Supporting Information for: Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring

Olivier Bouaziz

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France

1 Fast approximation of pseudo-observations for right-censored data using the infinitesimal jackknife

The classical jackknife procedure can be seen as giving a weight to each observation, then omitting an observation corresponds as giving it a weight of zero. This procedure can further be modified by giving the observation a weight slightly smaller than the other observations and then taking the limit as this weight tends to 0. This is the idea behind the infinitesimal jackknife as introduced by [1]. The weighted version of the Kaplan-Meier estimator is defined as:

$$\hat{S}_w(t) = \prod_{T_j \leq t} \left(1 - \frac{\sum_i w_i \triangle N_i(T_j)}{\sum_i w_i Y_i(T_j)}\right),$$

where $N_i(t) = I(T_i \leq t, \Delta_i = 1)$, $\triangle N_i(t) = N_i(t) - N_i(t-)$ and $Y_i(t) = I(T_i \geq t)$. From the results of [1], we have, as n tends to infinity that

$$n\hat{S}(t) - (n-1)\hat{S}^{(-l)}(t) = \hat{S}(t) + \hat{D}_l(t) + o_{\mathbb{P}}(1),$$

where

$$\hat{D}_l(t) = \frac{\partial}{\partial w_l} \hat{S}_w(t)\Big|_{w_j = 1/n,\, j=1,\ldots,n}.$$

In the following we will compute the derivative of $\hat{S}_w$ with respect to w_l, by separating the cases when the l^{th} observation is censored ($\Delta_l = 0$) and when it is observed ($\Delta_l = 1$).

(i) $\Delta_l = 0$
In that case, only the terms $\sum_i w_i Y_i(T_j)$ might contain w_l, since the term $\sum_i w_i \triangle N_i(T_j)$ is not 0 only for uncensored observations. We have:

$$\begin{aligned}\frac{\partial}{\partial w_l}\hat{S}_w(t)(1-\Delta_l) &= \sum_{T_j \neq T_l, T_j \leq t} Y_l(T_j) \frac{\sum_i w_i \triangle N_i(T_j)}{(\sum_i w_i Y_i(T_j))^2} \prod_{T_k \neq T_j, T_k \leq t} \left(1 - \frac{\sum_i w_i \triangle N_i(T_k)}{\sum_i w_i Y_i(T_k)}\right) \\ &\quad \times (1-\Delta_l) \\ &= \sum_{T_j \neq T_l, T_j \leq t} Y_l(T_j) \frac{\sum_i w_i \triangle N_i(T_j)}{\sum_i w_i Y_i(T_j)} \frac{\hat{S}_w(t)(1-\Delta_l)}{\sum_i w_i Y_i(T_j) - \sum_i w_i \triangle N_i(T_j)}, \quad (1)\end{aligned}$$

using the fact that

$$\prod_{T_k \neq T_j, T_k \leq t} \left(1 - \frac{\sum_i w_i \triangle N_i(T_k)}{\sum_i w_i Y_i(T_k)}\right) = \frac{\hat{S}_w(t) \sum_i w_i Y_i(T_j)}{\sum_i w_i Y_i(T_j) - \sum_i w_i \triangle N_i(T_j)}. \quad (2)$$

(ii) $\Delta_l = 1$

In that case, the term $\sum_i w_i \triangle N_i(T_j)$ might also contain w_l if the uncensored observation occurs before time t.

$$\begin{aligned}\frac{\partial}{\partial w_l}\hat{S}_w(t)\Delta_l = &\sum_{T_j \neq T_l, T_j \leq t} Y_l(T_j)\frac{\sum_i w_i \triangle N_i(T_j)}{\sum_i w_i Y_i(T_j)}\frac{\hat{S}_w(t)\Delta_l}{\sum_i w_i Y_i(T_j) - \sum_i w_i \triangle N_i(T_j)} \\ &+ I(T_l \leq t, \Delta_l = 1)\left(-\frac{1}{\sum_i w_i Y_i(T_l)} + \frac{\sum_i w_i \triangle N_i(T_l)}{(\sum_i w_i Y_i(T_l))^2}\right) \\ &\times \prod_{T_j \neq T_l, T_j \leq t}\left(1 - \frac{\sum_i w_i \triangle N_i(T_j)}{\sum_i w_i Y_i(T_j)}\right).\end{aligned}$$

