Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring

Olivier Bouaziz

MAP5 (CNRS 8145), Université Paris Cité

International Biometric Conference 2022

Outline

Introduction (Andersen, Klein, Rosthøj, 2003)

- $T_1^*, \dots, T_n^*$ i.i.d. variables of interest. $Z_1, \dots, Z_n$ i.i.d. covariates (in $\mathbb{R}^p$).
- $\theta := \mathbb{E}[h(T_i^*)]$, h a known function. $\theta_{(l)} := \mathbb{E}[h(T_l^* \mid Z_l)]$. For example :
  - $h(x) = \mathbb{1}_{x \geq t}$ gives $\theta = S(t)$ and $\theta_{(l)} = S(t \mid Z_l)$ (survival).
  - $h(x) = x \wedge \tau$ gives $\theta = \mathbb{E}(T^* \wedge \tau)$ and $\theta_{(l)} = \mathbb{E}(T_l^* \wedge \tau \mid Z_l)$ (RMST).
- Suppose there exists g known and invertible such that : $g(\theta_{(l)}) = Z_l^\top \beta$. We aim at estimating β.

Introduction (Andersen, Klein, Rosthøj, 2003)

- $T_1^*, \dots, T_n^*$ i.i.d. variables of interest. $Z_1, \dots, Z_n$ i.i.d. covariates (in $\mathbb{R}^p$).
- $\theta := \mathbb{E}[h(T_i^*)]$, h a known function. $\theta_{(l)} := \mathbb{E}[h(T_l^* \mid Z_l)]$. For example :
  - $h(x) = \mathbb{1}_{x \geq t}$ gives $\theta = S(t)$ and $\theta_{(l)} = S(t \mid Z_l)$ (survival).
  - $h(x) = x \wedge \tau$ gives $\theta = \mathbb{E}(T^* \wedge \tau)$ and $\theta_{(l)} = \mathbb{E}(T_l^* \wedge \tau \mid Z_l)$ (RMST).
- Suppose there exists g known and invertible such that : $g(\theta_{(l)}) = Z_l^\top \beta$. We aim at estimating β.
- Usually, T_i^* are not observed. Observations : $X_1, \dots, X_n$ i.i.d. Construct : $\hat{\theta} := \hat{\theta}(X_1, \dots, X_n)$.
- The l^{th} pseudo-observation is defined as :

$$\hat{\theta}_{(l)} = n\hat{\theta} - (n-1)\hat{\theta}^{(-l)},$$

where $\hat{\theta}^{(-l)}$ is the jackknife estimator of θ ($\hat{\theta}^{(-l)} := \hat{\theta}(X_1, \dots, X_{l-1}, X_{l+1}, \dots X_n)$).

Introduction (Andersen, Klein, Rosthøj, 2003)

- $T_1^*, \ldots, T_n^*$ i.i.d. variables of interest. $Z_1, \ldots, Z_n$ i.i.d. covariates (in $\mathbb{R}^p$).
- $\theta := \mathbb{E}[h(T_i^*)]$, h a known function. $\theta_{(l)} := \mathbb{E}[h(T_l^* \mid Z_l)]$. For example :
  - $h(x) = \mathbb{1}_{x \geq t}$ gives $\theta = S(t)$ and $\theta_{(l)} = S(t \mid Z_l)$ (survival).
  - $h(x) = x \wedge \tau$ gives $\theta = \mathbb{E}(T^* \wedge \tau)$ and $\theta_{(l)} = \mathbb{E}(T_l^* \wedge \tau \mid Z_l)$ (RMST).
- Suppose there exists g known and invertible such that : $g(\theta_{(l)}) = Z_l^\top \beta$.
  We aim at estimating β.
- Usually, T_i^* are not observed. Observations : $X_1, \ldots, X_n$ i.i.d. Construct :
  $\hat{\theta} := \hat{\theta}(X_1, \ldots, X_n)$.
- The l^{th} pseudo-observation is defined as :

$$\hat{\theta}_{(l)} = n\hat{\theta} - (n-1)\hat{\theta}^{(-l)},$$

  where $\hat{\theta}^{(-l)}$ is the jackknife estimator of θ ($\hat{\theta}^{(-l)} := \hat{\theta}(X_1, \ldots, X_{l-1}, X_{l+1}, \ldots X_n)$).
- Estimate β by treating $\hat{\theta}_{(l)}$ as the response in a regression model.
  Use Generalised Estimating Equation (GEE) : `geese` function from `geepack` R package.

