A comprehensive framework for time to event predictions with right censored data

Olivier Bouaziz
with Ariane Cwiling and Vittorio Perduca

MAP5 (CNRS 8145), Université Paris Cité

Séminaire Probabilités et Statistique, Laboratoire Paul Painlevé
Université de Lille

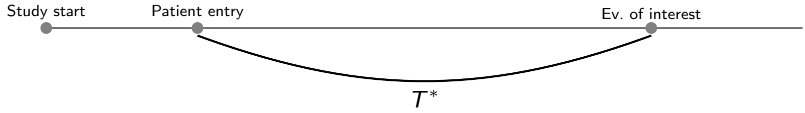
1. Background in time to event analysis
2. Prediction of the (restricted) time based on pseudo-observations
3. Evaluation of the quality of predictions
4. Conformal prediction intervals
5. Variable importance
6. Pseudo-observations and the super-learner

Outline

1. Background in time to event analysis
2. Prediction of the (restricted) time based on pseudo-observations
3. Evaluation of the quality of predictions
4. Conformal prediction intervals
5. Variable importance
6. Pseudo-observations and the super-learner

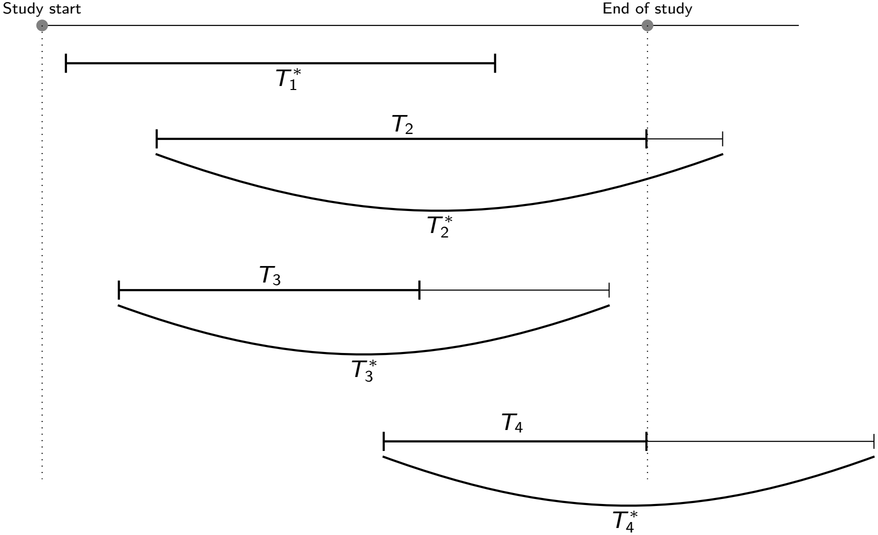
Background in time to event analysis

- We study a positive continuous time to event variable T^*.
- T^* represents the time difference between event of interest and patient entry.



- Examples : time to relapse of Leukemia patients, time to onset of cancer, time to death . . .

Background in time to event analysis : right censoring



Observations, assumptions, notations and quantities of interest

- Observations : for $i = 1, \dots, n$,

$$\begin{cases} T_i = T_i^* \wedge C_i \\ \Delta_i = \mathbb{1}_{T_i \leq C_i} \\ Z_i \in \mathbb{R}^d \end{cases}$$

- Independent censoring : $T^* \perp\!\!\!\perp C \mid Z$

- Quantities of interest :
  - The hazard rate :

$$\lambda(t \mid Z) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T^* < t + \triangle t \mid T^* \geq t, Z]}{\triangle t}$$

  - The survival function :

$$S(t \mid Z) := \mathbb{P}[T^* \geq t \mid Z]$$

  - The Restricted Mean Survival Time (RMST) :

$$\mu_\tau^*(Z) := \mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z) dt,$$

    for some $\tau > 0$.
  - . . .

Regression models

- The Cox model (proportional hazard) :

$$\lambda(t \mid Z) = \lambda_0(t) \exp(Z^\top \beta).$$

  For a binary covariate,

$$\frac{\lambda(t \mid Z=1)}{\lambda(t \mid Z=0)} = \exp(\beta).$$

- The survival function can be obtained by integrating out the hazard function :

$$S(t \mid Z) = \exp\left(-\int_0^t \lambda(t \mid Z) dt\right).$$

- The RMST can be obtained by computing $\int_0^\tau S(t \mid Z) dt$.
- References :
  - Direct modelling of the RMST Andersen P. K., Hansen M. G. and Klein J. P. **Lifetime Data Analysis** (2004) - Lu T., Zhao L. and Wei L. J. **Biostatistics** (2014) - Xin W. and Schaubel D. E. **Lifetime Data Analysis** (2018) - Zhao L. **Bioinformatics** (2021).
  - Random Survival Forests Ishwaran H and Kogalur U. B. **R news** (2007) - Ishwaran H, Kogalur U. B, Blackstone E. H. and Lauer M. S. **The Annals of Applied Statistics** (2008).
  - Super learner Golmakani M. K. and Polley E. C. **The International Journal of Biostatistics** (2020).

Objectives

1. Prediction of the (restricted) time using Machine Learning (ML) algorithms. (With a L^2 loss this is equivalent to estimating the RMST !)

2. Evaluation of the quality of prediction (MSE)

3. Computation of prediction intervals.

4. Variable importance assessment.

Outline

1. Background in time to event analysis

2. Prediction of the (restricted) time based on pseudo-observations

3. Evaluation of the quality of predictions

4. Conformal prediction intervals

5. Variable importance

6. Pseudo-observations and the super-learner

Backgrounds on pseudo-observations

Let $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$.

- $\hat{S}(t) := \hat{S}(X_1, \ldots, X_n)$ is the Kaplan-Meier (KM) estimator of the survival function.
- $\hat{S}^{(-\ell)}(t) := \hat{S}(X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots X_n)(t)$ is the jackknife (KM) estimator of the survival function.

Backgrounds on pseudo-observations

Let $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$.

- $\hat{S}(t) := \hat{S}(X_1, \ldots, X_n)$ is the Kaplan-Meier (KM) estimator of the survival function.
- $\hat{S}^{(-\ell)}(t) := \hat{S}(X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots X_n)(t)$ is the jackknife (KM) estimator of the survival function.
- The ℓ^{th} pseudo-observation is defined as :

$$\Gamma_\ell = n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{(-\ell)}(t)dt.$$

Backgrounds on pseudo-observations

Let $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$.

- $\hat{S}(t) := \hat{S}(X_1, \ldots, X_n)$ is the Kaplan-Meier (KM) estimator of the survival function.
- $\hat{S}^{(-\ell)}(t) := \hat{S}(X_1, \ldots, X_{\ell-1}, X_{\ell+1}, \ldots X_n)(t)$ is the jackknife (KM) estimator of the survival function.
- The ℓ^{th} pseudo-observation is defined as :

$$\Gamma_\ell = n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{(-\ell)}(t)dt.$$

- The goal is to estimate $\mu_\tau^*(Z_\ell) := \mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z_\ell)dt$ (RMST).
- We use Γ_ℓ as the response in a regression model.

Backgrounds on pseudo-observations

Let $X_i = (T_i, \Delta_i)$, $i = 1, \dots, n$.

- $\hat{S}(t) := \hat{S}(X_1, \dots, X_n)$ is the Kaplan-Meier (KM) estimator of the survival function.
- $\hat{S}^{(-\ell)}(t) := \hat{S}(X_1, \dots, X_{\ell-1}, X_{\ell+1}, \dots X_n)(t)$ is the jackknife (KM) estimator of the survival function.
- The ℓ^{th} pseudo-observation is defined as :

$$\Gamma_\ell = n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{(-\ell)}(t)dt.$$

- The goal is to estimate $\mu_\tau^*(Z_\ell) := \mathbb{E}[T^* \wedge \tau \mid Z] = \int_0^\tau S(t \mid Z_\ell)dt$ (RMST).
- We use Γ_ℓ as the response in a regression model.
- For example if there exists g known and invertible, a parameter $\beta \in \mathbb{R}^d$, such that $g(\mu_\tau^*(Z_\ell)) = Z_\ell^\top \beta$, then we can estimate β from least-square regression :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{\ell=1}^n \left(\hat{\Gamma}_{(\ell)} - g^{-1}(Z_\ell^\top \beta) \right)^2,$$

and $\hat{\mu}_{\tau,n}(Z_\ell) = g^{-1}(Z_\ell^\top \hat{\beta})$.

Andersen P. K., Klein J. P. and Rosthøj S. *Generalised linear models for correlated pseudo-observations, with applications to multi-state models*. **Biometrika** (2003).

