Non-parametric estimation of net survival under dependence assumptions.

ISCB 2024 @ Thessaloniki

Oskar Laverny[1], Nathalie Grafféo[1], and Roch Giorgi[1,2]

July 17, 2023

[1] Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.

[2] Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Hop Timone, BioSTIC, Biostatistique et Technologies de l'Information et de la Communication, Marseille, France.

Table of content

Introduction to 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 $\partial\Lambda_E(t) = -\partial \ln S_E(t)$.

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

(In)Dependence assumptions

Remark: The joint distribution of $(E, P, C, \boldsymbol{X})$ characterizes our observations.

Assumptions (Standard assumptions[1])

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

$E \perp\!\!\!\perp \boldsymbol{X}$

The distribution of $P|\boldsymbol{X}$ is known from standard life tables (at time 0).

Assumptions (Dependence structure of (E, P))

The $(\mathcal{H}_\mathcal{C})$ hypothesis states that (E, P) has survival copula the bivariate copula $\mathcal{C}$:

$$(\mathcal{H}_\mathcal{C}):\ S_O(t) = \mathcal{C}\left(S_E(t), S_P(t)\right) \tag{1}$$

Example: Denoting Π the independence copula, $(\mathcal{H}_\Pi) \iff E \perp\!\!\!\perp P$ 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).

Dataset, filtration, stochastic processes

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 $\mathcal{F}_t = \sigma\left\{\mathrm{X}_i, (T_i, \Delta_i) : T_i \geq t,\ \forall i \in 1, .., n\right\}$.

As standard in survival analysis[2,3], 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] Thomas R Fleming and David P Harrington. *Counting Processes and Survival Analysis*. Vol. 625. John Wiley & Sons, 2013.

[3] 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).

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

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

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

Integrating out P, we have:

$$\begin{aligned}
\partial N_E(t) &= \frac{1}{a(t)}\mathbb{E}\left(\partial N(t)|E, C\right) - \frac{b(t)}{a(t)c(t)}\mathbb{E}\left(Y(t)|E, C\right) \\
Y_E(t) &= \frac{1}{c(t)}\mathbb{E}\left(Y(t)|E, C\right) \\
\partial M_E(t) &= \frac{1}{a(t)}\mathbb{E}\left(\partial M(t)|E, C\right) \\
\partial\Lambda_E(t) &= \frac{c(t)}{a(t)}\left(\partial\Lambda_O(t) - \frac{b(t)}{c(t)}\right).
\end{aligned}$$

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!

Estimation of the excess hazard

We drop the previous conditional expectations to obtain:

$$\partial\widetilde{N}_{E,i}(t) = \frac{\partial N_i(t)}{a_i(t)} - \frac{b_i(t)}{a_i(t)c_i(t)}Y_i(t)$$
$$\widetilde{Y}_{E,i}(t) = \frac{Y_i(t)}{c_i(t)}$$
$$\partial\widetilde{M}_{E,i}(t) = \frac{\partial M_i(t)}{a_i(t)}$$
$$\partial\widetilde{\Lambda}_E(t) = \frac{\sum_{i=1}^n \partial\widetilde{N}_{E_i}(t)}{\sum_{i=1}^n \widetilde{Y}_{E_i}(t)}. \tag{2}$$

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* a solution of the differential equation

$$\partial\widehat{\Lambda}_E(t) = \frac{\sum_{i=1}^n \partial\widehat{N}_{E,i}(t)}{\sum_{i=1}^n \widehat{Y}_{E,i}(t)} = \frac{\sum_{i=1}^n \frac{1}{\widehat{a}_i(t)}\partial N_i(t) - \frac{\widehat{b}_i(t)}{\widehat{a}_i(t)\widehat{c}_i(t)} Y_i(t)}{\sum_{i=1}^n \frac{1}{\widehat{c}_i(t)} Y_i(t)}, \tag{3}$$

where for all $i \in 1, ..., n$,

$$\widehat{a}_i(t) = \mathcal{C}_1\left(e^{-\widehat{\Lambda}_E(t)}, S_{P_i}(t)\right),$$

$$\widehat{b}_i(t) = \mathcal{C}_2\left(e^{-\widehat{\Lambda}_E(t)}, S_{P_i}(t)\right)\left(-\partial S_{P_i}(t)\right) e^{\widehat{\Lambda}_E(t)},$$

$$\widehat{c}_i(t) = \mathcal{C}\left(e^{-\widehat{\Lambda}_E(t)}, S_{P_i}(t)\right) e^{\widehat{\Lambda}_E(t)}.$$

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.

Variance estimation

Excess Doob-Meyer decomposition.

Lemma (Doob-Meyer decompositions)

(i) The process $\widetilde{N}_{E,i}$ *admits the following Doob-Meyer decomposition:*

$$\partial \widetilde{N}_{E,i}(t) = \partial \widetilde{M}_{E,i}(t) + \widetilde{Y}_{E,i}(t) \partial \Lambda_E(t),$$

where $\partial \Lambda_E(t)$ *is the* true *excess hazard.*

(ii) 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:*

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

(ii) is derived from (i) which is derived from the DM decomposition of N_i*'s.*

Variance estimation

Standard techniques using optional processes.

