Non-parametric estimation of net survival under dependence assumptions

JS2O 2025 @ Perpignan

Oskar Laverny[1]

April 3, 2025

[1] Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.
Joint work with: R. Alhajal, N. Grafféo & R. giorgi

Table of content

Relative survival analysis ?

Relative survival context

Relative Survival Context: In population-based studies and/or cancer registries, the specific cause of death is often unidentified, unreliable or even unavailable.

Random Variable	Name	Observed ?
E	"Excess" lifetime	No
P	"Population" lifetime	No, but known distribution.
$O = E \wedge P$	"Overall" lifetime	No
C	"Censoring" time	No
$\boldsymbol{X}$	Vector of covariates	Yes
$T = O \wedge C$	Event time	Yes
$\Delta = \mathbb{1}\{T \leq C\}$	Event status	Yes
$\mathbb{1}\{E \leq P\}$	Cause of death	No

Goal: Estimate the distribution of E, say by it's hazard $\lambda_E(t) = \partial\Lambda_E(t) = -\partial \ln S_E(t)$.

Remark: With the missing cause of death indicatrix, we cannot use directly competing risks analysis..

Assumptions (Standard assumptions[1])

(i) $C \perp\!\!\!\perp (E, P, \boldsymbol{X})$

(ii) $\mathcal{L}(P \mid \boldsymbol{X})$ is known from life tables.

Assumptions (Dependence structure of (E, P))

The $(\mathcal{H}_{\mathcal{C}})$ hypothesis states that all couples (E_i, P_i) have the same survival copula $\mathcal{C}$:

$$(\mathcal{H}_{\mathcal{C}}) : \forall i \in 1, ..., n,\ S_{O_i}(t) = \mathcal{C}\left(S_E(t), S_{P_i}(t)\right) \tag{1}$$

Example: Denoting Π the independence copula, $(\mathcal{H}_{\Pi}) \iff \forall i\ E_i \perp\!\!\!\perp P_i$ was assumed in previous literature.

Issue: It would be reasonable to assume that $\mathcal{C} \neq \Pi$.. But remark that $\mathcal{C}$ is not identifiable !

[1] Maja Pohar Perme, Janez Stare, and Jacques Estève. "On Estimation in Relative Survival". In: *Biometrics* 68.1 (Mar. 2012), pp. 113–120. ISSN: 0006-341X, 1541-0420. DOI: 10.1111/j.1541-0420.2011.01640.x. (Visited on 11/05/2023).

Observations: Let $(\boldsymbol{X}_i, T_i, \Delta_i)_{i=1,\dots,n}$ be an observed, i.i.d., n-sample.

Filtered probability space: $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ with $\mathcal{F}_t = \sigma\{X_i, (T_i, \Delta_i) : T_i \leq t,\ \forall i \in 1, .., n\}$.

As standard in survival analysis[2], we define the following stochastic processes:

$$
\begin{aligned}
N(t) &= \mathbb{1}\{O \leq t, O \leq C\} && \textit{(Uncensored deaths process)} \\
Y(t) &= \mathbb{1}\{O \geq t, C \geq t\} && \textit{(At-risk process)} \\
M(t) &= N(t) - \int_0^t Y(s)\partial\Lambda_O(s) && \textit{(Martingale)} \\
N_E(t) &= \mathbb{1}\{E \leq t, E \leq C\} && \textit{(Excess uncensored deaths process)} \\
Y_E(t) &= \mathbb{1}\{E \geq t, C \geq t\} && \textit{(Excess at-risk process)}
\end{aligned}
$$

We similarly defined individual versions N_i, Y_i, M_i, N_{E_i} and Y_{E_i}.
Issue: N_{E_i} and Y_{E_i} are not observable.

[2] Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. New York, NY: Springer US, 1993. ISBN: 978-0-387-94519-4 978-1-4612-4348-9. DOI: 10.1007/978-1-4612-4348-9. (Visited on 02/22/2024).

Estimation of the excess hazard

Link between $(N_E, Y_E, \partial\Lambda_E)$ and $(N, Y, \partial\Lambda_O)$

Let $a(t) = \mathbb{P}(P \geq t | E = t)$, $b(t) = \mathbb{P}(P = t | E \geq t)$ and $c(t) = \mathbb{P}(P \geq t | E \geq t)$.

Lemma (Expressions of N_E, Y_E, Λ_E, Doob-meyer decomposition of N_E.)