Examples

1. Completely observed data : $X_i = T_i^*$, $\mu^*(Z_\ell) = \mathbb{E}[T_\ell^* \mid Z_\ell]$. Let $\theta = \mathbb{E}[T^*]$,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\theta}^{(-\ell)} = \frac{1}{n-1}\sum_{i\neq\ell}^{n} X_i,$$

$$\Gamma_\ell = n\hat{\theta} - (n-1)\hat{\theta}^{(-\ell)} = \sum_{i=1}^{n} X_i - \sum_{i\neq\ell}^{n} X_i = X_\ell.$$

Clearly : $\mathbb{E}[\Gamma_\ell \mid Z_\ell] = \mu^*(Z_\ell)$.

Examples

1. Completely observed data : $X_i = T_i^*$, $\mu^*(Z_\ell) = \mathbb{E}[T_\ell^* \mid Z_\ell]$. Let $\theta = \mathbb{E}[T^*]$,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\theta}^{(-\ell)} = \frac{1}{n-1}\sum_{i \neq \ell}^{n} X_i,$$

$$\Gamma_\ell = n\hat{\theta} - (n-1)\hat{\theta}^{(-\ell)} = \sum_{i=1}^{n} X_i - \sum_{i \neq \ell}^{n} X_i = X_\ell.$$

Clearly : $\mathbb{E}[\Gamma_\ell \mid Z_\ell] = \mu^*(Z_\ell)$.

2. Right censored data :

$$\Gamma_\ell = n\int_0^\tau \hat{S}(t)dt - (n-1)\int_0^\tau \hat{S}^{(-\ell)}(t)dt.$$

What can be said about $\mathbb{E}[\Gamma_\ell \mid Z_\ell]$?

Theoretical result for right-censored data

Proposition (Graw, Gerds, Schumacher 2009 ; Jacobsen, Martinussen 2016)

In the context of right-censored data (i.e. $X = (T, \Delta)$), assume that

- $C \perp\!\!\!\perp (T^*, Z)$,
- $\exists \tau > 0, \mathbb{P}[T \geq \tau] > 0$.

Then, for all $t \in [0, \tau]$, with $\hat{S}$ the Kaplan-Meier estimator,

$$n\hat{S}(t) - (n-1)\hat{S}^{(-\ell)}(t) = S(t) + \dot{\psi}(X_\ell, t) + O_{\mathbb{P}}(n^{-1/2}),$$

where $\dot{\psi}$ is the first order influence function defined as :

$$\dot{\psi}(X_\ell, t) = -S(t)\left(\frac{\mathbb{1}_{T_\ell \leq t, \Delta_\ell = 1}}{H(T_\ell)} - \int_0^{t \wedge T_\ell} \frac{dH_1(u)}{(H(u))^2}\right).$$

Moreover,

$$\mathbb{E}[\dot{\psi}(X_\ell, t) \mid Z_\ell] = -S(t)\left(1 - \frac{S(t \mid Z_\ell)}{S(t)}\right).$$

Notations : $H(\cdot) = \mathbb{P}[T \geq \cdot]$, $H_1(\cdot) = \mathbb{P}[T \leq \cdot, \Delta = 1]$.

Outline

1 Background in time to event analysis

2 Prediction of the (restricted) time based on pseudo-observations

3 Evaluation of the quality of predictions

4 Conformal prediction intervals

5 Variable importance

6 Pseudo-observations and the super-learner

Quality of prediction assessed with the Mean Squared Error (MSE)

- Observations : $D_n = \{O_i := (T_i, \Delta_i, Z_i), i = 1 \ldots, n\}$.
- Train and test sets : $D_n = D_{n_{\text{train}}} \cup D_{n_{\text{test}}}$, $D_{n_{\text{train}}} \cap D_{n_{\text{test}}} = \emptyset$, $n_{\text{train}} = \lfloor \rho n \rfloor$, $n_{\text{test}} = \lceil (1-\rho) n \rceil$, $0 < \rho < 1$.
    - We implement a learning algorithm $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{train}}}$
    - We assess the quality of prediction of $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{test}}}$
- We assume there exists $\tilde{\mu}_\tau$ such that $\hat{\mu}_{\tau, n_{\text{train}}} \to_{\mathbb{P}} \tilde{\mu}_\tau$, $n \to \infty$.

Quality of prediction assessed with the Mean Squared Error (MSE)

- Observations : $D_n = \{O_i := (T_i, \Delta_i, Z_i), i = 1 \ldots, n\}$.
- Train and test sets : $D_n = D_{n_{\text{train}}} \cup D_{n_{\text{test}}}$, $D_{n_{\text{train}}} \cap D_{n_{\text{test}}} = \emptyset$, $n_{\text{train}} = \lfloor \rho n \rfloor$, $n_{\text{test}} = \lceil (1-\rho) n \rceil$, $0 < \rho < 1$.
  - We implement a learning algorithm $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{train}}}$
  - We assess the quality of prediction of $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{test}}}$
- We assume there exists $\tilde{\mu}_\tau$ such that $\hat{\mu}_{\tau, n_{\text{train}}} \rightarrow_{\mathbb{P}} \tilde{\mu}_\tau$, $n \rightarrow \infty$.

1. In the absence of censoring

   We use the Residual Sum of Squares (RSS) :

$$\text{RSS}(\hat{\mu}_{\tau, n_{\text{train}}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(T_i^* \wedge \tau - \hat{\mu}_{\tau, n_{\text{train}}}(Z_i) \Big)^2,$$

   which converges in probability as n tends to infinity, to

$$\mathbb{E}\left[(T^* \wedge \tau - \tilde{\mu}_\tau(Z))^2 \right] = \underbrace{\mathbb{E}\Big[\big(\mu_\tau^*(Z) - \tilde{\mu}_\tau(Z) \big)^2 \Big]}_{\text{imprecision}} + \underbrace{\mathbb{E}\Big[\big(T^* \wedge \tau - \mu_\tau^*(Z) \big)^2 \Big]}_{\text{inseparability}}.$$

Quality of prediction assessed with the Mean Squared Error (MSE)

- Observations : $D_n = \{O_i := (T_i, \Delta_i, Z_i), i = 1 \ldots, n\}$.
- Train and test sets : $D_n = D_{n_{\text{train}}} \cup D_{n_{\text{test}}}$, $D_{n_{\text{train}}} \cap D_{n_{\text{test}}} = \emptyset$, $n_{\text{train}} = \lfloor \rho n \rfloor$, $n_{\text{test}} = \lceil (1-\rho) n \rceil$, $0 < \rho < 1$.
  - We implement a learning algorithm $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{train}}}$
  - We assess the quality of prediction of $\hat{\mu}_{\tau, n_{\text{train}}}$ on $D_{n_{\text{test}}}$
- We assume there exists $\tilde{\mu}_\tau$ such that $\hat{\mu}_{\tau, n_{\text{train}}} \to_{\mathbb{P}} \tilde{\mu}_\tau$, $n \to \infty$.

1. In the absence of censoring

   We use the Residual Sum of Squares (RSS) :

$$\text{RSS}(\hat{\mu}_{\tau, n_{\text{train}}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(T_i^* \wedge \tau - \hat{\mu}_{\tau, n_{\text{train}}}(Z_i) \Big)^2,$$

   which converges in probability as n tends to infinity, to

$$\mathbb{E}\left[(T^* \wedge \tau - \tilde{\mu}_\tau(Z))^2 \right] = \underbrace{\mathbb{E}\Big[\big(\mu_\tau^*(Z) - \tilde{\mu}_\tau(Z)\big)^2 \Big]}_{\text{imprecision}} + \underbrace{\mathbb{E}\Big[\big(T^* \wedge \tau - \mu_\tau^*(Z)\big)^2 \Big]}_{\text{inseparability}}.$$

2. With censored data

   We weight the observations using Inverse Probability of Censoring Weighting (IPCW).

Rationale behind IPCW

- For any measurable, bounded function h, we have (using the conditional independence assumption of censoring) :

$$\mathbb{E}\left[\frac{h(T_i, Z_i)\mathbb{1}_{T_i \leq \tau}\Delta_i}{1 - G(T_i - \mid Z_i)}\right] = \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau, T_i^* \leq C_i}}{1 - G(T_i^* - \mid Z_i)}\right]$$

Rationale behind IPCW

- For any measurable, bounded function h, we have (using the conditional independence assumption of censoring) :

$$\begin{aligned}\mathbb{E}\left[\frac{h(T_i, Z_i)\mathbb{1}_{T_i \leq \tau}\Delta_i}{1 - G(T_i - \mid Z_i)}\right] &= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau, T_i^* \leq C_i}}{1 - G(T_i^* - \mid Z_i)}\right] \\ &= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau}}{1 - G(T_i^* - \mid Z_i)}\mathbb{E}\left[\mathbb{1}_{T_i^* \leq C_i} \mid T_i^*, Z_i\right]\right]\end{aligned}$$

Rationale behind IPCW

- For any measurable, bounded function h, we have (using the conditional independence assumption of censoring) :

$$\begin{aligned}\mathbb{E}\left[\frac{h(T_i, Z_i)\mathbb{1}_{T_i \leq \tau}\Delta_i}{1 - G(T_i - \mid Z_i)}\right] &= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau, T_i^* \leq C_i}}{1 - G(T_i^* - \mid Z_i)}\right] \\ &= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau}}{1 - G(T_i^* - \mid Z_i)} \underbrace{\mathbb{E}\left[\mathbb{1}_{T_i^* \leq C_i} \mid T_i^*, Z_i\right]}_{1 - G(T_i^* - \mid Z_i)}\right] \\ &= \mathbb{E}\left[h(T_i^*, Z_i)\mathbb{1}_{T_i^* \leq \tau}\right],\end{aligned}$$

with $G(t \mid Z) := \mathbb{P}[C \leq t \mid Z]$.

