Estimation of the transition probabilities in non-Markov models: new contributions and software development

University of Minho, Portugal
Luís Meira-Machado

38th Annual Conference of the International Society for Clinical Biostatistics, 2017, Vigo, Spain

Outline

Mortality Model

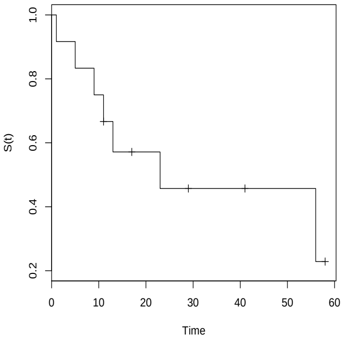
Mortality model for survival analysis.

Let T denote the survival times C a univariate right-censoring which we assume to be independent of T.

Because of censoring we only observe $(\widetilde{T}, \Delta)$ where $\widetilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$.

$S(T > y)$ may be consistently estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958):

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \equiv \prod_{i=1}^{n} \left(1 - \frac{\Delta_{[i]}}{n - i + 1}\right)^{I(\widetilde{T}_{(i)} \leq t)}$$



time:
1, 5, 9, 11, 11, 13, 17, 23, 29, 41, 56, 58
event:
1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0

Kaplan-Meier weights

$$\widehat{S}(t) = 1 - \sum_{i=1}^{n} W_i I(\widetilde{T}_{(i)} \leq t) \equiv 1 - \widehat{F}(t),$$

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

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

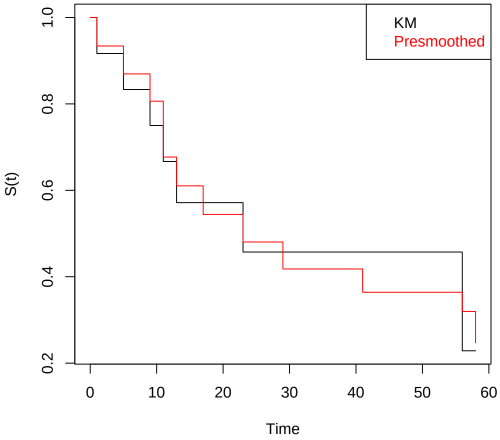
A presmoothed version of the Kaplan-Meier estimator:

$$\widetilde{S}(t) = 1 - \sum_{i=1}^{n} PW_i I(\widetilde{T}_{(i)} \leq t) \equiv 1 - \widetilde{F}(t),$$

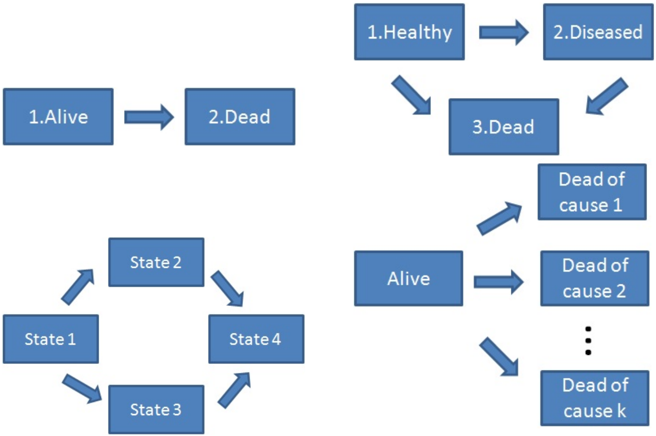
where PW_i are the presmoothed Kaplan-Meier weights:

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

Kaplan-Meier estimator



Common examples



Transition probabilities

Given two states i, j and $s < t$

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

Estimating these quantities is interesting, since they allow for long-term predictions of the process.

Markov assumption

The inference in multi-state models is traditionally performed under the Markov assumption, which states that past and future are independent given the present state.

A little of history on transition probabilities. . .

- Aalen and Johansen (SCAND. J. STAT. 1978) introduced a nonparametric estimator of $p_{ij}(s,t)$ for Markov models.
- Moreira et al (EJS 2013) propose a modification of the Aalen-Johansen estimator in the illness-death model based on presmoothing.
- Meira-Machado et al (LiDA 2006) introduce a substitute for the Aalen-Johansen estimator in the case of a non-Markov illness-death model.
- Amorim et al (SPL 2011) propose a modification of Meira-Machado et al (2006) estimator based on presmoothing, which allows for a variance reduction in the presence of censoring.
- Meira-Machado et al. (COST 2015) propose new estimators based on IPCW to deal with dependent censoring and that account for the influence of covariates.

A little of history on transition probabilities. . .