Then, $I(T_l \leq t, \Delta_l = 1)$ is replaced by $N_l(t)$, the product term is expressed as in Equation (2) and we obtain:

$$\begin{aligned}\frac{\partial}{\partial w_l}\hat{S}_w(t)\Delta_l = &\sum_{T_j \neq T_l, T_j \leq t} Y_l(T_j)\frac{\sum_i w_i \triangle N_i(T_j)}{\sum_i w_i Y_i(T_j)}\frac{\hat{S}_w(t)\Delta_l}{\sum_i w_i Y_i(T_j) - \sum_i w_i \triangle N_i(T_j)} \\ &- N_l(t)\frac{\hat{S}_w(t)}{\sum_i w_i Y_i(T_l)}. \qquad (3)\end{aligned}$$

Now, gathering the terms in Equations (1), (3) and evaluating the derivatives at $w_l = 1/n$, we get:

$$\begin{aligned}\hat{D}_l(t) &= \hat{S}(t)\left[\sum_{T_j \neq T_l, T_j \leq t}\frac{Y_l(T_j)\triangle\bar{N}(T_j)}{\hat{H}(T_j)(\hat{H}(T_j) - \triangle\bar{N}(T_j))} - \frac{N_l(t)}{\hat{H}(T_l)}\right] \\ &= \hat{S}(t)\left[\sum_{T_j \leq t}\frac{Y_l(T_j)\triangle\bar{N}(T_j)}{\hat{H}(T_j)(\hat{H}(T_j) - \triangle\bar{N}(T_j))} - \frac{N_l(t)\triangle\bar{N}(T_l)}{\hat{H}(T_l)(\hat{H}(T_l) - \triangle\bar{N}(T_l))} - \frac{N_l(t)}{\hat{H}(T_l)}\right] \\ &= \hat{S}(t)\left[\sum_{T_j \leq t}\frac{Y_l(T_j)\triangle\bar{N}(T_j)}{\hat{H}(T_j)(\hat{H}(T_j) - \triangle\bar{N}(T_j))} - \frac{N_l(t)}{\hat{H}(T_l) - \triangle\bar{N}(T_l)}\right], \qquad (4)\end{aligned}$$

where $\hat{H}(u) = \sum_i Y_i(u)/n$ (using the notation in the main document), $\bar{N}(u) = \sum_i N_i(u)/n$. If we assume that only a finite number of ties can occur at a given time point, then the terms $\triangle\bar{N}(T_l)$ and $\bar{N}(T_j)$ in the two denominators will tend to 0 as n tends to infinity. Then, writing this result using the standard counting process we obtain:

$$\begin{aligned}n\hat{S}(t) - (n-1)\hat{S}^{(-l)}(t) &= \hat{S}(t) + \hat{S}(t)\left[\sum_{T_j \leq t}\frac{Y_l(T_j)\triangle\bar{N}(T_j)}{(\hat{H}(T_j))^2} - \frac{N_l(t)}{\hat{H}(T_l)}\right] + o_{\mathbb{P}}(1) \\ &= \hat{S}(t) + \hat{S}(t)\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{Y_l(u)dN_i(u)}{(\hat{H}(u))^2} - \int_0^t \frac{dN_l(u)}{\hat{H}(u)}\right] + o_{\mathbb{P}}(1) \\ &= \hat{S}(t) - \hat{S}(t)\int_0^t \frac{d\hat{M}_l(u)}{\hat{H}(u)} + o_{\mathbb{P}}(1),\end{aligned}$$

where $\hat{M}_l$ is the martingale residual, as defined in the main document. As it turns out, this is exactly the same approximation as the one proposed in the main document, in Proposition 1. Pseudo observations based on the infinitesimal jackknife are implemented in the `survival` library through the `pseudo` function. This function takes a `survfit` object as input which can

be computed in two different ways, depending on the value of the `stype` argument. The `pseudo` function returns the pseudo-observations obtained using Formula (4) when the `survfit` object has been implemented with `stype`= 1 (the default for the `survfit` function).

An alternative formula for the pseudo observations using the infinitesimal jackknife can also be obtained using the Breslow estimator of the survival function (the exponential of minus the Nelson-Aalen estimator). The weighted version of this estimator is defined as:

$$\tilde{S}_w(t) = \exp\left(-\int_0^t \frac{\sum_i w_i dN_i(u)}{\sum_i w_i Y_i(u)}\right).$$