Rationale behind IPCW

- For any measurable, bounded function h, we have (using the conditional independence assumption of censoring) :

$$\begin{aligned}
\mathbb{E}\left[\frac{h(T_i, Z_i)\mathbb{1}_{T_i \le \tau}\Delta_i}{1 - G(T_i - \mid Z_i)}\right] &= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \le \tau, T_i^* \le C_i}}{1 - G(T_i^* - \mid Z_i)}\right] \\
&= \mathbb{E}\left[\frac{h(T_i^*, Z_i)\mathbb{1}_{T_i^* \le \tau}}{1 - G(T_i^* - \mid Z_i)} \underbrace{\mathbb{E}\left[\mathbb{1}_{T_i^* \le C_i} \mid T_i^*, Z_i\right]}_{1 - G(T_i^* - \mid Z_i)}\right] \\
&= \mathbb{E}\left[h(T_i^*, Z_i)\mathbb{1}_{T_i^* \le \tau}\right],
\end{aligned}$$

with $G(t \mid Z) := \mathbb{P}[C \le t \mid Z]$.

- In practice, we have to estimate G by $\hat{G}_n$ such that :

$$\frac{1}{n}\sum_{i=1}^{n} \frac{h(T_i, Z_i)\mathbb{1}_{T_i \le \tau}\Delta_i}{1 - \hat{G}_n(T_i - \mid Z_i)} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{h(T_i, Z_i)\mathbb{1}_{T_i \le \tau}\Delta_i}{1 - G(T_i - \mid Z_i)} \xrightarrow[n \to \infty]{\mathbb{P}} \mathbb{E}\left[h(T_i^*, Z_i)\mathbb{1}_{T_i^* \le \tau}\right].$$

Evaluation of the quality of prediction with censored data

- Observations : $D_n = \{O_i := (T_i, \Delta_i, Z_i), i = 1\ldots, n\}$.
- Train and test sets : $D_n = D_{n_{\text{train}}} \cup D_{n_{\text{test}}}$, $D_{n_{\text{train}}} \cap D_{n_{\text{test}}} = \emptyset$.
  - We implement a learning algorithm $\hat{\mu}_{\tau,n_{\text{train}}}$ on $D_{n_{\text{train}}}$ ($n_{\text{train}} = \lfloor \rho n \rfloor$),
  - We assess the quality of prediction of $\hat{\mu}_{\tau,n_{\text{train}}}$ on $D_{n_{\text{test}}}$ ($n_{\text{test}} = \lceil (1-\rho)n \rceil$).
- We use the following Weighted Residual Sum of Squares (WRSS) :

$$\text{WRSS}(\hat{\mu}_{\tau,n_{\text{train}}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(T_i \wedge \tau - \hat{\mu}_{\tau,n_{\text{train}}}(Z_i) \Big)^2 \hat{\omega}_i,$$

$$\hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{1 - \hat{G}_n(T_i - \mid Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{1 - \hat{G}_n(\tau \mid Z_i)} \text{ (IPCW)}.$$

Evaluation of the quality of prediction with censored data

- Observations : $D_n = \{O_i := (T_i, \Delta_i, Z_i), i = 1\dots, n\}$.
- Train and test sets : $D_n = D_{n_{\text{train}}} \cup D_{n_{\text{test}}}$, $D_{n_{\text{train}}} \cap D_{n_{\text{test}}} = \emptyset$.
  - We implement a learning algorithm $\hat{\mu}_{\tau,n_{\text{train}}}$ on $D_{n_{\text{train}}}$ ($n_{\text{train}} = \lfloor \rho n \rfloor$),
  - We assess the quality of prediction of $\hat{\mu}_{\tau,n_{\text{train}}}$ on $D_{n_{\text{test}}}$ ($n_{\text{test}} = \lceil (1-\rho)n \rceil$).
- We use the following Weighted Residual Sum of Squares (WRSS) :

$$\text{WRSS}(\hat{\mu}_{\tau,n_{\text{train}}}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(T_i \wedge \tau - \hat{\mu}_{\tau,n_{\text{train}}}(Z_i) \Big)^2 \hat{\omega}_i,$$

$$\hat{\omega}_i = \frac{\mathbb{1}(T_i \le \tau)\Delta_i}{1 - \hat{G}_n(T_i- \mid Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{1 - \hat{G}_n(\tau \mid Z_i)} \text{ (IPCW)}.$$

Assume :

- $C \perp\!\!\!\perp T^* \mid Z$,
- $\iint \big|\hat{G}_n(s \mid z) - G(s \mid z)\big| d\mathbb{P}(s,z) \xrightarrow[n\to\infty]{\mathbb{P}} 0$, $\int |\hat{\mu}_{\tau,n_{\text{train}}}(z) - \tilde{\mu}_\tau(z)| d\mathbb{P}(z) \xrightarrow[n\to\infty]{\mathbb{P}} 0$.

Then, $\text{WRSS}(\hat{\mu}_{\tau,n_{\text{train}}})$ converges in probability, as $n \to \infty$, towards :

$$\mathbb{E}\Big[(T^* \wedge \tau - \tilde{\mu}_\tau(Z))^2\Big] = \underbrace{\mathbb{E}\Big[\big(\mu^*_\tau(Z) - \tilde{\mu}_\tau(Z)\big)^2\Big]}_{\text{imprecision}} + \underbrace{\mathbb{E}\Big[\big(T^* \wedge \tau - \mu^*_\tau(Z)\big)^2\Big]}_{\text{inseparability}}.$$

Illustration on simulated data

- **Scenario A** : Linear model.

  $$T_i^* = \tilde{\beta}_0^\top Z_i + \varepsilon_i,$$

  where $\tilde{\beta}_0 = (5.5, 2.5, 2.5)^\top$, $Z_i = (1, Z_i^1, Z_i^2)^\top$, $Z_i^1, Z_i^2 \sim \mathcal{B}(0.5)$, $\varepsilon_i \sim U[-3, 3]$.

- Closed form for the RMST :

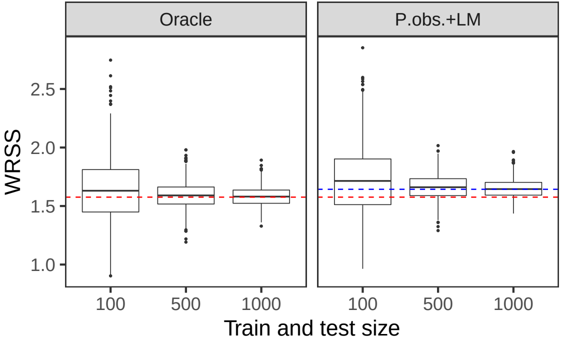
  $$\mu_\tau^*(Z) = \beta_{00} + \beta_{01} Z^1(1 - Z^2) + \beta_{10} Z^2(1 - Z^1) + \beta_{11} Z^1 Z^2,$$

  where $\beta_0 = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})^\top = (5.5, 2.1, 2.1, 3.2)^\top$, $\tau = 8.8$.

  - scheme **A1** : censoring is independent from the covariates, exponential law with parameter $\alpha = 0.07$
  - scheme **A2** : $C \sim$ Cox model, $\lambda(t \mid Z) = \lambda_0(t) \exp(\beta_1 Z^1 + \beta_2 Z^2)$ with Weibull baseline hazard $\mathcal{W}(\nu, \kappa)$, $\nu = 6, \kappa = 12$.

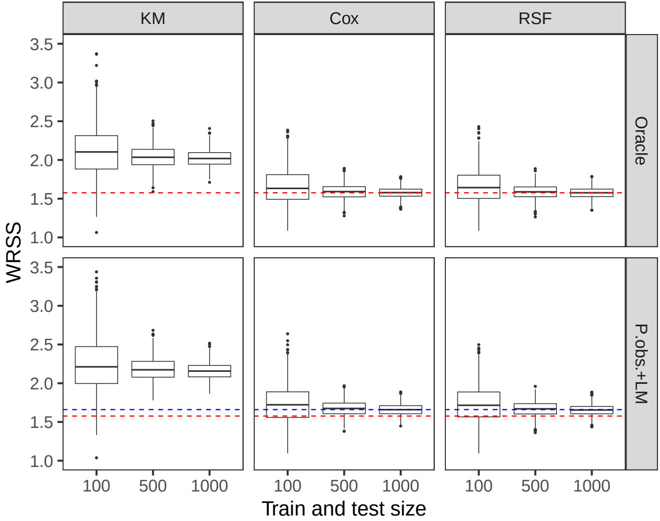
44% of censored data in the two settings.