- Allignol (LiDA 2014) propose a simplified representation of the LiDA estimator in terms of the limiting probability of a particular competing risks process.
- de Uña-Álvarez and Meira-Machado (Biometrics 2015) propose new estimators based on landmarking.
- Titman (Biometrics 2015) propose a conditional Pepe (1991) estimator.
- Putter and Spitoni (SMMR 2016) propose a landmark Aalen-Johansen estimator.
- Meira-Machado (SORT 2016) propose presmoothed landmark estimators.

Illness-death model

$Z = T_{12} \wedge T_{13}$ *time in state 1*

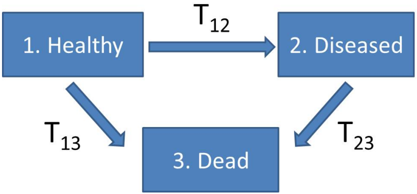
$\rho = I(T_{12} \leq T_{13})$ *indicator of visiting state 2*

$T = Z + \rho T_{23}$ *total survival time*

C *censoring time*

One observes $(\widetilde{Z}, \widetilde{T}, \triangle_1, \triangle)$ where

$$\widetilde{Z} = Z \wedge C$$
$$\triangle_1 = I(Z \leq C)$$
$$\widetilde{T} = T \wedge C$$
$$\triangle = I(T \leq C)$$



Nonparametric estimators

$$p_{11}(s,t) = P(Z > t \mid Z > s) = \frac{P(Z > t)}{P(Z > s)}$$

$$p_{12}(s,t) = P(Z \leq t, T > t \mid Z > s) = \frac{P(s < Z \leq t, T > t)}{P(Z > s)}$$

$$p_{13}(s,t) = P(T \leq t \mid Z > s) = \frac{P(Z > s, T \leq t)}{P(Z > s)}$$

$$p_{22}(s,t) = P(T > t \mid Z < s, T > s) = \frac{P(Z \leq s, T > t)}{P(Z \leq s, T > s)}$$

$$p_{23}(s,t) = P(T \leq t \mid Z < s, T > s) = \frac{P(Z < s, s < T \leq t)}{P(Z \leq s, T > s)}$$

LiDA estimators

The estimators can be written using Kaplan-Meier Weights:

$$\widehat{p}_{11}(s,t) = \frac{1 - \sum_{i=1}^{n} W_i^1 I(\widetilde{Z}_i \leq t)}{1 - \sum_{i=1}^{n} W_i^1 I(\widetilde{Z}_i \leq s)}$$

$$\widehat{p}_{12}(s,t) = \frac{\sum_{i=1}^{n} W_i I(s < \widetilde{Z}_i \leq t, \widetilde{T}_i > t)}{1 - \sum_{i=1}^{n} W_i^1 I(\widetilde{Z}_i \leq s)}$$

$$\widehat{p}_{22}(s,t) = \frac{\sum_{i=1}^{n} W_i I(\widetilde{Z}_i \leq s, \widetilde{T}_i > t)}{\sum_{i=1}^{n} W_i I(\widetilde{Z}_i \leq s, \widetilde{T}_i > s)}$$

Where W_i^1 and W_i are the Kaplan-Meier weights of Z and T, respectively.

LiDA revised expressions

To avoid problems in the right tail where uncensored data are scarce we can use the use of the following alternative expressions:

$$p_{11}(s,t) = \frac{P(Z > t)}{P(Z > s)}$$

$$p_{12}(s,t) = \frac{P(s < Z \leq t) - P(s < Z \leq t, T \leq t)}{P(Z > s)}$$

$$p_{22}(s,t) = \frac{P(Z \leq s) - P(Z \leq s, T \leq t)}{P(Z \leq s) - P(T \leq s)}$$

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

Landmark estimators

$$
\begin{aligned}
\widehat{p}_{11}(s,t) &= \widehat{S}_Z^{(s)}(t) \\
\widehat{p}_{12}(s,t) &= \widehat{S}_T^{(s)}(t) - \widehat{S}_Z^{(s)}(t) \\
\widehat{p}_{13}(s,t) &= 1 - \widehat{S}_T^{(s)}(t)
\end{aligned}
$$

where $S_Z^{(s)}$ and $S_T^{(s)}$ are the survival functions of the first sojourn time and total time, respectively; computed from the sample $\{i : \widetilde{Z}_i > s\}$.

$$\widehat{p}_{22}(s,t) = \widehat{S}_T^{[s]}(t) \qquad \widehat{p}_{23}(s,t) = 1 - \widehat{S}_T^{[s]}(t)$$

where $S_T^{[s]}$ is the survival functions of the total time computed from the sample $\{i : \widetilde{Z}_i \leq s, \widetilde{T}_i > s\}$.

