Non-parametric estimation of net survival under dependence between death causes

Oskar Laverny [1] & Nathalie Graffféo [1] & Roch Giorgi [2]

[1] *Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.*
oskar.laverny@univ-amu.fr
[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.*

Résumé. L'analyse de survie relative traite un problème de risques compétitifs où la cause de la mort est inconnue. Ce manque d'information arrive régulièrement dans les études de cohortes liées au cancer. L'estimation non paramétrique de la survie nette est possible via l'estimateur de Pohar Perme, qui prend en compte les autres causes de mortalité. Dérivé de manière analogue à Kaplan-Meier, cet estimateur repose cependant sur une hypothèse d'indépendance non testable. Nous proposons ici de relâcher cette hypothèse et fournissons un estimateur généralisé qui fonctionne sous d'autres structures de dépendances, en exploitant les processus de comptage et les martingales sous-jacentes. Notre approche fournit de nouvelles perspectives sur l'estimateur de Pohar Perme et l'acceptabilité de ses hypothèses. Nous montrons la différence entre les deux estimateurs sur des données relatives au cancer colorectal et discutons finalement de possibles extensions de la méthodologie.

Mots-clés. Analyse de survie, Surive nette, Estimateurs non-paramétriques, Copules.

Abstract. Relative survival analysis deals with a competing risks survival model where the cause of death is unknown. This lack of information occurs regularly in population-based cancer studies. Non-parametric estimation of the net survival is possible through the Pohar Perme estimator, taking other causes of mortality into account. Derived similarly to Kaplan-Meier, it nevertheless relies on untestable independence assumptions. We propose here to relax these assumptions and provide a generalized estimator that works for other dependence structures, by leveraging the underlying counting process and martingales. Our approach provides a new perspective on the Pohar Perme estimator and the acceptability of this assumption. We showcase the difference between the two estimators on population-based colorectal cancer registry, and discuss potential extensions of the methodology.

Keywords. Survival analysis; Net survival; Non-parametric estimators; Copulas

1 Net survival analysis

Survival analysis produces valuable tools for prognosis of cancer patients. However, in population-based cancer studies, the cause of death – assumed binary, studied cancer or

not – is usually unreliable or unavailable. Relative survival analysis takes this particularity into account to evaluate the excess mortality – due to cancer – with respect to population life tables.

Let E, P and $O = E \wedge P$ be random times to death from (resp) the Excess, Population and Overall mortalities. Let $\mathbf{X}$ be a vector of covariates, C the time to censorship, and denote $T = O \wedge C$ and $\Delta = \mathbf{1}\{T \leq C\}$. Only $(\mathbf{X}, T, \Delta)$ is observable. In particular, we do not observe a potential ordering indicatrix $\mathbf{1}\{E \geq P\}$. Alike standard approaches, we suppose the distribution of $P|\mathbf{X}$ known from population life tables. To simplify our exposition, assume that the rest does not depend on covariates.

Let $(\mathbf{X}_i, T_i, \Delta_i)_{i=1,..,n}$ be an observed n-sample and $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ the associated filtered probability space, with

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

The mutual independence of (E, P, C) is a central assumption in survival literature (see e.g., Czado & Van Keilegom (2023) for recent discussion), even if the epidemiological interpretation makes it hard to justify. We here suppose only independent censorship, which, denoting $\mathcal{C}$ the survival copula (see Nelsen (2006)) of the random vector (E, P), writes:

$$(\mathcal{H}_{\mathcal{C}}): \ S_{P \wedge E}(t) = \mathcal{C}(S_P(t), S_E(t)).$$

In particular, the standard independence assumption writes simply $(\mathcal{H}_\Pi)$ for $\Pi(u_1, u_2) = u_1 u_2$. There are a few standard non-parametric estimators under $(\mathcal{H}_\Pi)$, from Ederer (1961), Hakulinen (1982), to more recently Pohar Perme & al (2012), but none under $(\mathcal{H}_{\mathcal{C}})$. If Adatorwovor & al (2023) provides parametric estimations under $(\mathcal{H}_{\mathcal{C}})$, we propose here a non-parametric estimator that generalizes the Pohar Perme estimator from $(\mathcal{H}_\Pi)$ to $(\mathcal{H}_{\mathcal{C}})$.