Scenario A1



- Censoring is estimated using the Kaplan-Meier estimator.
- In blue : imprecision+inseparability.
- In red : inseparability.

Scenario A2



Censoring distribution is estimated using the Kaplan-Meier estimator, the Cox model or the Random Survival Forests (RSF).

Outline

1. Background in time to event analysis
2. Prediction of the (restricted) time based on pseudo-observations
3. Evaluation of the quality of predictions
4. Conformal prediction intervals
5. Variable importance
6. Pseudo-observations and the super-learner

Presentation of split conformal intervals

$D_n = D_{n_1} \cup D_{n_2}$, $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$.

1. In the absence of censoring :
   - Train $\hat{\mu}_{\tau,n_1}$ on D_{n_1}.
   - Compute the residuals $R_i = |T_i^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|$, for i in D_{n_2}.

Presentation of split conformal intervals

$D_n = D_{n_1} \cup D_{n_2}$, $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$.

1. In the absence of censoring :
   - Train $\hat{\mu}_{\tau,n_1}$ on D_{n_1}.
   - Compute the residuals $R_i = |T_i^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|$, for i in D_{n_2}.

   Then compute the following estimators,

$$\mathcal{R}_{n_2}(t) := \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \mathbb{1}_{R_i \leq t}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

Presentation of split conformal intervals

$D_n = D_{n_1} \cup D_{n_2}$, $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$.

1. In the absence of censoring :
   - Train $\hat{\mu}_{\tau,n_1}$ on D_{n_1}.
   - Compute the residuals $R_i = |T_i^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|$, for i in D_{n_2}.

   Then compute the following estimators,

$$\mathcal{R}_{n_2}(t) := \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \mathbb{1}_{R_i \leq t}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

   For a new individual O_j, define :

$$\mathcal{C}_{n_2}^{\text{split}}(Z_j) = [\hat{\mu}_{\tau,n_1}(Z_j) - \hat{q}_{n_2}(\alpha), \hat{\mu}_{\tau,n_1}(Z_j) + \hat{q}_{n_2}(\alpha)].$$

Presentation of split conformal intervals

$D_n = D_{n_1} \cup D_{n_2},\ D_{n_1} = \{O_i : i \in \mathcal{I}_1\},\ D_{n_2} = \{O_i : i \in \mathcal{I}_2\},\ \mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset.$

1. In the absence of censoring :
   - Train $\hat{\mu}_{\tau,n_1}$ on D_{n_1}.
   - Compute the residuals $R_i = |T_i^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|$, for i in D_{n_2}.

   Then compute the following estimators,

$$\mathcal{R}_{n_2}(t) := \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \mathbb{1}_{R_i \leq t}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

   For a new individual O_j, define :

$$\mathcal{C}_{n_2}^{\text{split}}(Z_j) = [\hat{\mu}_{\tau,n_1}(Z_j) - \hat{q}_{n_2}(\alpha), \hat{\mu}_{\tau,n_1}(Z_j) + \hat{q}_{n_2}(\alpha)].$$

   Using the exchangeability property, we can prove that :

$$\mathbb{P}\left[T_j^* \wedge \tau \in \mathcal{C}_{n_2}^{\text{split}}(Z_j)\right] \geq 1 - \alpha.$$

Vovk V., Gammerman A. and Shafer G. *Algorithmic learning in a random world*. **Springer** (2005).

Lei, J., G'Sell M., Rinaldo A., Tibshirani R. J. and Wasserman L. *Distribution-Free Predictive Inference for Regression*. **JASA** (2018).

Presentation of split conformal intervals

$D_n = D_{n_1} \cup D_{n_2}$, $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$.

1. In the absence of censoring :
   - Train $\hat{\mu}_{\tau,n_1}$ on D_{n_1}.
   - Compute the residuals $R_i = |T_i^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|$, for i in D_{n_2}.

   Then compute the following estimators,

$$\mathcal{R}_{n_2}(t) := \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \mathbb{1}_{R_i \leq t}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

   For a new individual O_j, define :

$$\mathcal{C}_{n_2}^{\text{split}}(Z_j) = [\hat{\mu}_{\tau,n_1}(Z_j) - \hat{q}_{n_2}(\alpha), \hat{\mu}_{\tau,n_1}(Z_j) + \hat{q}_{n_2}(\alpha)].$$

   Using the exchangeability property, we can prove that :

$$\mathbb{P}\left[T_j^* \wedge \tau \in \mathcal{C}_{n_2}^{\text{split}}(Z_j)\right] \geq 1 - \alpha.$$

   Vovk V., Gammerman A. and Shafer G. *Algorithmic learning in a random world*. **Springer** (2005).

   Lei, J., G'Sell M., Rinaldo A., Tibshirani R. J. and Wasserman L. *Distribution-Free Predictive Inference for Regression*. **JASA** (2018).

2. With censoring : we use IPCW.

Split conformal intervals with right-censored data

Let $\hat{G}_n$ be a consistent estimator of G. For $i \in \mathcal{I}_2$ define :

$$R_i = |T_i \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|, \quad \hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{1 - \hat{G}_{n_2}(T_i - \mid Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{1 - \hat{G}_{n_2}(\tau \mid Z_i)} \text{ (IPCW)},$$

$$\mathcal{R}_{n_2}(t) := \frac{\sum_{i \in \mathcal{I}_2} \hat{\omega}_i \mathbb{1}_{R_i \leq t}}{\sum_{i \in \mathcal{I}_2} \hat{\omega}_i}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

For a new individual O_j, define :

$$\mathcal{C}_{n_2}^{\text{split}}(Z_j) = [\hat{\mu}_{\tau,n_1}(Z_j) - \hat{q}_{n_2}(\alpha), \hat{\mu}_{\tau,n_1}(Z_j) + \hat{q}_{n_2}(\alpha)]$$

Split conformal intervals with right-censored data

Let $\hat{G}_n$ be a consistent estimator of G. For $i \in \mathcal{I}_2$ define :

$$R_i = |T_i \wedge \tau - \hat{\mu}_{\tau,n_1}(Z_i)|, \quad \hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{1 - \hat{G}_{n_2}(T_i - \mid Z_i)} + \frac{\mathbb{1}(T_i > \tau)}{1 - \hat{G}_{n_2}(\tau \mid Z_i)} \text{ (IPCW)},$$

$$\mathcal{R}_{n_2}(t) := \frac{\sum_{i \in \mathcal{I}_2} \hat{\omega}_i \mathbb{1}_{R_i \leq t}}{\sum_{i \in \mathcal{I}_2} \hat{\omega}_i}, \quad \hat{q}_{n_2}(\alpha) := \inf\{t : \mathcal{R}_{n_2}(t) \geq 1 - \alpha\}.$$

For a new individual O_j, define :

$$\mathcal{C}_{n_2}^{\text{split}}(Z_j) = [\hat{\mu}_{\tau,n_1}(Z_j) - \hat{q}_{n_2}(\alpha), \hat{\mu}_{\tau,n_1}(Z_j) + \hat{q}_{n_2}(\alpha)]$$

Theoretical result for the split conformal prediction interval

Assume

- $C \perp\!\!\!\perp T^* \mid Z$
- $\sup_{s \leq \tau, z \in \mathbb{R}^d} \left|\hat{G}_{n_2}(s \mid z) - G(s \mid z)\right| \xrightarrow[n_2 \to \infty]{\mathbb{P}} 0.$

Then

$$\lim_{n_2 \to \infty} \mathbb{P}\left[T_j^* \wedge \tau \in \mathcal{C}_{n_2}^{\text{split}}(Z_j)\right] \geq 1 - \alpha.$$

Split conformal prediction intervals : simulation design

Scenario B : Cox model.

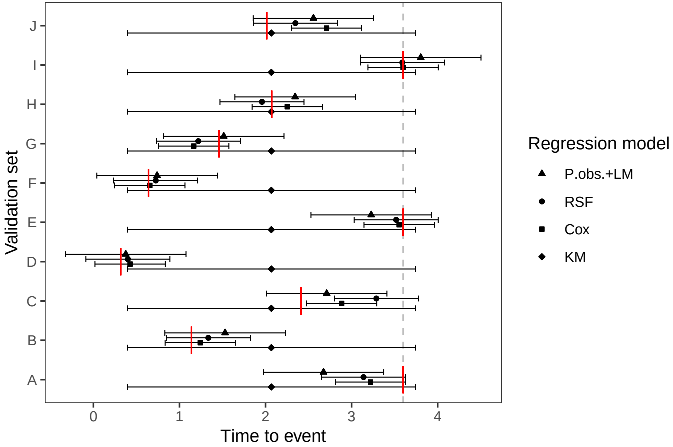
- $T_i^* \sim \lambda_0(t)\exp(\beta_0^\top Z_i)$, $\lambda_0 \sim \mathcal{W}(\nu, \kappa)$, $\nu = 6, \kappa = 2$.
- $Z \in \mathbb{R}^3$, $Z^k \sim \mathcal{U}[-5, 5]$, $k = 1, 2, 3$, $\beta_0 = (2, 1, 0)^\top$.
- $C \sim \mathcal{E}(0.3)$, censoring rate : 47%.
- $\tau = 3.6$.