Integrating out P, we have:

$$\partial N_E(t) = \mathbb{E}\left(\frac{\partial N(t)}{a(t)} - \frac{b(t)Y(t)}{a(t)c(t)} \,\middle|\, E, C\right)$$
$$Y_E(t) = \mathbb{E}\left(\frac{Y(t)}{c(t)} \,\middle|\, E, C\right)$$
$$\partial M_E(t) = \mathbb{E}\left(\frac{\partial M(t)}{a(t)} \,\middle|\, E, C\right)$$
$$\partial\Lambda_E(t) = \frac{c(t)}{a(t)}\left(\partial\Lambda_O(t) - \frac{b(t)}{c(t)}\right).$$

Furthermore, the process N_E admits the following Doob-Meyer decomposition:

$$\partial N_E(t) = \partial M_E(t) + Y_E(t)\partial\Lambda_E(t),$$

Warning: These conditional expectations (and thus N_E, Y_E) are still not observable!

We drop the previous conditional expectations to obtain:

$$\partial\widetilde{N}_{E,i}(t) = \frac{\partial N_i(t)}{a_i(t)} - \frac{b_i(t)}{a_i(t)c_i(t)} Y_i(t)$$
$$\widetilde{Y}_{E,i}(t) = \frac{Y_i(t)}{c_i(t)}$$
$$\partial\widetilde{M}_{E,i}(t) = \frac{\partial M_i(t)}{a_i(t)}$$
$$\partial\widetilde{\Lambda}_E(t) = \frac{\sum_{i=1}^n \partial\widetilde{N}_{E_i}(t)}{\sum_{i=1}^n \widetilde{Y}_{E_i}(t)}.$$

However, note that the constants can be expressed as follow:

$$a_i(t) = \mathcal{C}_1\left(S_E(t), S_{P_i}(t)\right)$$
$$b_i(t) = \mathcal{C}_2\left(S_E(t), S_{P_i}(t)\right) \frac{-\partial S_{P_i}(t)}{S_E(t)}$$
$$c_i(t) = \mathcal{C}(S_E(t), S_{P_i}(t)) \frac{1}{S_E(t)},$$

Problem: $\widetilde{\Lambda}_E(t)$ is still not observable since it depends on unknow S_E.

Exception: Uner $(\mathcal{H}_\Pi)$, $\widetilde{\Lambda}_E(t)$ is observable !

Definition (Generalized PPE)

We call *generalized Pohar Perme estimator* the solution $\widehat{\Lambda}_E$ of the differential equation

$$\partial\widehat{\Lambda}_E(t) = \frac{\sum_{i=1}^n \partial\widehat{N}_{E,i}(t)}{\sum_{i=1}^n \widehat{Y}_{E,i}(t)}, \text{ where:} \tag{2}$$

$$\widehat{N}_{E,i}(t) = \frac{\partial N_i(t)}{\widehat{a}_i(t)} - \frac{\widehat{b}_i(t)Y_i(t)}{\widehat{a}_i(t)\widehat{c}_i(t)}, \qquad \widehat{a}_i(t) = \mathcal{C}_1\left(\widehat{S}_E(t), S_{P_i}(t)\right),$$

$$\widehat{Y}_{E,i}(t) = \frac{Y_i(t)}{\widehat{c}_i(t)}, \qquad \widehat{b}_i(t) = \mathcal{C}_2\left(\widehat{S}_E(t), S_{P_i}(t)\right)\frac{-\partial S_{P_i}(t)}{\widehat{S}_E(t)},$$

$$\widehat{S}_E(t) = \exp\left\{-\widehat{\Lambda}_E(t)\right\}, \qquad \widehat{c}_i(t) = \frac{\mathcal{C}\left(\widehat{S}_E(t), S_{P_i}(t)\right)}{\widehat{S}_E(t)}.$$

Remark: Under $(\mathcal{H}_\Pi)$, $\mathcal{C}(u,v) = uv, \mathcal{C}_1(u,v) = v$ and $\mathcal{C}_2(u,v) = u$, and the differential equation is separable. It is called the Pohar Perme estimator, consistent and asymptotically unbiased estimator of the excess hazard.

Second order

Lemma (Doob-Meyer decompositions)

The process $\widetilde{\Lambda}_E$ admits the following Doob-Meyer decomposition:

$$\widetilde{\Lambda}_E(t) = \Lambda_E(t) + \Xi(t),$$

where the local square integrable martingale Ξ is defined by:

$$\partial\Xi(t) = \frac{\sum_{i=1}^{n} \frac{1}{a_i(t)} \partial M_i(t)}{\sum_{i=1}^{n} \frac{Y_i(t)}{c_i(t)}}.$$

This is derived from the DM decomposition of N_i's.