Landmark estimators

Variance estimates:

- A simple bootstrap can be used to obtain variance estimates.
- Asymptotic estimates are also possible using moment-type variance estimators as developed by Pepe (1991).
- Greenwood estimator can be used for almost all transition probabilities.

A presmoothed version of the landmark estimator (SORT 2016) can be used to reduce variability of the estimator when:

- Sample size is small.
- Censoring is high.
- Higher values of s.

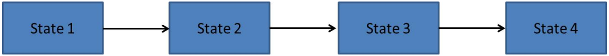
Landmark estimators: occupation probabilities

$$P_j(t) = p_{ij}(0,t), j = 1, 2, 3.$$

$$\widehat{P}_1(t) = \widehat{p}_{11}(0,t) = \widehat{S}_Z(t)$$
$$\widehat{P}_2(t) = \widehat{p}_{12}(0,t) = \widehat{S}_T(t) - \widehat{S}_Z(t)$$
$$\widehat{P}_3(t) = \widehat{p}_{13}(0,t) = 1 - \widehat{S}_T(t)$$

The estimators are very simple and intuitive. They are equivalent to Pepe's estimator (1991).

Landmark estimators: k-state progressive model



Let (T_1, T_2, T_3) denote the event times:

$$\widehat{p}_{11}(s,t) = \widehat{P}(T_1 > t | T_1 > s) = \widehat{S}_1^{(s)}(t)$$

$$\widehat{p}_{12}(s,t) = \widehat{P}(T_1 \le t, T_2 > t | T_1 > s) = \widehat{S}_2^{(s)}(t) - \widehat{S}_1^{(s)}(t)$$

$$\widehat{p}_{13}(s,t) = \widehat{P}(T_2 \le t, T_3 > t | T_1 > s) = \widehat{S}_3^{(s)}(t) - \widehat{S}_2^{(s)}(t)$$

$$\widehat{p}_{14}(s,t) = \widehat{P}(T_3 \le t | T_1 > s) = 1 - \widehat{S}_3^{(s)}(t)$$

$$\widehat{p}_{22}(s,t) = \widehat{P}(T_2 > t | T_1 \le s, T_2 > s) = \widehat{S}_2^{[s]}(t)$$

$$\widehat{p}_{23}(s,t) = \widehat{P}(T_2 \le t, T_3 > t | T_1 \le s, T_2 > s) = \widehat{S}_3^{[s]}(t) - \widehat{S}_2^{[s]}(t)$$

$$\widehat{p}_{24}(s,t) = \widehat{P}(T_3 \le t | T_1 \le s, T_2 > s) = 1 - \widehat{S}_3^{[s]}(t)$$

$$\widehat{p}_{33}(s,t) = \widehat{P}(T_3 > t | T_2 \le s, T_3 > s) = \widehat{S}_3^{]s[}(t)$$

$$\widehat{p}_{34}(s,t) = \widehat{P}(T_3 \le t | T_2 \le s, T_3 > s) = 1 - \widehat{S}_3^{]s[}(t)$$

Including covariates

One important goal is to estimate the transition probabilities given continuous covariate(s).

- Estimators based on a Cox's model fitted marginally to each type of transitions, with the corresponding baseline hazard function estimated by the Breslow's method.
- In the paper by Meira-Machado et al. (COST 2015) nonparametric regression estimators are introduced where local smoothing is done by introducing kernel weights that are based on Nadaraya-Watson regression.
- A single-index model is one effective tool to avoid the curse of dimensionality.

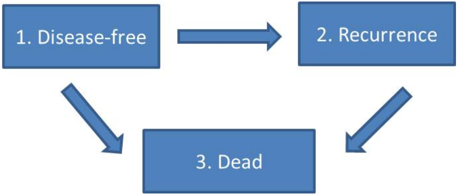
R-based packages

1. the msm package can be used to obtain estimates for the transition probabilities in time-homogeneous Markov models.
2. the etm package computes and displays the transition probabilities for the Aalen-Johansen estimator.
3. the mstate package computes and displays the transition probabilities for the landmark Aalen-Johansen estimator.
4. the msSurv package estimate the state occupation probabilities.

R-based packages

5. the p3state.msm package enables the user to perform inference in the illness-death model. The main feature of the package is its ability for obtaining non-Markov estimates for the transition probabilities.
6. the TPmsm package computes and displays the transition probabilities for several methods.
7. the TP.idm package computes and displays the transition probabilities for the landmark estimator and the Aalen-Johansen estimator.
8. the survidm package for inference and prediction in an illness-death model.