Learning models :

- Linear model applied on pseudo-observations.
- Random Survival Forests.
- Cox model.
- Kaplan-Meier estimator.

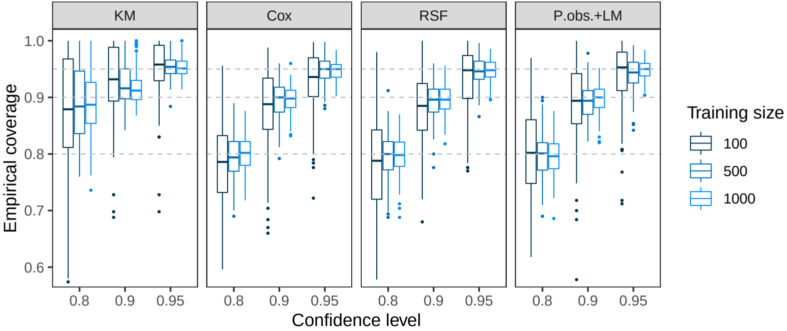
Illustrations with IPCW Rank-One-Out (ROO) split conformal intervals.

Split conformal prediction intervals : illustration on 10 observations



- Prediction intervals at the 90% level.
- $n = 4,000$ in IPCW ROO split conformal intervals.
- $T_i^* \wedge \tau$ is represented in red.
- $\tau = 3.6$ is represented by a dotted vertical line.

Split conformal prediction intervals : coverage property



- IPCW Rank-One-Out (ROO) split conformal intervals.
- Test size : 500.
- Number of replications : 200.

Outline

1 Background in time to event analysis

2 Prediction of the (restricted) time based on pseudo-observations

3 Evaluation of the quality of predictions

4 Conformal prediction intervals

5 Variable importance

6 Pseudo-observations and the super-learner

Global variable importance based on Leave One Covariate Out (LOCO)

Let $\hat{\mu}_{\tau,n_1}^{(-k)}(Z)$ be the learner trained on D_{n_1} without covariate Z^k. Define :

$$\delta_k(T^*, Z) = |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}^{(-k)}(Z)| - |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z)|$$

and

$$\begin{aligned} m_k &= \text{median}\left[\delta_k(T^*, Z) \mid T^* \leq \tau\right], \\ p_k &= \mathbb{P}[\delta_k(T^*, Z) \geq 0 \mid T^* \leq \tau]. \end{aligned}$$

Global variable importance based on Leave One Covariate Out (LOCO)

Let $\hat{\mu}_{\tau,n_1}^{(-k)}(Z)$ be the learner trained on D_{n_1} without covariate Z^k. Define :

$$\delta_k(T^*, Z) = |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}^{(-k)}(Z)| - |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z)|$$

and

$$\begin{aligned} m_k &= \text{median}\left[\delta_k(T^*, Z) \mid T^* \le \tau\right], \\ p_k &= \mathbb{P}[\delta_k(T^*, Z) \ge 0 \mid T^* \le \tau]. \end{aligned}$$

Goals :

- Construct asymptotic confidence intervals for p_k.
- Construct an asymptotic test

  $$(H_0) : m_k \le 0 \text{ versus } (H_1) : m_k > 0,$$

  or equivalently

  $$(H_0) : p_k \le 1/2 \text{ versus } (H_1) : p_k > 1/2.$$

  ⚠ (H_0) is composite.

Global variable importance based on Leave One Covariate Out (LOCO)

Let $\hat{\mu}_{\tau,n_1}^{(-k)}(Z)$ be the learner trained on D_{n_1} without covariate Z^k. Define :

$$\delta_k(T^*, Z) = |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}^{(-k)}(Z)| - |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z)|$$

and

$$m_k = \text{median}\left[\delta_k(T^*, Z) \mid T^* \leq \tau\right], \\ p_k = \mathbb{P}[\delta_k(T^*, Z) \geq 0 \mid T^* \leq \tau].$$

Goals :

- Construct asymptotic confidence intervals for p_k.
- Construct an asymptotic test

  $$(H_0) : m_k \leq 0 \text{ versus } (H_1) : m_k > 0,$$

  or equivalently

  $$(H_0) : p_k \leq 1/2 \text{ versus } (H_1) : p_k > 1/2.$$

  ⚠ (H_0) is composite.

The derivation of the test is based on Kaplan-Meier integral properties.

Statistical test for global variable importance

- $p_k = \mathbb{P}[\delta_k(T^*, Z) \geq 0 \mid T^* \leq \tau]$.
- Let

$$\Phi_k(u, z) = \mathbb{1}_{|u - \hat{\mu}^{(-k)}_{\tau,n_1}(z)| - |u - \hat{\mu}_{\tau,n_1}(z)| \geq 0, 0 \leq u \leq \tau},$$

  such that

$$p_k = \frac{1}{1 - S(\tau)} \iint \Phi_k(u, z) dF(u, z).$$

Statistical test for global variable importance

- $p_k = \mathbb{P}[\delta_k(T^*, Z) \geq 0 \mid T^* \leq \tau]$.
- Let

$$\Phi_k(u,z) = \mathbb{1}_{|u-\hat{\mu}^{(-k)}_{\tau,n_1}(z)|-|u-\hat{\mu}_{\tau,n_1}(z)|\geq 0, 0\leq u\leq\tau},$$

such that

$$p_k = \frac{1}{1-S(\tau)} \iint \Phi_k(u,z)dF(u,z).$$

The test statistic is :

$$\mathcal{T}_{n_2} = \sqrt{\frac{n_2}{\hat{\sigma}^2(\Phi_k)}}\left(\frac{1}{1-\hat{S}_{n_2}(\tau)} \underbrace{\iint \Phi_k(u,z)d\hat{F}_{n_2}(u,z)}_{\frac{1}{n_2}\sum_{i\in\mathcal{I}_2}\Phi_k(T_i,Z_i)\hat{\omega}_i} - \frac{1}{2}\right),$$

- $\hat{\omega}_i$, $i \in \mathcal{I}_2$ are the IPCW weights

$$\hat{\omega}_i = \frac{\mathbb{1}(T_i \leq \tau)\Delta_i}{1-\hat{G}_{n_2}(T_i-)},$$

- $\hat{S}_{n_2}$: Kaplan-Meier estimator of the survival function of T^*,
- $1-\hat{G}_{n_2}$: Kaplan-Meier estimator of the survival function of C,
- $\hat{\sigma}^2(\Phi_k)$: estimator of the asymptotic variance.

Ingredients for deriving the asymptotic distribution of the test statistic

1. Decomposition of Kaplan-Meier integrals as sums of i.i.d. terms.

$$
\begin{aligned}
&\frac{\sqrt{n_2}}{1-S(\tau)}\iint \Phi_k(u,z)d(\hat{F}_{n_2}-F)(u,z)\\
&=\frac{\sqrt{n_2}}{1-S(\tau)}\frac{1}{n_2}\sum_{i\in\mathcal{I}_2}\left(\Phi_k(T_i,Z_i)\frac{\Delta_i\mathbb{1}_{T_i\leq\tau}}{1-G(T_i-)}-\iint \Phi_k(u,z)dF(u,z)+\gamma_1(\Phi_k;T_i,\Delta_i)\right)\\
&\quad+O_{\mathbb{P}}(n_2^{-1/2}),
\end{aligned}
$$

with $\mathbb{E}\left[\gamma_1(\Phi_k;T_i,\Delta_i)\right]=0$.

Ingredients for deriving the asymptotic distribution of the test statistic

1. Decomposition of Kaplan-Meier integrals as sums of i.i.d. terms.

$$\frac{\sqrt{n_2}}{1-S(\tau)}\iint \Phi_k(u,z)d(\hat{F}_{n_2}-F)(u,z)$$
$$=\frac{\sqrt{n_2}}{1-S(\tau)}\frac{1}{n_2}\sum_{i\in\mathcal{I}_2}\left(\Phi_k(T_i,Z_i)\frac{\Delta_i\mathbb{1}_{T_i\leq\tau}}{1-G(T_i-)}-\iint\Phi_k(u,z)dF(u,z)+\gamma_1(\Phi_k;T_i,\Delta_i)\right)$$
$$+O_{\mathbb{P}}(n_2^{-1/2}),$$

with $\mathbb{E}\left[\gamma_1(\Phi_k;T_i,\Delta_i)\right]=0$.