Settings

Right-censoring

- Observations : $X_i = (T_i, \Delta_i)$, $Z_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$,

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

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

Interval-censoring (mixed case)

- Observations : $X_i = (L_i, R_i)$, $Z_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$,
  - $0 = L_i < R_i < +\infty$ for left-censored observations,
  - $0 < L_i < R_i < +\infty$ for interval-censored observations,
  - $0 < L_i < R_i = +\infty$ for right-censored observations.
  - $0 < L_i = R_i < +\infty$ for exact observations.

- Assumption : $\begin{cases} \mathbb{P}(T^* \in [L, R]) = 1 \\ \mathbb{P}(T^* \leq t \mid L = l, R = r, Z) = \mathbb{P}(T^* \leq t \mid l \leq T^* \leq r, Z) \end{cases}$

Theoretical result for right-censored data

Graw, Gerds, Schumacher 2009 ; Jacobsen, Martinussen 2016 ; Overgaard, Parner, Pedersen 2017.

Proposition

Consider $\theta = S(t)$, $\theta_{(l)} = S(t \mid Z_l)$. Let $\hat{S}$ be the Kaplan-Meier estimator. Then, for all $t \in [0, \tau]$,

$$\hat{\theta}_{(l)} = n\hat{S}(t) - (n-1)\hat{S}^{(-l)}(t) = S(t) + \dot{\psi}(X_l, t) + O_{\mathbb{P}}(n^{-1/2}),$$

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

$$\dot{\psi}(X_l, t) = -S(t) \int_0^t \frac{dM_l(u)}{H(u)}.$$

Moreover,

$$S(t) + \mathbb{E}(\dot{\psi}(X_l, t) \mid Z_l) = S(t \mid Z_l)$$

Notations : $H(\cdot) = \mathbb{P}(T \geq \cdot)$, $M_l(\cdot) = \mathbb{1}_{T_l \leq \cdot, \Delta_l = 1} - \int_0^{\cdot} \mathbb{1}_{T_l \geq u} d\Lambda(u)$, Λ is the cumulative hazard function, $S(\cdot \mid Z)$ is the conditional survival function.

Approximation of the pseudo-values

- Right-censoring (RC).
  - We can approximate the pseudo-observations by :

$$\hat{S}(t) - \hat{S}(t) \int_0^t \frac{d\hat{M}_l(u)}{\hat{H}(u)},$$

    $\hat{S}$ is the Kaplan-Meier estimator, $\hat{H}(t) = \sum_{i=1}^n \mathbb{1}_{T_i \geq t}/n$, $\hat{M}_l(\cdot) = \mathbb{1}_{T_l \leq \cdot, \Delta_l = 1} - \int_0^{\cdot} \mathbb{1}_{T_l \geq u} d\hat{\Lambda}(u)$.
  - Time reduction : the pseudo-observations can be approximated without implementing the jackknife !

- Interval-censoring (IC).
  - S is estimated by the non-parametric estimator (**Turnbull 1976**).
  - Slow rate of convergence : $O_{\mathbb{P}}(n^{-1/3})$ or $O_{\mathbb{P}}((n \log n)^{-1/3})$ (**Groeneboom, Wellner 1992**).
  - $\sqrt{n}$ rate of convergence in **Huang 1999** where it is further assumed that $\#$exact obs.$/n$ tends to a positive constant. However no closed form for the Von Mises expansions.
  - Approximations based on Von-Mises expansions are not applicable for deriving approximated pseudo-observations !

Outline

Parametric modelling and maximum likelihood estimators

- We assume the distribution of T^* (f^*, S, Λ) depends on a parameter $\alpha_0 \in \mathbb{R}^d$.
- Let the observed sample $X_1, \ldots, X_n$ i.i.d. $\sim f(\cdot; \alpha_0)$.

Parametric modelling and maximum likelihood estimators

- We assume the distribution of T^* (f^*, S, Λ) depends on a parameter $\alpha_0 \in \mathbb{R}^d$.
- Let the observed sample $X_1, \ldots, X_n$ i.i.d. $\sim f(\cdot; \alpha_0)$.
- Define the maximum likelihood estimator of α_0 by :

$$\hat{\alpha} = \arg\max_{\alpha} \ell_n(\alpha) := \sum_{i=1}^{n} \log f(X_i; \alpha).$$

Example : for interval-censored data $X_i = (L_i, R_i)$ and

$$f(X_i; \alpha) = (S(L_i; \alpha) - S(R_i; \alpha)) I(L_i \neq R_i) + f^*(L_i; \alpha) I(L_i = R_i),$$

with the slight abuse of notation $S(R_i; \alpha) = 0$ if $R_i = \infty$.