Standard techniques using optional processes.

Property (Variance of $\widetilde{\Lambda}_E(t)$)

$$\mathrm{Var}\left(\widetilde{\Lambda}_E(t)\right) = \mathbb{E}\left([\Xi](t)\right) = \mathbb{E}\left(\int_0^t \frac{\sum_{i=1}^n \frac{1}{a_i(t)^2}\partial N_i(t)}{\left(\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}\right)^2}\right)$$

Thus, a good estimator for the variance of $\widetilde{\Lambda}_E(t)$ is simply $[\Xi](t)$.

Definition (Estimator of $\widetilde{\Lambda}_E(t)$'s variance)

$$\widetilde{\sigma}^2_E(t) = [\Xi](t) = \int_0^t \frac{\sum_{i=1}^n \frac{1}{a_i(t)^2}\partial N_i(t)}{\left(\sum_{i=1}^n \frac{Y_i(t)}{c_i(t)}\right)^2} \quad \text{and} \quad \widehat{\sigma}^2_E(t) = \int_0^t \frac{\sum_{i=1}^n \frac{1}{\widehat{a}_i(t)^2}\partial N_i(t)}{\left(\sum_{i=1}^n \frac{1}{\widehat{c}_i(t)}Y_i(t)\right)^2}$$

Under $(\mathcal{H}_\Pi)$, $\widetilde{\sigma}^2_E(t)$ is feasible, already obtained in previous litterature. However, under $(\mathcal{H}_\mathcal{C})$, $\widetilde{\sigma^2_E}(t)$ is not feasible, and thus we propose to use the straightforward plug-in estimator $\widehat{\sigma}^2_E(t)$.

Assymptotics & tests

Groups and test statistic

Let $G = \{g_1, .., g_r\}$ be a partition of $1, ..., n$. We want to check the hypothesis:

$$(H_0) : \forall g \in G, \forall i \in g,\ \Lambda_{E_i} = \Lambda_E.$$

Let us denote $\widetilde{Y}_{E,g} = \sum_{i \in g} \widetilde{Y}_{E,i}$ for any group $g \in G$, and $\widetilde{Y}_{E,\bullet} = \sum_{g \in G} \widetilde{Y}_{E,g}$. Similarly, denote $\widetilde{N}_{E,g} = \sum_{i \in g} \widetilde{N}_{E,i}$ and $\widetilde{N}_{E,\bullet} = \sum_{g \in G} \widetilde{N}_{E,g}$.

Define finally the vectors $\boldsymbol{R}(t), \boldsymbol{Z}(t)$, the matrix $\boldsymbol{\Gamma}(t)$ and the test statistic $\widetilde{\chi}(T)$ by:

$$\begin{aligned}
R_g(t) &= \frac{\widetilde{Y}_{E,g}(t)}{\widetilde{Y}_{E,\bullet}(t)} \\
Z_g(t) &= \widetilde{N}_{E,g}(t) - \int_0^t R_g(s) \partial \widetilde{N}_{E,\bullet}(s) \\
\Gamma_{g,h}(t) &= \sum_{\ell \in G} \int_0^t (\delta_{\ell,g} - R_g(s)) (\delta_{\ell,h} - R_h(s)) \sum_{i \in \ell} \frac{\partial N_i(s)}{a_i(s)^2}. \\
\widetilde{\chi}(T) &= \boldsymbol{Z}(T)' \boldsymbol{\Gamma}(T)^{-1} \boldsymbol{Z}(T)
\end{aligned}$$

Property

Under (H_0), assuming the existence of an $\epsilon > 0$ such that $a_i(t) > \epsilon$ and $c_i(t) > \epsilon$ over $t \in [0, T]$, we have

$$\widetilde{\chi}(T) \xrightarrow[n\to\infty]{\mathcal{D}} \texttt{Chi2}\left(|G| - 1\right).$$

Lemma (Elements of proofs, using Robolledo's Martingale CLT)

Let $T < \infty$. Under (H_0), assuming that there exists an $\epsilon > 0 :\ a_i(t) > \epsilon, c_i(t) > \epsilon$ over $t \in [0, T]$, the following points hold over $t \in [0, T]$,

(i) $\boldsymbol{Z}$ *is a centered local square integrable martingale*

(ii) $\mathrm{Cov}(\boldsymbol{Z}(t)) = \mathbb{E}\left(\boldsymbol{\Gamma}(t)\right)$

(iii) $n^{-1}\boldsymbol{\Gamma}(t) \xrightarrow[n\to\infty]{\mathbb{P}} \boldsymbol{V}(t)$, $\boldsymbol{V}$ *is deterministic, and both* $\boldsymbol{\Gamma}(t)$ *and* $\boldsymbol{V}(t)$ *are semi-definite positives.*

(iv) $n^{-\frac{1}{2}}\boldsymbol{Z}(t) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \boldsymbol{V}(t)\right)$

(v) $\mathrm{Ker}(\boldsymbol{V}(t)) = \mathrm{Vect}(1)$

Short example

Dataset: Colorectal cancer

The dataset we have consists of french patients with colorectal cancer, well described in Wolski & Al[3]. See also this page of `NetSurvival.jl`'s documentation.

