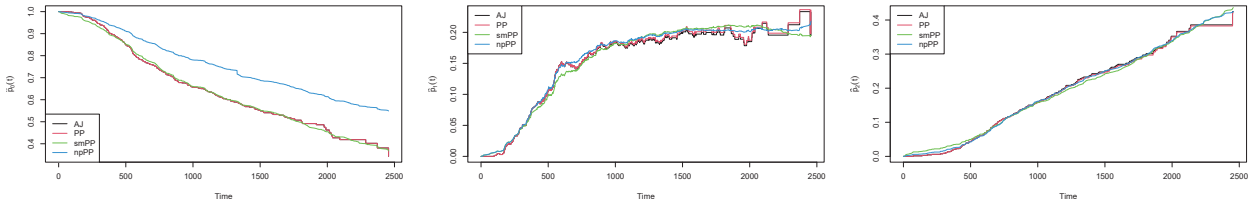

FIGURE 2. Estimated state occupation probabilities. Breast cancer data.

DISCUSSION

This article investigated the relative performance of presmoothed estimators for state occupation probabilities. The findings suggest that the proposed estimators can offer competitive alternatives that might surpass the original estimators, yielding estimates with reduced variability. It's important to note that presmoothing may introduce some bias into estimation while reducing variance. This bias tends to be more pronounced when the selected parametric model is misaligned with the data. However, practical control over the risk of substantial bias due to model mispecification can be achieved through the application of goodness-of-fit tests to validate the chosen model. Nonparametric presmoothing proves valuable when no suitable parametric candidate exists for the presmoothing function. Similar findings were obtained in the paper by Soutinho, Meira-Machado and Oliveira (2022) but for a different target.

ACKNOWLEDGMENTS

This research was financed within the research grants PTDC/MAT-STA /28248/2017 and PD/BD/142887/2018.

REFERENCES

[1] P. K. Andersen, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes* (Springer-Verlag, New York, 1993).
[2] O. O. Aalen and S. Johansen, Scandinavian Journal of Statistics **5**, 141–150 (1978).
[3] A. A. Araújo, J. Roca-Pardiñas, and L. Meira-Machado, Journal of Statistical Software **62(4)**, 1–29 (2014).
[4] S. Datta and G. A. Satten, Statistics & Probability Letters **55(4)**, 403–411 (2001).
[5] J. de Uña-Álvarez and L. Meira-Machado, Biometrics **71(2)**, 364–375 (2015).
[6] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (John Wiley & Sons, New York, 1989).
[7] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen, Statistical Methods in Medical Research **18(2)**, 195–222 (2009).
[8] L. Meira-Machado and M. Sestelo, Biometrical Journal **61(2)**, 245–263 (2019).
[9] L. Meira-Machado, J. de Uña-Álvarez, and S. Datta, Computational Statistics **30**, 377–397 (2015).
[10] A. Moreira, J. de Uña-Álvarez, and L. Meira-Machado, Electronical Journal of Statistics **7**, 1491–1516 (2013).
[11] M. S. Pepe, Journal of the American Statistical Association **86**, 770–778 (1991).
[12] G. Soutinho, L. Meira-Machado, and P. Oliveira, Communications in Statistics - Simulation and Computation **51**, 5202–5221 (2022).