2. Decomposition of the Kaplan-Meier estimator as a martingale process.

$$\sqrt{n_2}\left(\hat{S}_{n_2}(\tau)-S(\tau)\right)=-S(\tau)\frac{1}{\sqrt{n_2}}\sum_{i\in\mathcal{I}_2}\int_0^\tau\frac{dM_i(u)}{1-H(u)}+o_{\mathbb{P}}(1),$$

where

$$N_i(t)=\mathbb{1}_{T_i\leq t,\Delta_i=1},\quad Y_i(t)=\mathbb{1}_{T_i\geq t},\quad 1-H(t)=\mathbb{P}[T\geq t],$$
$$M_i(t)=N_i(t)-\int_0^t Y_i(u)\lambda(u)du,$$

$M_i(t)$ is a martingale with respect to the filtration $\mathcal{F}_t^i=\sigma(N_i(u),Y_i(u):0\leq u\leq t)$.

Asymptotic results for the global variable importance based on LOCO

- D_{n_1} is fixed,
- $\delta_k(T^*, Z) = |T^* \wedge \tau - \hat{\mu}^{(-k)}_{\tau,n_1}(Z)| - |T^* \wedge \tau - \hat{\mu}_{\tau,n_1}(Z)|$,
- $p_k = \mathbb{P}[\delta_k(T^*, Z) \geq 0 \mid T^* \leq \tau]$.

Asymptotic confidence intervals and asymptotic distribution of the test statistic

Assume that

- $\mathbb{P}[T > \tau] > 0$,
- $C \perp\!\!\!\perp (T^*, Z)$.

Then,

$$\lim_{n_2 \to \infty} \mathbb{P}_{H_0}\left[\mathcal{T}_{n_2} > q_{1-\alpha}^{\mathcal{N}(0,1)}\right] \geq \alpha.$$

We also have

$$\lim_{n_2 \to \infty} \mathbb{P}\left(p_k \in \left[\frac{1}{1 - \hat{S}_{n_2}(\tau)} \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \Phi_k(T_i, Z_i)\hat{\omega}_i \pm \sqrt{\frac{\hat{\sigma}^2(\Phi_k)}{n_2}} q_{1-\alpha/2}^{\mathcal{N}(0,1)}\right]\right) \geq 1 - \alpha,$$

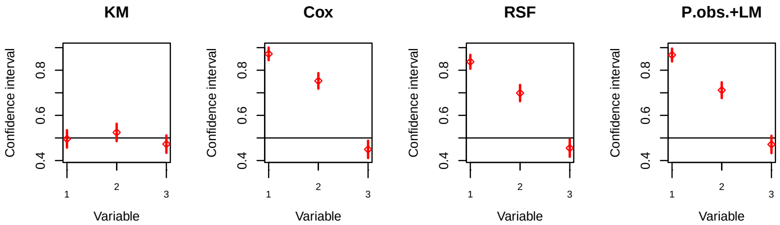
where $q_{1-\alpha}^{\mathcal{N}(0,1)}$ is the quantile of order $1 - \alpha$ of the $\mathcal{N}(0, 1)$ distribution.

Simulations - Scenario B

- $n_1 = 500$, D_{n_1} is fixed.
- True values of p_k obtained from 10^5 Monte-Carlo simulations :

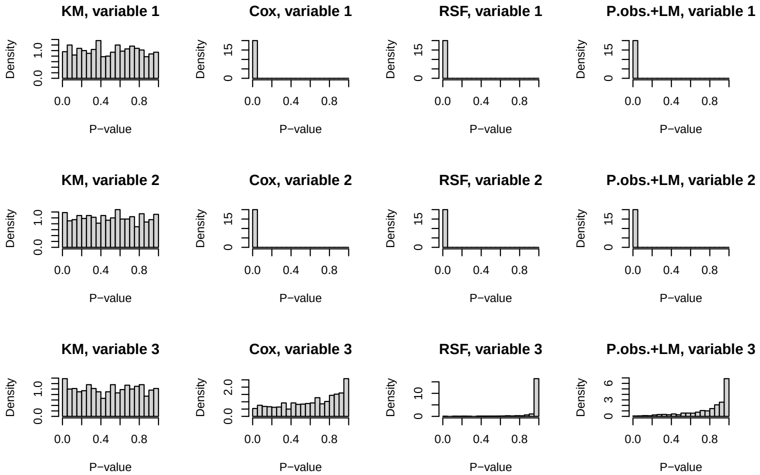
Learning models	p_1	p_2	p_3
Kaplan-Meier	0.50	0.50	0.50
Cox	0.87	0.79	0.49
Random Survival Forests	0.82	0.71	0.44
Pseudo-observations and linear model	0.84	0.70	0.46

- Confidence intervals at the 90% level on a single sample $n_2 = 500$:



Distribution of the p-values under (H_0) and (H_1)

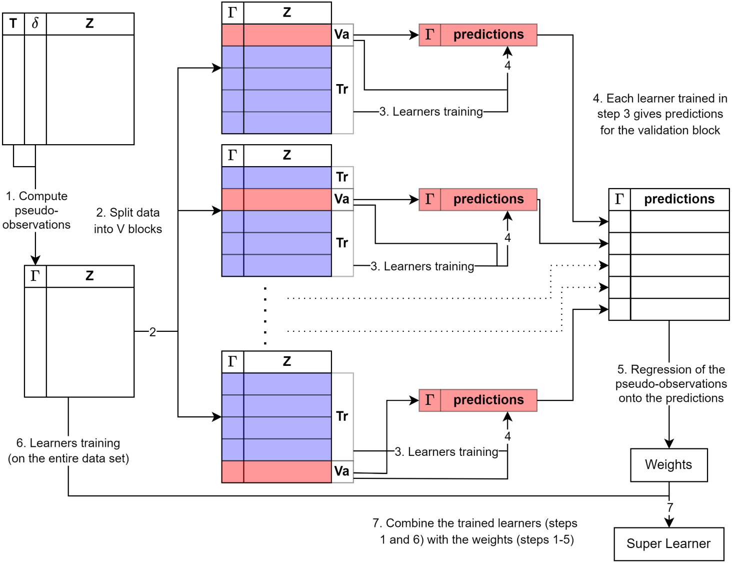
- $n_1 = 500$ (fixed), $n_2 = 500$, $1,000$ Monte-Carlo replications.
- For KM estimator : (H_0) is true for all variables.
- For all other algorithms : (H_0) is true for variable 3, (H_1) is true for variables 1 and 2.



Outline

1. Background in time to event analysis
2. Prediction of the (restricted) time based on pseudo-observations
3. Evaluation of the quality of predictions
4. Conformal prediction intervals
5. Variable importance
6. Pseudo-observations and the super-learner

Pseudo-Observations based Super Learner


T
δ
Z
1. Compute pseudo-observations
Γ
Z
2
2. Split data into V blocks
Γ
Z
Va
Tr
Γ
predictions
4
3. Learners training
Γ
Z
Tr
Va
Tr
Γ
predictions
4
3. Learners training
Γ
Z
Tr
Va
Γ
predictions
4
3. Learners training
4. Each learner trained in step 3 gives predictions for the validation block
Γ
predictions
5. Regression of the pseudo-observations onto the predictions
Weights
6. Learners training (on the entire data set)
7
7. Combine the trained learners (steps 1 and 6) with the weights (steps 1-5)
Super Learner

Split pseudo-observations

- The proof for the validity of the super learner relies on a Bernstein's inequality for **independent variables**.
  ⚠ but the pseudo-observations are not independent !

Split pseudo-observations

- The proof for the validity of the super learner relies on a Bernstein's inequality for **independent variables**.
  ⚠ but the pseudo-observations are not independent !
- We introduce the split pseudo-observations. $D_n = D_{n_1} \cup D_{n_2}$. $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Then for $i \in \mathcal{I}_2$, define :

  $$\Gamma_i(D_{n_1}) := (n_1 + 1) \int_0^\tau \hat{S}_{D_{n_1}, O_i}(t) dt - n_1 \int_0^\tau \hat{S}_{D_{n_1}}(t) dt,$$

  where
  - $\hat{S}_{D_{n_1}}$ is the Kaplan-Meier estimator computed on D_{n_1}
  - $\hat{S}_{D_{n_1}, O_i}$ is the Kaplan-Meier estimator computed on D_{n_1} and $O_i (\in D_{n_2})$.

  We have : $\Gamma_i(D_{n_1}) \perp\!\!\!\perp \Gamma_j(D_{n_1}) \mid D_{n_1}$, for $i \neq j$.