Characteristics of the dataset:

- 10 years of follow-up before administrative censoring
- Demographic covariates $\boldsymbol{X}$: age, sex, date of birth, enough to fetch P_i's distributions.
- Extra covariates: the primary tumor location, left or right.

Main question on this data: Does the tumor location affect significantly the net survival ?

State of the art: Previous literature, restricted to $(\mathcal{H}_\Pi)$, conclude that it does not. But $(\mathcal{H}_\Pi)$ is known to be false..

[3] Anna Wolski, Nathalie Grafféo, Roch Giorgi, and the CENSUR working survival group. "A Permutation Test Based on the Restricted Mean Survival Time for Comparison of Net Survival Distributions in Non-Proportional Excess Hazard Settings". In: *Statistical Methods in Medical Research* 29.6 (June 2020), pp. 1612–1623. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280219870217. (Visited on 12/13/2023).

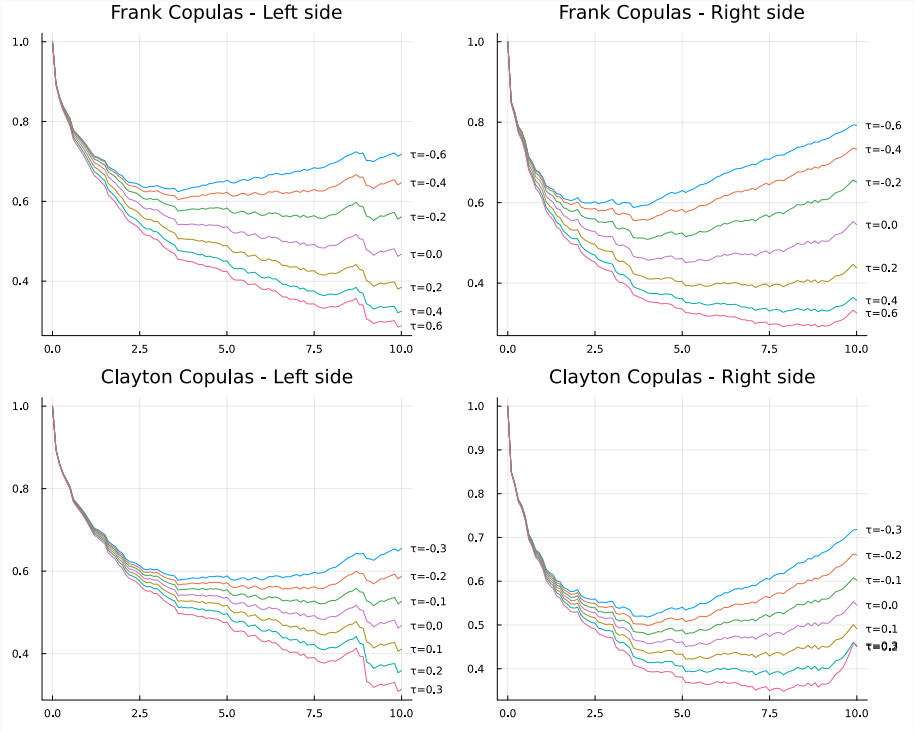

Figure 1: $\widehat{S}_E$ for several $(\mathcal{H}_\mathcal{C})$. Data was split w.r.t. tumor location (left or right), and several copulas $\mathcal{C}$ are proposed: Frank copulas (top), Clayton copulas (bottom), with varying Kendall τ. In each graph, $\tau = 0 \iff \mathcal{C} = \Pi$

