The negative log-likelihood loss and cross-validation with censored data

Anders Munch

PhD Student, Section of Biostatistics
University of Copenhagen

June 14, 2022

Outline

Model and hyper-parameter selection for survival models

The least false model in the presence of censoring

Hold-out samples and survival model estimators

Selecting a model from a collection of candidate models

Consider estimation of the parameter

$$\theta(P) := \operatorname*{argmin}_{f \in \mathcal{F}} P[L(f, \cdot)], \quad \text{where} \quad P[g] := \int_{\mathcal{O}} g(o) P(\mathrm{d}o),$$

for some loss function $L \colon \mathcal{F} \times \mathcal{O} \to \mathbb{R}_+$.

Selecting a model from a collection of candidate models

Consider estimation of the parameter

$$\theta(P) := \operatorname*{argmin}_{f \in \mathcal{F}} P[L(f, \cdot)], \quad \text{where} \quad P[g] := \int_{\mathcal{O}} g(o) P(\mathrm{d}o),$$

for some loss function $L \colon \mathcal{F} \times \mathcal{O} \to \mathbb{R}_+$.

Maximum likelihood estimator (MLE)

If $\mathcal{F}$ is a collection of densities on $\mathcal{O}$ and $L(f, O) := -\log(f(O))$, then $\theta(\hat{\mathbb{P}}_n)$ is the MLE for the model $\mathcal{F}$, where $\hat{\mathbb{P}}_n$ denotes the empirical measure.

Selecting a model from a collection of candidate models

Consider estimation of the parameter

$$\theta(P) := \operatorname*{argmin}_{f\in\mathcal{F}} P[L(f,\cdot)], \quad \text{where} \quad P[g] := \int_{\mathcal{O}} g(o)P(\mathrm{d}o),$$

for some loss function $L\colon \mathcal{F}\times\mathcal{O}\to\mathbb{R}_+$.

Maximum likelihood estimator (MLE)

If $\mathcal{F}$ is a collection of densities on $\mathcal{O}$ and $L(f,O) := -\log(f(O))$, then $\theta(\hat{\mathbb{P}}_n)$ is the MLE for the model $\mathcal{F}$, where $\hat{\mathbb{P}}_n$ denotes the empirical measure.

Hyper-parameter selection

For estimation in high-dimensional settings we often introduce a regularization parameter λ (e.g., LASSO, kernel smoothing). To select a value for λ we would typically split the data $\mathcal{D}_n = \{O_1,\dots,O_n\}$ randomly in two, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$, and calculate

$$\operatorname*{argmin}_{\lambda\in\Lambda} \hat{\mathbb{P}}_n^2[L(\hat{f}_\lambda^1,\cdot)],$$

where $\hat{\mathbb{P}}_n^2$ denotes the empirical measure based on the sample $\mathcal{D}_n^2$, and $\hat{f}_\lambda^1$ denotes an estimator calculated on $\mathcal{D}_n^1$ with regularization parameter λ.

A loss function for survival data

$O = (\tilde{T}, \Delta, X) \sim P \in \mathcal{P}$ Oberved data with $\mathcal{O} = \mathbb{R}_+ \times \{0, 1\} \times \mathbb{R}^p$.

$(T, X) \sim Q \in \mathcal{Q}$ The distribution Q (or a feature of it) is of interest.

Assuming coarsening at random [Gill et al., 1997] we can write

$$\mathcal{P} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\},$$

where $\mathcal{G}$ denotes a collection of conditional distributions for the censoring mechanism, and the likelihood factorizes as

$$\ell(P_{Q,G}, O) = \ell_F(Q, O) \cdot \ell_{\mathcal{C}}(G, O),$$

with

$$\ell_F(Q, O) := q(\tilde{T} \mid X)^{\Delta} \bar{Q}(\tilde{T} \mid X)^{1-\Delta} m(X),$$

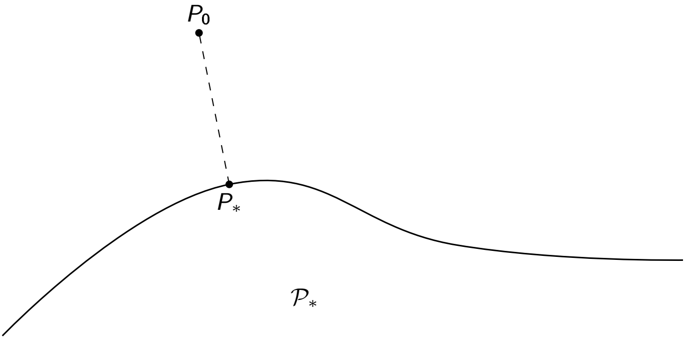
where q and $\bar{Q}$ are the conditional density and survivor function, respectively, and m the marginal distribution of X.