Split pseudo-observations

- The proof for the validity of the super learner relies on a Bernstein's inequality for **independent variables**.
  ⚠ but the pseudo-observations are not independent !
- We introduce the split pseudo-observations. $D_n = D_{n_1} \cup D_{n_2}$. $D_{n_1} = \{O_i : i \in \mathcal{I}_1\}$, $D_{n_2} = \{O_i : i \in \mathcal{I}_2\}$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Then for $i \in \mathcal{I}_2$, define :

$$\Gamma_i(D_{n_1}) := (n_1 + 1) \int_0^\tau \hat{S}_{D_{n_1}, O_i}(t)dt - n_1 \int_0^\tau \hat{S}_{D_{n_1}}(t)dt,$$

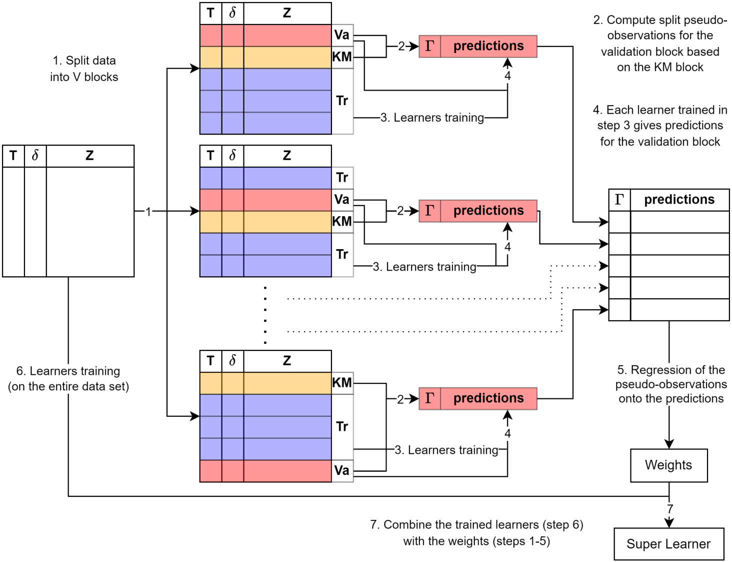
  where
  - $\hat{S}_{D_{n_1}}$ is the Kaplan-Meier estimator computed on D_{n_1}
  - $\hat{S}_{D_{n_1}, O_i}$ is the Kaplan-Meier estimator computed on D_{n_1} and $O_i (\in D_{n_2})$.

  We have : $\Gamma_i(D_{n_1}) \perp\!\!\!\perp \Gamma_j(D_{n_1}) \mid D_{n_1}$, for $i \neq j$.

We prove a similar result as in Jacobsen M., Martinussen T. (2016) :

$$\mathbb{E}[\Gamma_i(D_{n_1}) \mid Z_i, D_{n_1}] = \mathbb{E}[T_i^* \wedge \tau \mid Z_i] + o_{\mathbb{P}}(1).$$

Split pseudo-observations based Super Learner



1. Split data into V blocks

2. Compute split pseudo-observations for the validation block based on the KM block

3. Learners training

3. Learners training

3. Learners training

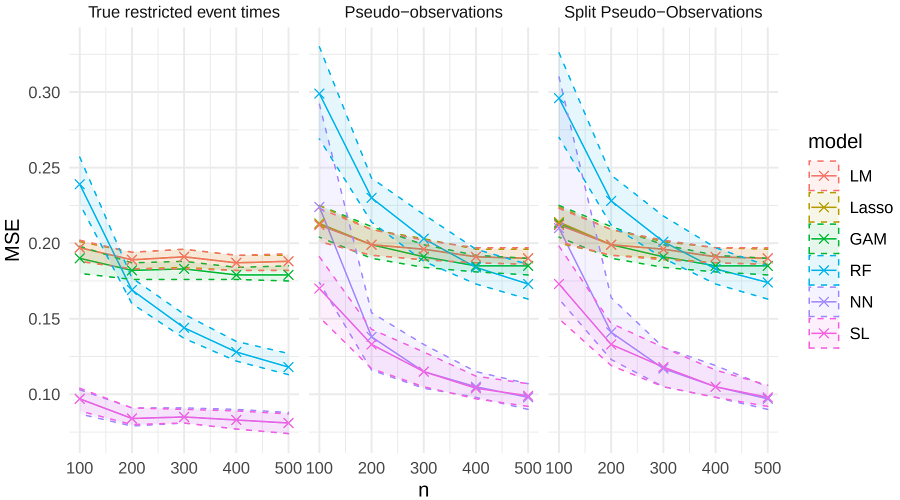
4. Each learner trained in step 3 gives predictions for the validation block

5. Regression of the pseudo-observations onto the predictions

6. Learners training (on the entire data set)

7. Combine the trained learners (step 6) with the weights (steps 1-5)

Standard VS split pseudo-observations based Super Learners : Scenario B



Simulation design

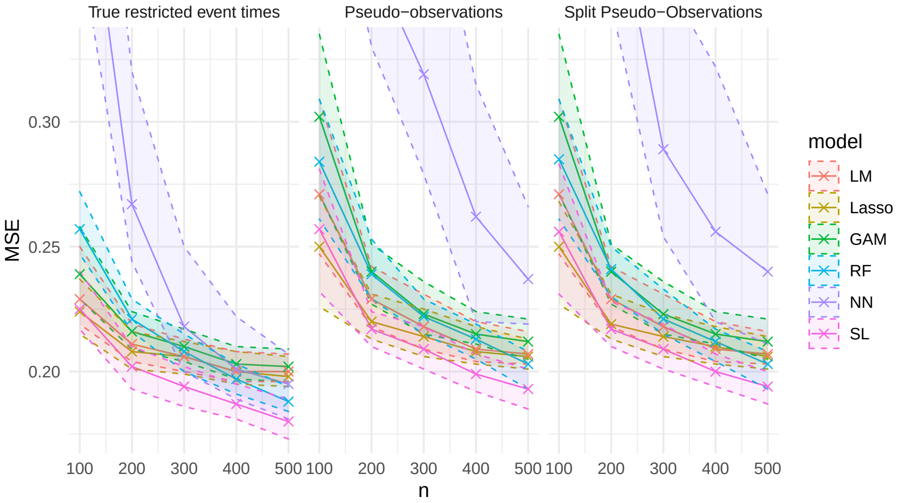
Scenario C : Cox model with interactions.

- $\lambda(t \mid Z) = \lambda_0(t) \exp(g(Z))$, $\lambda_0 \sim \mathcal{W}(\nu, \kappa)$, $\nu = 6, \kappa = 2$ and

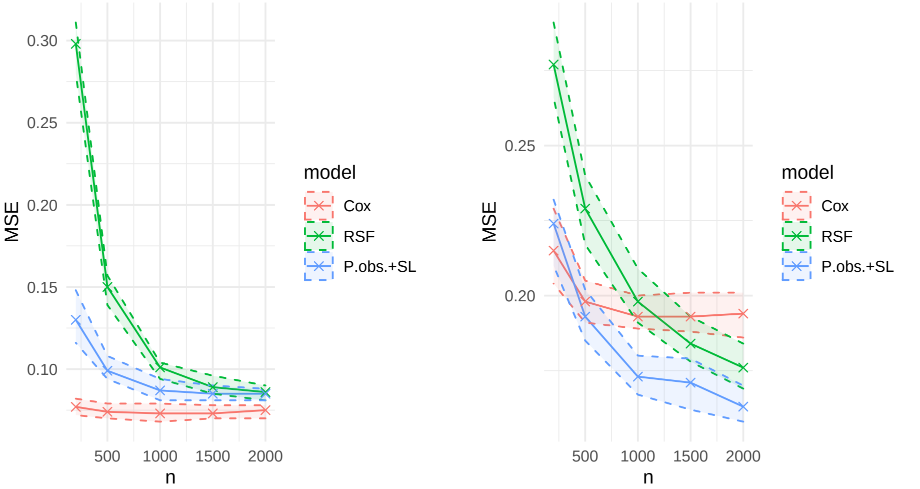
$$g(Z) = Z^3 - 3Z^5 + 2Z^1 Z^{10} + 4Z^2 Z^7 + 3Z^4 Z^5 - 5Z^6 Z^{10} + 3Z^8 Z^9 + Z^1 Z^4 - 2Z^6 Z^9 - 4Z^3 Z^4 - Z^7 Z^8,$$

- we simulate the covariates $Z = (Z^1, \dots, Z^{15})^\top$ with $Z^j \sim \mathcal{B}(0.4)$ for $j \in \{2, 4, 6, 9, 11, 12\}$ and $Z^j \sim U[0, 1]$, for $j \in \{1, 3, 5, 7, 8, 10, 13, 14, 15\}$. Only the first 10 covariates are associated with the event times.
- censoring rate : 47%.
- $\tau = 2.8$.