2 Estimation of the net survival under $(\mathcal{H}_{\mathcal{C}})$

Let's define the following stochastic processes:

$$\begin{aligned}
N(t) &= \mathbf{1}\{O \leq t, O \leq C\} \quad \textit{(Uncensored deaths process)} \\
Y(t) &= \mathbf{1}\{O \geq t, C \geq t\} \quad \textit{(At-risk process)} \\
N_E(t) &= \mathbf{1}\{E \leq t, E \leq C\} \quad \textit{(Excess uncensored deaths process)} \\
Y_E(t) &= \mathbf{1}\{E \geq t, C \geq t\} \quad \textit{(Excess at-risk process)}
\end{aligned}$$

Unfortunately, $(N_{E_i}, Y_{E_i})_{i \in 1,...,n}$ are not observable, but we show that

$$\begin{aligned}
\partial N_E(t) &= \frac{1}{a_t}\mathbb{E}(\partial N(t)|E, C) - \frac{b_t}{a_t c_t}\mathbb{E}(Y(t)|E, C), \\
Y_E(t) &= \frac{1}{c_t}\mathbb{E}(Y(t)|E, C),
\end{aligned}$$

where $a_t = \mathbb{P}(P \geq t|E = t)$, $b_t = \mathbb{P}(P = t|E \geq t)$ and $c_t = \mathbb{P}(P \geq t|E \geq t)$.

Assuming that (P, E) is an absolutely continuous random vector, a_t, b_t, c_t can be computed from partial derivatives $\mathcal{C}_i(\mathbf{u}) = \frac{\partial \mathcal{C}}{\partial u_i}(\mathbf{u}), i \in 1, 2$ of $\mathcal{C}$. Remark that under $(\mathcal{H}_\Pi)$, $a_t = c_t = S_P(t)$ and $b_t = -\partial S_P(t)$ do not depend on (unknown) $S_E(t)$, while they generally do under $(\mathcal{H}_\mathcal{C})$. However, like in classical survival analysis, we show under $(\mathcal{H}_\mathcal{C})$ that $\mathbb{E}\left(\partial N_E(t)\right) = \mathbb{E}\left(Y_E(t)\partial\Lambda_E(t)\right)$, and moreover that $N_{E_i} - \int_0^t Y_{E_i}\partial\Lambda_{E_i}$ are $\mathcal{F}_t$-martingales. A natural estimator for $\partial\Lambda_E(t)$ can therefore be simply constructed as

$$\widehat{\partial\Lambda_E}(t) = \frac{\frac{1}{n}\sum_{i=1}^n \partial N_{E_i}(t)}{\frac{1}{n}\sum_{i=1}^n Y_{E_i}(t)}.$$

However, $(\partial N_{E_i}, Y_{E_i})$ are not directly observable and need to be estimated. For that, replace first unobservable conditional expectations by their stochastic counterpart: $\mathbb{E}\left(\partial N_i(t)|E_i, C_i\right)$ by $\partial N_i(t)$ and $\mathbb{E}\left(Y_i(t)|E_i, C_i\right)$ by $Y_i(t)$. This plug-in is enough to make the estimator computable under $(\mathcal{H}_\Pi)$ (it is the Pohar Perme estimator), but not under $(\mathcal{H}_\mathcal{C})$. We call *generalized Pohar Perme estimator* a solution of the differential equation:

$$\widehat{\partial\Lambda_E}(t) = \frac{\sum_{i=1}^n \frac{1}{\widehat{a_{i,t}}}\partial N_i(t) - \frac{\widehat{b_{i,t}}}{\widehat{a_{i,t}}\widehat{c_{i,t}}}Y_i(t)}{\sum_{i=1}^n \frac{1}{\widehat{c_{i,t}}}Y_i(t)}, \tag{1}$$

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

$$\begin{aligned}
\widehat{a_{i,t}} &= \mathcal{C}_2\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right), \\
\widehat{b_{i,t}} &= -\mathcal{C}_1\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right)\partial S_{P_i}(t)e^{\widehat{\Lambda_E}(t)}, \\
\widehat{c_{i,t}} &= \mathcal{C}(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)})e^{\widehat{\Lambda_E}(t)}.
\end{aligned}$$

Unfortunately, the differential equation 1 is now non-separable, and a non-linear equation in $\widehat{\partial\Lambda_E}(t)$ needs to be solved at each time step. Alike previous estimators under $(\mathcal{H}_\Pi)$, the obtained $\widehat{\partial\Lambda_E}$ process is piecewise continuous, with jumps at event times $T_1, ..., T_n$. It is moreover always negative, except at jump points, making the produced survival curve increasing between jumps. The solving scheme must therefore be performed on a very dense mesh $t_1, ..., t_N$ that includes observed times $T_1, ...T_n$. These characteristics were already present under $(\mathcal{H}_\Pi)$.