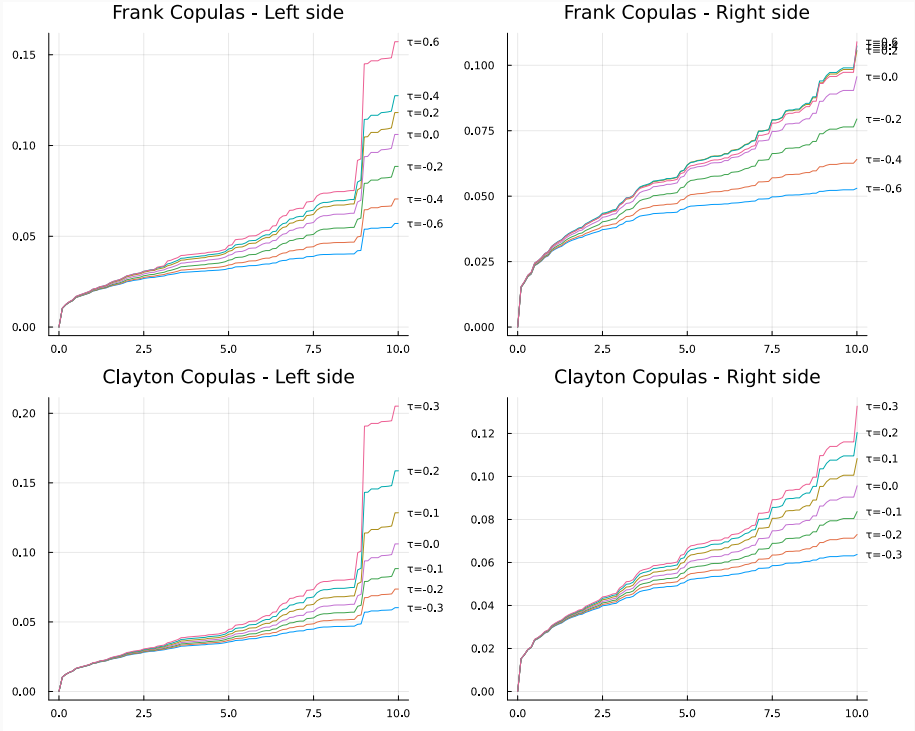

Figure 2: Estimated standard errors $\sqrt{\widehat{\sigma}_E^2(t)}$. Again, for both the frank and Clayton copula, $\tau = 0$ represents the Pohar Perme-estimated variance. Multiply by ≈ 4 to get wideness of assymptotic CIs

Tests results for several Frank copulas.

Table 1: Obtained p-value for the generalized log-rank-type test for $\mathcal{C} = \texttt{Frank}(\tau)$, at various horizons T (in years).

τ	$T = 3$		$T = 5$		$T = 8$		$T = 10$	
-0.6	0.05266	+	0.20128		0.90222		0.66067	
-0.5	0.03689	*	0.13102		0.77497		0.75530	
-0.4	0.02417	*	0.07991	+	0.64116		0.85883	
-0.3	0.01476	*	0.04493	*	0.49968		0.98195	
-0.2	0.00845	**	0.02329	*	0.35883		0.86804	
-0.1	0.00461	**	0.01127	*	0.23305		0.69194	
0.0	0.00244	**	0.00522	**	0.13575		0.50419	
0.1	0.00129	**	0.00240	**	0.07163	+	0.33148	
0.2	0.00070	***	0.00114	**	0.03537	*	0.19859	
0.3	0.00040	***	0.00058	***	0.01724	*	0.11324	
0.4	0.00025	***	0.00034	***	0.00889	**	0.06671	+
0.5	0.00018	***	0.00023	***	0.00533	**	0.04642	*
0.6	0.00015	***	0.00021	***	0.00435	**	0.04985	*

Remarks on these results

(i) The variance explodes after $T = 8$, so maybe the $T = 10$ case is irrelevant.

(ii) We enforced the same copula on both left and right side...

(iii) Experts think that the true dependence structures should be concordant ($\tau > 0$) in this dataset.

(iv) Same kind of results with Claytons and Gumbels.

Conclusion and perspectives

Conclusion

So far:

(i) The relative survival field usually assumes $(\mathcal{H}_\Pi) : E \perp\!\!\!\perp P$, which is known to be false.

(ii) The true dependence structure is not estimable from available data.

(iii) However, even small dependencies ($\tau = 0.2$ or 0.3) can have large impact on results of estimators and tests, and thus on public health decisions.

(iv) Removing the assumption would in many case yield a confidence interval as wide as the unit interval for the survival function...

For all these reasons, we recommend that further analysis is made to craft acceptable dependence structures for these datasets.

Shameless propaganda:

(i) Full paper available on arXiv, full code merged in `JuliaSurv/NetSurvival.jl`.

(ii) NetPlus & LostLife projects on $L_1, ..., L_n$ i.i.d such that $O_i = P_i - L_i$.

(iii) The `JuliaSurv` community awaits you :)

Thanks !