- Available as part of the R survival package.
- 929 patients underwent a curative surgery for colorectal cancer.
- 468 developed recurrence - 414 died; 38 died without recurrence.
- States: “Alive and Disease-Free”; “Recurrence”; “Death”.
- Covariates: Age (years)



Transition Probabilities

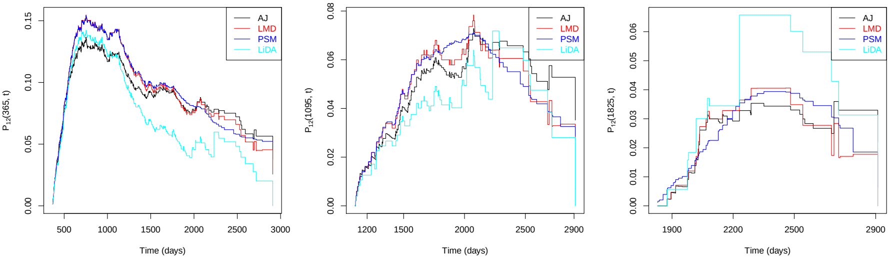


Figure: Estimates of the transition probabilities $p_{12}(s,t)$ for fixed values of s. Colon cancer data.

Transition Probabilities

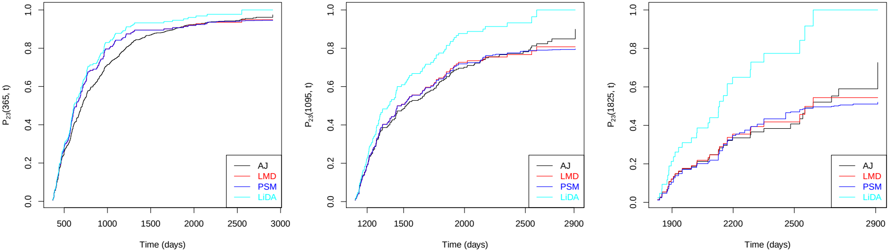


Figure: Estimates of the transition probabilities $p_{23}(s, t)$ for fixed values of s. Colon cancer data.

Testing the Markov Assumption

The Markov assumption is that future evolution only depends on the state occupied at current time.

The Markov assumption may be checked:

- by including covariates in the modelling process.
- log-rank test statistics for each of the relevant transition intensities can be combined to construct a local test of Markovianity.
- Rodriguez-Girondo and de Uña-Álvarez (2012) proposed a non-parametric test of Markovianity based upon the Kendall's τ between the time of exit from the healthy state and time of death.

Some 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.
- Pepe, M.S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* **86**, 770-778.
- 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.
- Amorim AP, de Uña-Álvarez J, Meira-Machado L (2011) Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters,* **81(7)**, 797-806.
- Van Keilegon I, de Uña-Álvarez J, Meira-Machado L. (2011) Nonparametric location-scale models for successive survival times under dependent censoring. *Journal of Statistical Planning and Inference,* **141**, 1118-1131.
- Meira-Machado L, and Roca-Pardiñas J (2011). p3state.msm: Analyzing survival data from an illness-death model. *Journal of Statistical Software,* **38(3)**, 1-18.
- Moreira A, de Uña-Álvarez J, Meira-Machado L (2013). Presmoothing the Aalen-Johansen estimator in the illness-death model. *Electronic Journal of Statistics,* **Volume 7**, 1491-1516.
- Meira-Machado L, Roca-Pardiñas R, Van Keilegon I, Cadarso-Suárez C (2013) Bandwidth selection for the estimation of transition probabilities in the location-scale progressive three-state model. *Computational Statistics,* **28**, 2185-2210.

Some references

- Allignol A, Beyersmann J, Gerds T and Latouch A (2014). A competing risks approach for non-parametric estimation of transition probabilities in a non-Markov illness-death model, *Lifetime Data Analysis*, **20**, 495–513.
- Meira-Machado L, de Uña-Álvarez J, Somnath D. (2015). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Computational Statistics*, *30(2)*, 377-397.
- Araújo A, Meira-Machado L, Roca-Pardiñas J (2015). TPmsm: Estimation of the transition probabilities in 3-state models. *Journal of Statistical Software*, **62(4)**
- Titman, A (2015) Transition probability estimates for non-Markov multi-state models. *Biometrics*, **11**, 1034-1041.
- de Uña-Álvarez J, Meira-Machado L. (2015). Nonparametric Estimation of Transition Probabilities in the Non-Markov Illness-Death Model: A Comparative Study, *Biometrics*, **71**, 364–375.
- Putter H and Spitoni C. (2016). Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator, *Statistical Methods in Medical Research*.

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 the Project *UID*/*MAT*/00013/2013.

Thank you for your attention!