Natural to use the negative partial log-likelihood $-\log \ell_F$ as loss function, or even only the first part concerning the conditional distribution of T given X.

Kullback-Leibler divergence and partial likelihoods

Maximum likelihood estimation is connected to minimizing the Kullback-Leibler divergence and gives an interpretation of the MLE under mis-specified models.

$$D_{\mathrm{KL}}(P_0 \,||\, P) := P_0\left[\log \frac{p_0}{p}\right], \quad \text{where} \quad P_0 = p_0 \cdot \mu, P = p \cdot \mu.$$



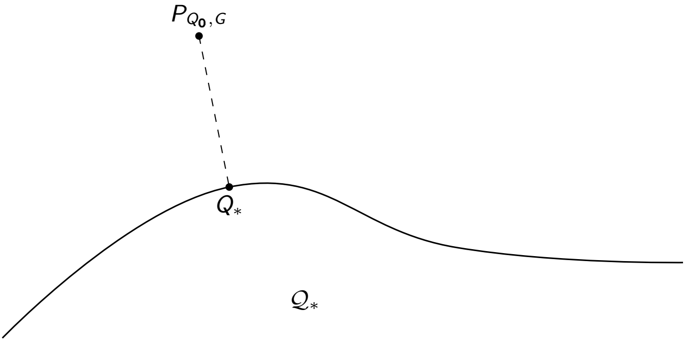
Kullback-Leibler divergence and partial likelihoods

Maximum likelihood estimation is connected to minimizing the Kullback-Leibler divergence and gives an interpretation of the MLE under mis-specified models.

$$D_{\mathrm{KL}}(P_0 \,||\, P) := P_0\left[\log \frac{p_0}{p}\right], \quad \text{where} \quad P_0 = p_0 \cdot \mu, P = p \cdot \mu.$$

For a partial likelihood we are minimizing

$$Q \longmapsto D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q,G}), \quad \text{with} \quad Q \in \mathcal{Q}_*.$$



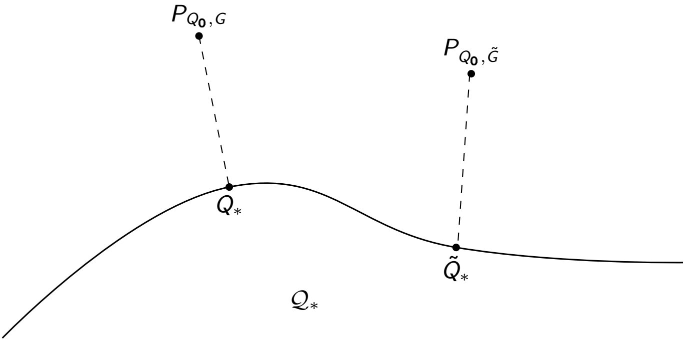
Kullback-Leibler divergence and partial likelihoods

Maximum likelihood estimation is connected to minimizing the Kullback-Leibler divergence and gives an interpretation of the MLE under mis-specified models.

$$D_{\mathrm{KL}}(P_0 \,||\, P) := P_0\left[\log \frac{p_0}{p}\right], \quad \text{where} \quad P_0 = p_0 \cdot \mu, P = p \cdot \mu.$$

For a partial likelihood we are minimizing

$$Q \longmapsto D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q,G}), \quad \text{with} \quad Q \in \mathcal{Q}_*.$$



Least false model depends on the censoring distribution

For any value $G \in \mathcal{G}$ we have that $D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q_0,G}) = 0$, so the correct model Q_0 is ranked better than any other model independently of $G \in \mathcal{G}$. However, if $Q_0 \notin \mathcal{Q}_*$ the minimizer might depend on the value of G.[1]

[1] This is mentioned in Whitney et al. [2019] and van der Laan and Dudoit [2003], and a similar phenomenon is well studied for the Cox model [Struthers and Kalbfleisch, 1986, Hjort, 1992, Fine, 2002].