Parametric modelling and maximum likelihood estimators

- We assume the distribution of T^* (f^*, S, Λ) depends on a parameter $\alpha_0 \in \mathbb{R}^d$.
- Let the observed sample $X_1, \dots, X_n$ i.i.d. $\sim f(\cdot; \alpha_0)$.
- Define the maximum likelihood estimator of α_0 by :

$$\hat{\alpha} = \arg\max_{\alpha} \ell_n(\alpha) := \sum_{i=1}^{n} \log f(X_i; \alpha).$$

Example : for interval-censored data $X_i = (L_i, R_i)$ and

$$f(X_i; \alpha) = (S(L_i; \alpha) - S(R_i; \alpha))I(L_i \neq R_i) + f^*(L_i; \alpha)I(L_i = R_i),$$

with the slight abuse of notation $S(R_i; \alpha) = 0$ if $R_i = \infty$.

- Define the jackknife estimators $\hat{\alpha}^{(-l)}$, for $l = 1, \dots, n$.
- The pseudo-observations for the survival function are :

$$nS(t; \hat{\alpha}) - (n-1)S(t; \hat{\alpha}^{(-l)}), \quad l = 1, \dots, n.$$

- Implementation based on the jackknife in **Sabathé, Andersen, Helmer, Gerds, Jacqmin-Gadda, Joly, 2019**, for IC data with the Weibull and spline models for f^*.

Parametric modelling and maximum likelihood estimators

Proposition 1

Under standard regularity conditions for maximum likelihood theory, we have :

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\alpha_0) - S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top I^{-1}\nabla\log f(X_l;\alpha_0) + o_{\mathbb{P}}(1),$$

with $I = -\mathbb{E}[\nabla^2 \log f(X_i;\alpha_0)]$ is the Fisher information.

- We can estimate the pseudo-observations by :

$$S(t;\hat{\alpha}) - S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})^\top \hat{I}^{-1}\nabla\log f(X_l;\hat{\alpha}).$$

- Time reduction : no need to implement the jackknife !

Parametric modelling and maximum likelihood estimators

Proposition 1

Under standard regularity conditions for maximum likelihood theory, we have :

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = \underbrace{S(t;\alpha_0) - S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top I^{-1}\nabla \log f(X_l;\alpha_0)}_{\varphi(X_l,t;\alpha_0)} + o_{\mathbb{P}}(1),$$

with $I = -\mathbb{E}[\nabla^2 \log f(X_i;\alpha_0)]$ is the Fisher information.

- We can estimate the pseudo-observations by :

$$S(t;\hat{\alpha}) - S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})^\top \hat{I}^{-1}\nabla \log f(X_l;\hat{\alpha}).$$

- Time reduction : no need to implement the jackknife !

Proposition 2

Assume that $T^* \mid Z = z$ follows the same parametric distribution as T^* but with different parameters. We then have $\mathbb{E}(\varphi(X_l, t;\alpha_0) \mid Z_l = z) = S(t \mid z) + R_z(t)$.

- In general $R_z(t) \neq 0$!

Outline

Likelihood maximisation for IC data - two possible approaches

We use the piecewise constant hazard (pch) model. Let $0 = c_0 < c_1 < \cdots < c_K = +\infty$. The model is defined as :

$$\lambda(t;\alpha) = \sum_{k=1}^{K} \alpha_k \mathbb{1}_{c_{k-1}<t\leq c_k}, \quad \alpha = (\alpha_1,\ldots,\alpha_K)^\top \in (R_+^*)^K.$$

Observations : $X_i = (L_i, R_i)$, $i = 1,\ldots,n$.

1. Direct maximisation of the observed log-likelihood ($S(+\infty) = 0$) :

$$\ell_n(\alpha) = \sum_{i=1}^{n} \log\left(S(L_i;\alpha) - S(R_i;\alpha)\right) I(L_i \neq R_i) + f^*(L_i;\alpha) I(L_i = R_i).$$

   - No explicit maximiser for the pch model.
   - Requires the Newton-Raphson algorithm but the Hessian is not diagonal and of full rank !
   - Inversion of the Hessian might be intractable for K large !

2. Maximisation of the complete likelihood $\ell_n^{\text{comp}}(\alpha) = \sum_{i=1}^{n} \log f^*(T_i^*;\alpha)$ based on the EM algorithm.
   - Treat the true event time T^* as a missing variable.
   - The M-step is explicit !

We choose option 2 !