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

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

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

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

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

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

Log rank test and asymptotics

Let $G = \{g_1, .., g_r\}$ be a partition of $1, ..., n$. For any symbol $X \in \left\{\Lambda_E, \widetilde{N}_E, \widetilde{M}_E, \widetilde{Y}_E, \widetilde{\Lambda}_E\right\}$, denote first $X_\bullet = \sum_{i=1}^n X_i$ and then for any group $g \in G$, denote $X_g = \sum_{i \in g} X_i$.

We want to check the hypothesis:

$$(H_0) : \forall g \in G,\ \Lambda_{E,g} = \Lambda_{E,\bullet}.$$

For any group $g, h \in G$, define

$$R_g(t) = \frac{\widetilde{Y}_{E,g}(t)}{\widetilde{Y}_{E,\bullet}(t)}$$

$$\partial Z_g(t) = \partial \widetilde{N}_{E,g}(t) - R_g(t)\partial \widetilde{N}_{E,\bullet}(t)$$

$$\partial \Gamma_{g,h}(t) = \sum_{\ell \in G} \left(\delta_{\ell,g} - R_g(t)\right)\left(\delta_{\ell,h} - R_h(t)\right) \sum_{i \in \ell} \frac{\partial N_i(t)}{a_i(t)^2}$$

Property (Expectation and variance of Z)

Under (H_0), the multivariate process $\boldsymbol{Z} = (Z_g, g \in G)$ is centered, with variance-covariance matrix defined by

$$\mathrm{Cov}(Z_g(t), Z_h(t)) = \mathbb{E}\left(\Gamma_{g,h}(t)\right).$$

Property

As $n \to \infty$,

$$\frac{\boldsymbol{Z}(t)}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \boldsymbol{\Sigma}(t)\right),$$

where $\boldsymbol{\Sigma}(t)$ is a square matrix with entries

$$\Sigma_{g,h}(t) = \sum_{\ell \in G} \left(\delta_{g,\ell} - \omega_g\right)\left(\delta_{\ell,h} - \omega_h\right)\sigma_\ell^2$$

where $\omega_g = \lim_{n\to\infty} \frac{|g|}{n}$ and $\sigma_g^2 = \frac{1}{n}\sum_{i\in g} \mathbb{E}\left(a_{i,T_i}^{-2}\Delta_i\right)$.

Result: $\boldsymbol{Z}'\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}$ follows assymptiotically a $\chi^2(|G|-1)$ under (H_0).

Short example

The dataset we have consists of french patients with colorectal cancer, followed for up to 10 years, well described in Wolski & Al[4].

Demographic covariates $\boldsymbol{X}$: age, sex, date of birth.
Extra covariate: Tumor location, left or right.

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

With known routines under $(\mathcal{H}_\Pi)$, they conclude that yes it does. But $(\mathcal{H}_\Pi)$ is known to be false..

[4] 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).

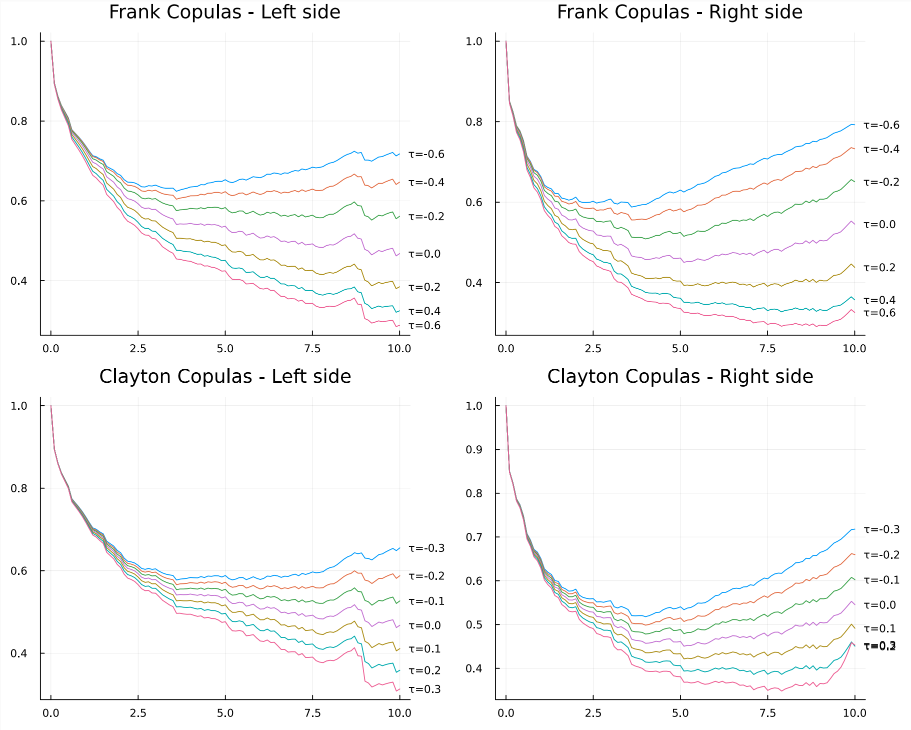

Figure 1: $\widehat{S}_E$ for several $(\mathcal{H}_{\mathcal{C}})$. Data was split w.r.t. tumor location (left or right), and several copulas $\mathcal{C}$ are proposed: Frank copulas (top), Clayton copulas (bottom), with varying Kendall τ. In each graph, $\tau = 0 \iff \mathcal{C} = \Pi$