Least false model depends on the censoring distribution

For any value $G \in \mathcal{G}$ we have that $D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q_0,G}) = 0$, so the correct model Q_0 is ranked better than any other model independently of $G \in \mathcal{G}$. However, if $Q_0 \notin \mathcal{Q}_*$ the minimizer might depend on the value of G.[1]

For the simple survival case with no baseline covariates, we have the following result stating that for a mis-specified model Q we can alway find an alternative model $\tilde{Q}$ that is ranked better under one censoring regime but worse under another.

Let Q_0 and G be given together with some $Q \neq Q_0$. Then (under regularity conditions) we can find $\tilde{Q}$ and $\tilde{G}$ such that

$$D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{Q,G}) < D_{\mathrm{KL}}(P_{Q_0,G} \,||\, P_{\tilde{Q},G}),$$

and

$$D_{\mathrm{KL}}(P_{Q_0,\tilde{G}} \,||\, P_{Q,\tilde{G}}) > D_{\mathrm{KL}}(P_{Q_0,\tilde{G}} \,||\, P_{\tilde{Q},\tilde{G}}).$$

[1] This is mentioned in Whitney et al. [2019] and van der Laan and Dudoit [2003], and a similar phenomenon is well studied for the Cox model [Struthers and Kalbfleisch, 1986, Hjort, 1992, Fine, 2002].

Sketch of proof

- Divide $(0, \tau)$ into $(0, \tau_0)$ and $[\tau_0, \tau)$, where $(0, \tau)$ is the support of T.
- Construct $\tilde{Q}$ such that it performs better than Q on $(0, \tau_0)$ but worse on (τ_0, τ) under the censoring regime G.
- Construct $\tilde{G}$ such that observations on (τ_0, τ) are less likely than under G.

Sketch of proof

- Divide $(0, \tau)$ into $(0, \tau_0)$ and $[\tau_0, \tau)$, where $(0, \tau)$ is the support of T.
- Construct $\tilde{Q}$ such that it performs better than Q on $(0, \tau_0)$ but worse on (τ_0, τ) under the censoring regime G.
- Construct $\tilde{G}$ such that observations on (τ_0, τ) are less likely than under G.

Whether the alternative model $\tilde{Q}$ can be constructed such that $\tilde{Q} \in \mathcal{Q}_*$ for some model class $\mathcal{Q}_*$ will depend on the model class and on Q_0 and $\mathcal{G}$.

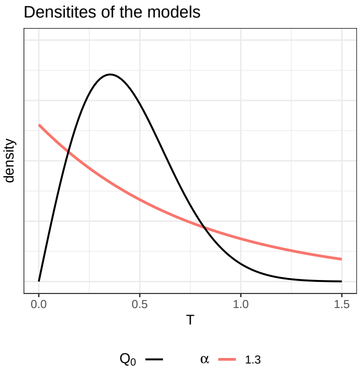
A simple example with mis-specified survival models

Assume the data generating distribution given by

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma),$$

and consider the four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3,\ 1.5,\ 1.8,\ 2\}.$$



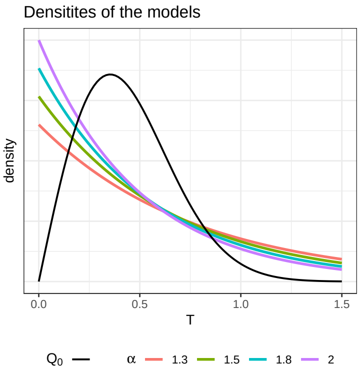
A simple example with mis-specified survival models

Assume the data generating distribution given by

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma),$$

and consider the four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3,\ 1.5,\ 1.8,\ 2\}.$$



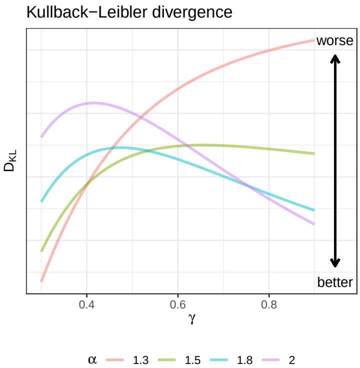
A simple example with mis-specified survival models

Assume the data generating distribution given by

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma),$$

and consider the four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3,\ 1.5,\ 1.8,\ 2\}.$$