3 Illustration

We use data on colorectal cancer patients extensively described in Wolski & al (2020). Population is separated along the primary tumor location (left or right). Using several dependence structures, we obtain net survival curves from Figure 1. If Wolski & al (2020) found the overall survival to be significantly different between left and right, net survival might not be if we take into account the uncertainty in $\mathcal{C}$. Further analysis to derive proper log-rank-type tests under $(\mathcal{H}_\mathcal{C})$ is possible.

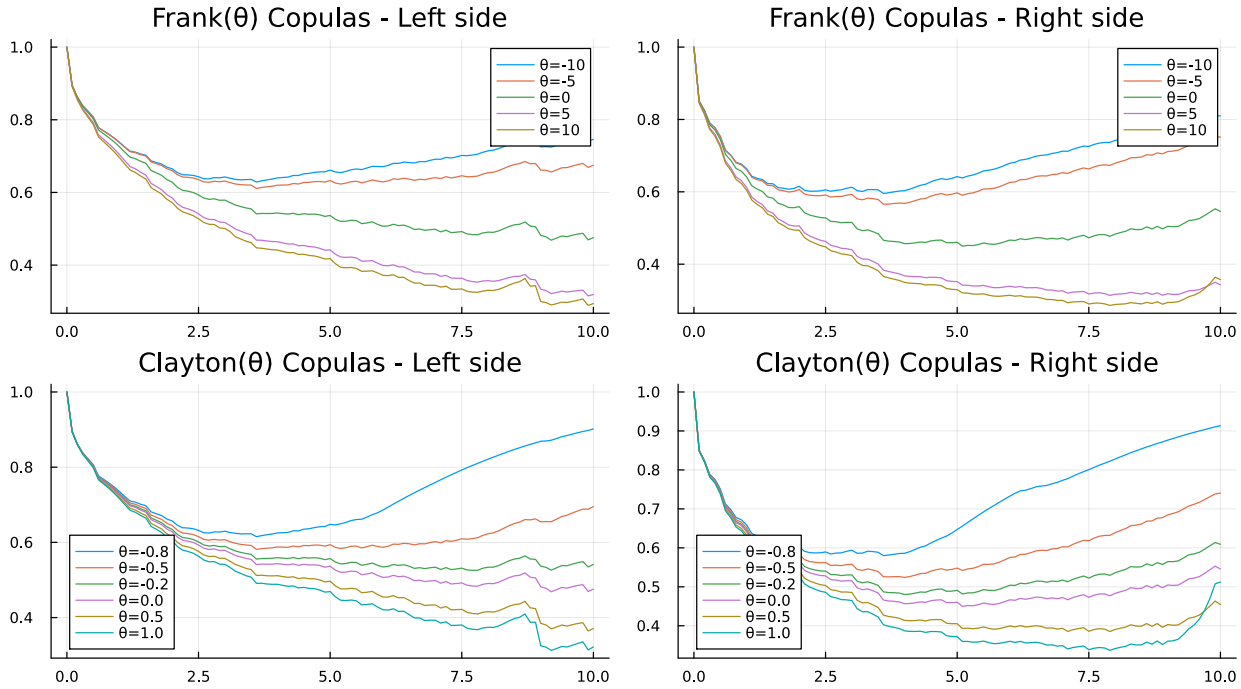

Figure 1: Obtained $\widehat{S_E}$ for several ($\mathcal{H}_{\mathcal{C}}$). Data was split w.r.t. tumor location (left or right) as in Wolski & al (2020), and several runs were done on several copulas: Frank copulas on the top and Clayton copulas on the bottom, with varying parameters θ. In both cases, $\theta = 0 \iff \mathcal{C} = \Pi$, and this curve represents the Pohar Perme estimator.

Bibliographie

Adatorwovor, R., Latouche, A., and Fine, J. P. (2023). A parametric approach to relaxing the independence assumption in relative survival analysis. In: *The international journal of biostatistics*, 18(2), 577-592.

Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on copulas. In: *Biometrika* 110(3), 721-738.

Ederer, F. (1961). The relative survival rate: a statistical methodology. In: *Natl. Cancer Inst. Monogr.*, vol. 6, p. 101-121, 1961.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. In : *Biometrics* 38, 933–942.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Pohar Perme, M., Stare, J., and Estève, J. (2012). On estimation in relative survival. In: *Biometrics*, vol. 68, no 1, p. 113-120

Wolski, A., Grafféo, N., Giorgi, R., and the CENSUR working survival group (2020). A permutation test based on the restricted mean survival time for comparison of net

survival distributions in non-proportional excess hazard settings. In: *Stat Methods Med Res*, vol. 29, no 6, p. 1612-1623