Taking the derivative with respect to w_l gives

$$\frac{\partial}{\partial w_l}\tilde{S}_w(t) = \int_0^t \left(-\frac{dN_l(u)}{\sum_i w_i Y_i(u)} + \frac{Y_l(u)\sum_i w_i dN_i(u)}{(\sum_i w_i Y_i(u))^2}\right)\tilde{S}_w(t),$$

and evaluating this expression in $w_j = 1/n$ for $j = 1, \ldots, n$, leads to

$$\begin{aligned}\left.\frac{\partial}{\partial w_l}\tilde{S}_w(t)\right|_{w_j=1/n,\, j=1,\ldots,n} &= \left(-\int_0^t \frac{dN_l(u)}{\hat{H}(u)} + \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{Y_l(u)dN_i(u)}{(\hat{H}(u))^2}\right)\tilde{S}(t), \\ &= -\tilde{S}(t)\int_0^t \frac{d\hat{M}_l(u)}{\hat{H}(u)}. \end{aligned} \tag{5}$$

We therefore retrieve the same approximation as the one obtained using the Von-Mises expansion developed in the current paper in Proposition 1, but with the Breslow estimator instead of the Kaplan-Meier. Those two estimators are very similar (even for small sample sizes), in particular they are asymptotically equivalent. The `pseudo` function returns the pseudo-observations obtained using Formula (5) when the `survfit` object has been implemented with `stype`= 2.

The `pseudo` function also allows to compute pseudo-observations for the Restricted Mean Survival Time using the infinitesimal jackknife. This has been implemented by computing the integral of the pseudo-observation for the survival function using the simple trapezoid rule.

2 Supplementary simulations for interval censored when τ is equal to infinity

This section contains a supplementary simulation analysis for interval-censored data. In this scenario we assume a standard linear model for the time of interest:

$$T_i^* = \beta_0 + \beta_1 Z_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{6}$$

where $\beta_0 = 6$, $\beta_1 = 4$, $Z_i \sim \mathcal{U}[0, 2]$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. Here $\tau = \infty$, and for the interval-censored data the values of L_i and R_i were determined through a visit process with a total of $K = 5$ simulated visits such that $V_1 \sim \mathcal{U}[0, 10]$ and $V_k = V_{k-1} + U[0, 4]$, for $k = 2, \ldots, K$. The left, interval and right-censored observations where obtained as in Section 5.2 of the main document. This simulation setting corresponds to 10% of left-censoring, 26% of right-censoring and 64% of interval-censoring. For interval-censored data, the average length of the intervals was approximately equal to 3.5. The results are presented in Table 1 where the pch estimator was used with cuts equal to $6, 8, 10, 12, 14$.

This scenario is challenging both due to the fact that τ equals infinity (and thus causing estimation problems in the tails) and to the width of the intervals that are larger on average than in Section 5.2 of the main document. As a result, the algorithm seems to fail in some rare cases for $n = 500$ and generates a drastic overestimation of the parameter value. This seemed to be caused by the generation of samples for which too few values of L_i and R_i satisfy the

regularity conditions of Equation (5) of the main document. In the simulations for $n = 500$, this situation occurred for one sample for both the jackknife method and the approximated formula and for 3 other samples for the jackknife method. Table 1 only displays the results with those 4 samples generating drastic overestimations removed, which shows very similar performances of the two methods. All samples were kept for $n = 1,000$ and the results are identical for both methods. We also compared the absolute difference between the two estimators of β_0 componentwise: 99% of those values are less than 2.039×10^{-2} for the first component and less than 2.760×10^{-2} for the second component, when $n = 500$, while 99% of those values are less than 5.521×10^{-3} for the first component and less than 7.453×10^{-3} for the second component when $n = 1,000$. In terms of computation time, the approximated formula is 258 and 426 times faster than the jackknife method for $n = 500$ and $n = 1,000$ respectively. Again, those results emphasise the importance of verifying the regularity conditions in Equation (5) of the main document for the pch model in choosing the number and location of the cuts.

	Jackknife				Approximated formula			
n	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
500	-0.116	0.168	0.042	4.086 min	-0.105	0.203	0.052	0.949 s
	0.083	0.150	0.029		0.078	0.151	0.029	
1,000	-0.108	0.119	0.026	11.216 min	-0.105	0.119	0.025	1.580 s
	0.077	0.106	0.017		0.075	0.106	0.017	

Table 1: Simulation results for the estimation of β in the RMST model (6) based on pseudo-regression with 10% of left-censored data, 26% of interval-censored data and 64% of right-censored data. The piecewise constant hasard model with cuts equal to $6, 8, 10, 12, 14$ was used for the estimation of the survival function in the computation of the pseudo-observations. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates.

References

[1] Louis A Jaeckel. *The infinitesimal jackknife.* Bell Telephone Laboratories, 1972.