Standard VS Split pseudo-observations based Super Learners : Scenario C



Simulations : scenarios B (left) and C (right)



Discussion and future works

Assessment of the quality of prediction :

- The method is robust against model misspecification.
- All the tools are based on IPCW. Censoring distribution must be accurately modelled. Often it only depends on few or no covariates.
- Censoring assumption : for the test statistic, we impose $C \perp\!\!\!\perp Z$!
- Conformal intervals and test are split dependent $\hookrightarrow$ multi-splitting.
- Variable importance measure is dependent of the chosen model !

Cwiling A.,Perduca V. and Bouaziz O. *A Comprehensive Framework for Evaluating Time to Event Predictions using the Restricted Mean Survival Time*.

https://hal.science/hal-04143419v1/document

Discussion and future works

Assessment of the quality of prediction :

- The method is robust against model misspecification.
- All the tools are based on IPCW. Censoring distribution must be accurately modelled. Often it only depends on few or no covariates.
- Censoring assumption : for the test statistic, we impose $C \perp\!\!\!\perp Z$!
- Conformal intervals and test are split dependent $\hookrightarrow$ multi-splitting.
- Variable importance measure is dependent of the chosen model !

Cwiling A.,Perduca V. and Bouaziz O. *A Comprehensive Framework for Evaluating Time to Event Predictions using the Restricted Mean Survival Time*.
https://hal.science/hal-04143419v1/document

Pseudo-observations + super-learner :

- Split pseudo-observations are used to obtain the theoretical results for the super-learner.
- Split pseudo-observations and classical pseudo-observations are almost identical.
- The super learner automatically selects the best learner
  - among all the candidates (discrete super learner),
  - or provides the best combination of the candidates (continuous super learner).
- Extensions to other types of incomplete data : recurrent events, competing risks, left-truncation . . . .

Some references

[1] Per Kragh Andersen, John P. Klein, and Susanne Rosthoj. Generalised linear models for correlated pseudo-observations,with applications to multi-state models. *Biometrika*, 90 :15–27, 2003.

[2] Cyrus J. DiCiccio, Thomas J. DiCiccio, and Joseph P. Romano. Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166 :108865, 2020.

[3] Sandrine Dudoit and Mark J van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, 2(2) :131–154, 2005.

[4] Martin Jacobsen and Torben Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3) :845–862, 2016.

[5] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523) :1094–1111, 2018.

[6] Mark J. Van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1,25), 2007.

[7] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media New York, 2005.

Some references

[1] Per Kragh Andersen, John P. Klein, and Susanne Rosthoj. Generalised linear models for correlated pseudo-observations,with applications to multi-state models. *Biometrika*, 90 :15–27, 2003.

[2] Cyrus J. DiCiccio, Thomas J. DiCiccio, and Joseph P. Romano. Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166 :108865, 2020.

[3] Sandrine Dudoit and Mark J van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, 2(2) :131–154, 2005.

[4] Martin Jacobsen and Torben Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3) :845–862, 2016.

[5] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523) :1094–1111, 2018.

[6] Mark J. Van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1,25), 2007.

[7] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media New York, 2005.

Thank you for your attention

Additional slides

Theoretical results for the super learner

▷ Cross-validation : Observations are divided according to an independent random vector $B_n = (B_n(i) : i = 1, \dots, n) \in \{0, 1, 2\}^n$:

$$\begin{cases} \{O_i : i, B_n(i) = 0\}, \text{ size } n_0, \text{ empirical law } P^0_{Bn} : \text{ training set} \\ \{O_i : i, B_n(i) = 1\}, \text{ size } n_1, \text{ empirical law } P^1_{Bn} : \text{ KM set} \\ \{O_i : i, B_n(i) = 2\}, \text{ size } n_2, \text{ empirical law } P^2_{Bn} : \text{ validation set} \end{cases}$$

↪ Empirical law of the data : $P_{B_n} = \{P^0_{Bn}, P^1_{Bn}, P^2_{Bn}\}$

▷ $\Gamma_O(P^1_{Bn})$ = split pseudo-observation

▷ Estimators of the RMST : $\{\hat{\psi}_k : k = 1, \dots, K_n\}$

↪ Trained estimators : $\{\hat{\psi}_k(P^0_{Bn}) : k = 1, \dots, K_n\}$

▷ Quadratic loss : $L(\psi, O^*) = (T^* \wedge \tau - \psi(Z))^2$, where $O^* = (T^*, Z) \sim P$

↪ The RMST $\psi^*(Z) = \mathbb{E}[T^* \wedge \tau \mid Z]$ minimizes the risk $\mathbb{E}_{Bn} \int L(\psi, o) dP(o)$

▷ Cross-validated risk : $\tilde{\theta}(k) = \mathbb{E}_{B_n} \int L(\hat{\psi}_k(P^0_{Bn}), o) dP(o)$

↪ Cross-validated oracle selector : $\tilde{k} = \arg\min_{k \in \{1, \dots, K_n\}} \tilde{\theta}(k)$

▷ Cross-validated risk estimator : $\hat{\theta}^{\text{po}}_n(k) = \frac{1}{n_2} \sum_{i : B_n(i) = 2} (\Gamma_{O_i}(P^1_{Bn}) - \hat{\psi}_k(P^0_{Bn})(Z_i))^2$

↪ Cross-validated selector : $\hat{k}^{\text{po}} = \arg\min_{k \in \{1, \dots, K_n\}} \hat{\theta}^{\text{po}}_n(k)$

Theoretical results for the super learner

Theorem

Suppose that there exists $\tau \leq M < \infty$ such that

$$|\Gamma_O(P_{Bn}^1)| \leq M \text{ and } \sup_{Z \in \mathcal{Z}, \psi \in \Psi} |\psi(Z)| \leq M \text{ almost surely.}$$

Suppose that the censoring time C and the pair of variables (T^*, Z) are independent.

If $\log(K_n)/n_2 \underset{n \to \infty}{\longrightarrow} 0$, then

$$\mathbb{E}[\tilde{\theta}_n(\hat{k}^{\text{po}}) - \tilde{\theta}_n(\tilde{k})] \underset{n \to \infty}{\longrightarrow} 0,$$

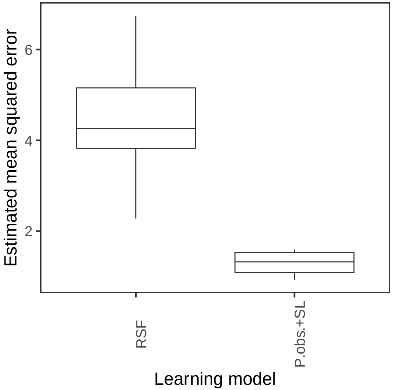
and

$$\tilde{\theta}_n(\hat{k}^{\text{po}}) - \tilde{\theta}_n(\tilde{k}) \underset{n \to \infty}{\longrightarrow} 0 \text{ in probability.}$$

Analysis of the maintenance dataset

- $1,000$ machines
- T^* : number of weeks in activity
- $T = T^* \wedge C$ ranges from 1 to 93 weeks
- 40% of censored data
- $\tau = 88$
- Five covariates : pressure (cont.), moisture (cont.), temperature (cont.), team (three levels), manufacturer (four levels)
- Algorithms :
  - RSF
  - pseudo-observations + super learner based on LM, LASSO, GAM, RF.

Analysis of the maintenance dataset



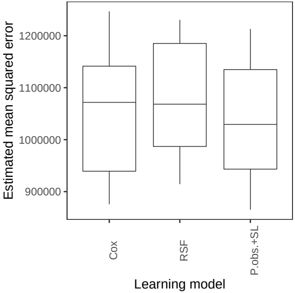
Variable	P-value RSF	P-value P.obs.+SL
Pressure	1	1
Moisture	0.257	1
Temperature	0.003	1
Team	0	0
Provider	0	0

TABLE – Tests for variable importance based on 40 multi-splits.

Analysis of the colon dataset

- 888 patients
- T^* : time elapsed from randomisation to minimum between recurrence and death.
- $T = T^* \wedge C$ ranges from 8 to 3329 days (9.12 years)
- 46% of censored data
- $\tau = 2672$
- Ten covariates : type of treatment administrated (three levels), sex (binary), age (in years), obstruction indicator of the colon by the tumor (binary), whether the colon was perforated or not (binary), whether or not it adhered to nearby organs (binary), number of lymph nodes with detectable cancer (integer value that ranges from 0 to 33), level of differentiation of the tumor (three levels), extent of local spread (four levels), whether the time from surgery to registration was short or long (binary).
- Algorithms :
  - Cox model
  - RSF
  - pseudo-observations + super learner based on LM, LASSO, GAM, RF.

Analysis of the colon dataset



Variable	P-value Cox	P-value RSF	P-value P.obs.+SL
Treatment	0	0.084	0.002
Sex	0.981	0.724	0.622
Age	1	1	1
Obstruction	1	1	1
Perforation	1	1	1
Adherence	1	1	1
Nodes	1	0.637	0.662
Differentiation	1	1	1
Spread	0	0.063	0.002
Surgery	1	1	1

TABLE – Tests for variable importance based on 40 multi-splits.