Outline

Estimation of the RMST with interval-censored data

Scenario 1 : Restricted Mean Survival Time (RMST) model.

$$\mathbb{E}(T_i^* \wedge \tau \mid Z_i) = \beta_{01} + \beta_{02} Z_{i,1}(1 - Z_{i,2}) + \beta_{03} Z_{i,2}(1 - Z_{i,1}) + \beta_{04} Z_{i,1} Z_{i,2},$$

- $\beta_0 = (4.98, 0.14, 0.14, 0.27)^\top$, $Z_{i,1},\ Z_{i,2} \underset{\text{i.i.d.}}{\sim} \mathcal{B}(0.25)$, $\tau = 6$ (54.2% quantile of T^*)
- LC 14.6%, IC 52.07%, RC 33.33%.
- Average length of IC intervals ≈ 1.34

Scenario 2 : Linear model ($\tau = \infty$).

$$\mathbb{E}(T_i^* \mid Z_i) = \beta_{00} + \beta_{01} Z_i,$$

- $\beta = (6, 4)^\top$, $Z_i \sim \mathcal{U}[0, 2]$
- LC 10%, IC 64%, RC 26%.
- Average length of IC intervals ≈ 3.5

Parametric model for T^* : we use the piecewise-constant hazard model.

Simulation results for interval-censored data

- First scenario. Four cuts for the pch model : $c_1 = 4, c_2 = 5, c_3 = 6, c_4 = 7$.

n	Jackknife Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Approximated formula Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
200	-0.188	0.231	0.090	6.219 min	-0.187	0.232	0.089	0.221 s
	0.026	0.320	0.103		0.027	0.319	0.102	
	0.045	0.325	0.107		0.045	0.323	0.106	
	0.096	0.296	0.097		0.094	0.295	0.096	
500	-0.187	0.152	0.058	23.589 min	-0.187	0.152	0.058	0.664 s
	0.048	0.208	0.046		0.048	0.208	0.046	
	0.038	0.209	0.045		0.038	0.209	0.045	
	0.080	0.192	0.043		0.080	0.192	0.043	
1,000	-0.189	0.106	0.047	87.717 min	-0.189	0.106	0.047	1.349 s
	0.043	0.137	0.021		0.043	0.137	0.021	
	0.043	0.145	0.023		0.043	0.145	0.023	
	0.074	0.138	0.025		0.074	0.138	0.025	

- Second scenario. Five cuts for the pch model : $c_1 = 6, c_2 = 8, c_3 = 10, c_4 = 12, c_5 = 14$.

n	Jackknife Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Approximated formula Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
500	-0.130	0.202	0.058	24.461 min	-0.114	0.186	0.047	0.552 s
	0.094	0.174	0.039		0.083	0.153	0.030	
1,000	-0.113	0.113	0.025	68.998 min	-0.110	0.113	0.025	1.092 s
	0.080	0.102	0.017		0.078	0.102	0.016	

Summary, comments and perspectives

- Proposed approximations are extremely accurate (as compared to jackknife) and fast.
- Fast approximations for other quantities of interest are possible : cumulative incidence functions (competing risks), state probabilities (multi-state models) etc.
- Pseudo-values for parametric models are theoretically not valid.

$$nS(t; \hat{\alpha}) - (n-1)S(t; \hat{\alpha}^{(-l)}) = \cdot + o_{\mathbb{P}}(1),$$

  with $\mathbb{E}[\cdot \mid Z_l = z] = S(t \mid Z_l = z) + R_z$.
- However, the remainder term seems to be small in practice (see also **Sabathé, Andersen, Helmer, Gerds, Jacqmin-Gadda, Joly, 2019** with the Weibull or spline models).
- There is a special property for the pch model : when the number of cuts tends to infinity, the remainder term tends to zero !
- The pch model can be combined with penalised data-driven methods for choosing the number and location of the cuts.
  See **Bouaziz, Lauridsen, Nuel, 2021** and use my pchsurv GitHub package.
- There exists another fast approximation for the Kaplan-Meier estimator based on the infinitesimal jackknife in the `pseudo` function, `survival` package.

My GitHub package : https://github.com/obouaziz/FastPseudo

Bibliography

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

[2] Olivier Bouaziz, Eva Lauridsen, and Grégory Nuel. Regression modelling of interval-censored data based on the adaptive-ridge procedure. *Journal of Applied Statistics*, pages 1–25, In press.

[3] Frederik Graw, Thomas A Gerds, and Martin Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2) :241–255, 2009.

[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] Martin Nygård Johansen, Søren Lundbye-Christensen, and Erik Thorlund Parner. Regression models using parametric pseudo-observations. *Statistics in Medicine*, 39(22) :2949–2961, 2020.

[6] Camille Sabathe, Per K Andersen, Catherine Helmer, Thomas A Gerds, Hélène Jacqmin-Gadda, and Pierre Joly. Regression analysis in an illness-death model with interval-censored data : A pseudo-value approach. *Statistical methods in medical research*, 29(3) :752–764, 2020.

My preprint : Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring. https://arxiv.org/abs/2109.02959

Thank you for your attention