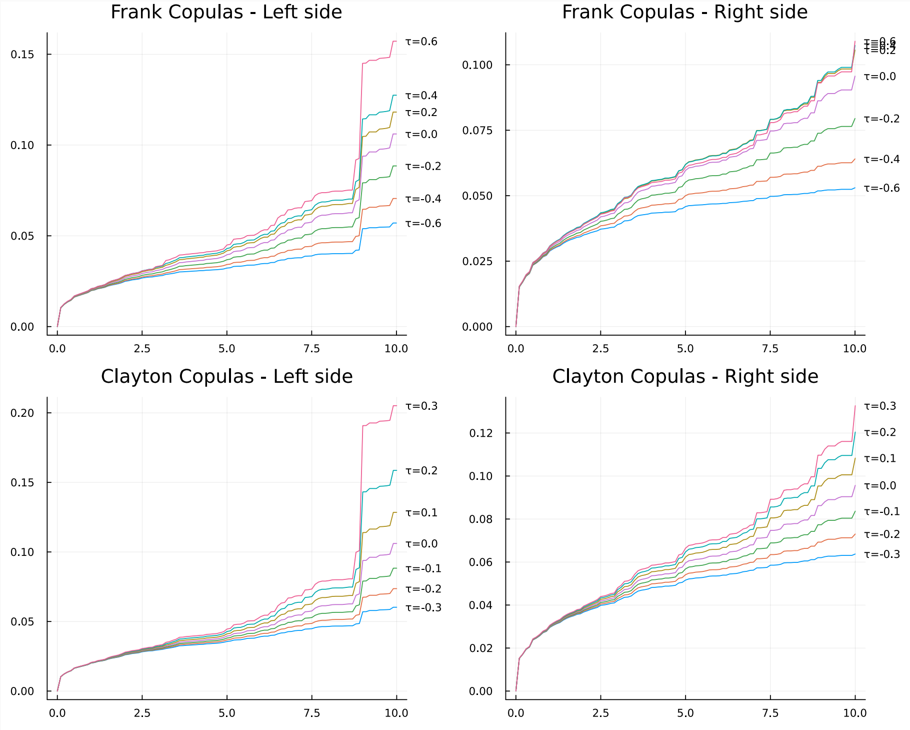

Figure 2: Estimated standard errors $\sqrt{\widehat{\sigma}_E^2(t)}$. Again, for both the frank and Clayton copula, $\tau = 0$ represents the Pohar Perme-estimated variance. Multiply by ≈ 4 to get wideness of assymptotic CIs

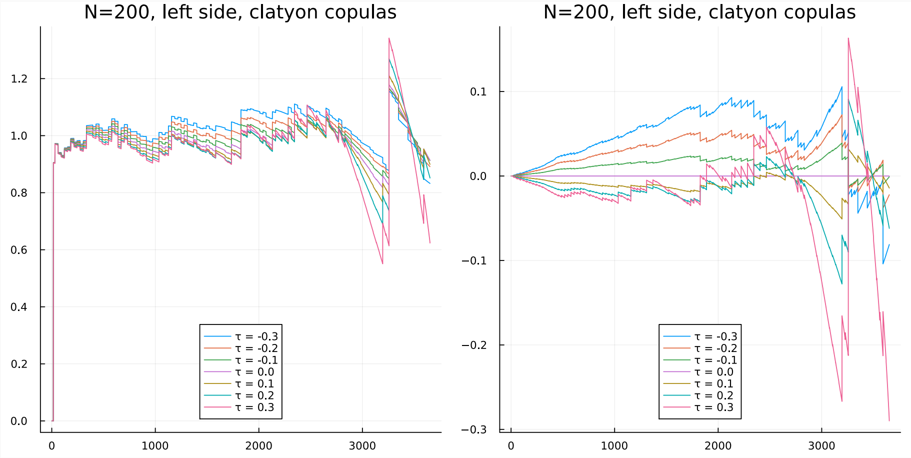

Figure 3: Left: Ratio of $\sqrt{\widehat{\sigma}_E^2(t)}$ and a bootstrap estimate (on N=200 resamples). Right: same ratios, recentered on the $\tau = 0$ curve (the Pohar Perme variance estimate), since this one does not suffer the plug-in biais.

Conclusion

Conclusion

So far:

(i) Net survival estimation 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 recommand that further analysis is made to craft acceptable dependence structures for these datasets.

Shameless propaganda:

(i) Our code will soon land in the Julia package `JuliaSurv/NetSurvival.jl`.

(ii) The `JuliaSurv` GitHub organization is rising, contributions welcomed !

(iii) Currently hiring on related topics... Please reach me !

Thanks !