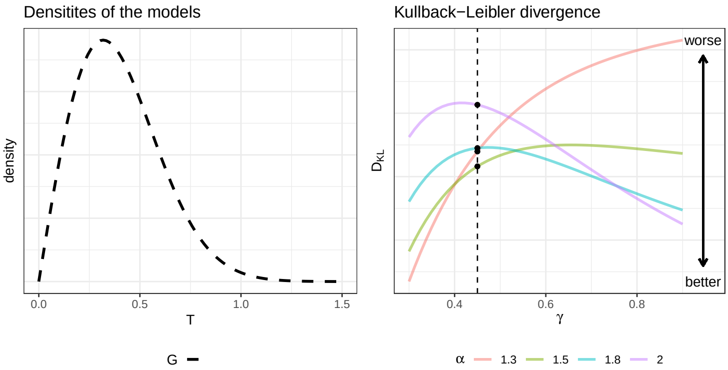
A simple example with mis-specified survival models

Assume the data generating distribution given by

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma),$$

and consider the four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3,\, 1.5,\, 1.8,\, 2\}.$$



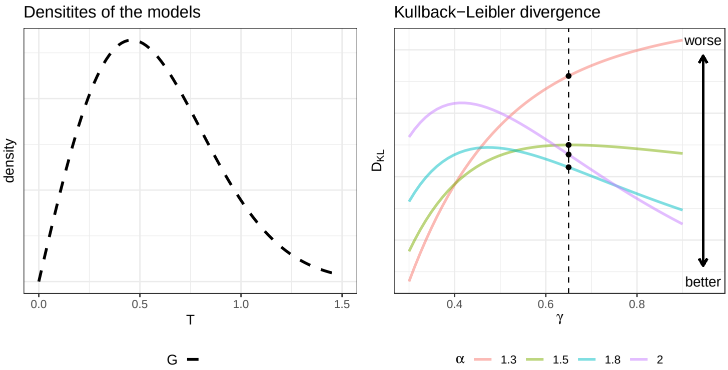
A simple example with mis-specified survival models

Assume the data generating distribution given by

$$Q_0 = \text{Weibull}(2, 0.5), \quad \text{and} \quad G_\gamma = \text{Weibull}(2, \gamma),$$

and consider the four candidate models indexed by α,

$$Q_\alpha = \text{Exp}(\alpha), \quad \text{with} \quad \alpha \in \{1.3,\, 1.5,\, 1.8,\, 2\}.$$



Survival curve estimators evaluated on hold-out samples

Consider the problem of selecting a hyper-parameter or model using cross-validation. We split the data $\mathcal{D}_n = \{O_1, \ldots, O_n\}$ in two, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$.

On split $\mathcal{D}_n^1$ Fit models $\{\hat{f}_\lambda : \lambda \in \Lambda\}$ or $\{\hat{f}_1, \hat{f}_2, \ldots \hat{f}_k\}$.

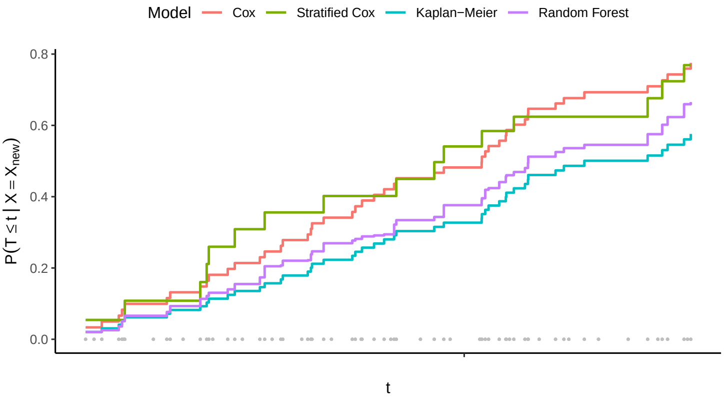
On split $\mathcal{D}_n^2$ Evalute the performance using a loss function L.

Survival curve estimators evaluated on hold-out samples

Consider the problem of selecting a hyper-parameter or model using cross-validation. We split the data $\mathcal{D}_n = \{O_1, \dots, O_n\}$ in two, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$.

On split $\mathcal{D}_n^1$ Fit models $\{\hat{f}_\lambda : \lambda \in \Lambda\}$ or $\{\hat{f}_1, \hat{f}_2, \dots \hat{f}_k\}$.

On split $\mathcal{D}_n^2$ Evalute the performance using a loss function L.

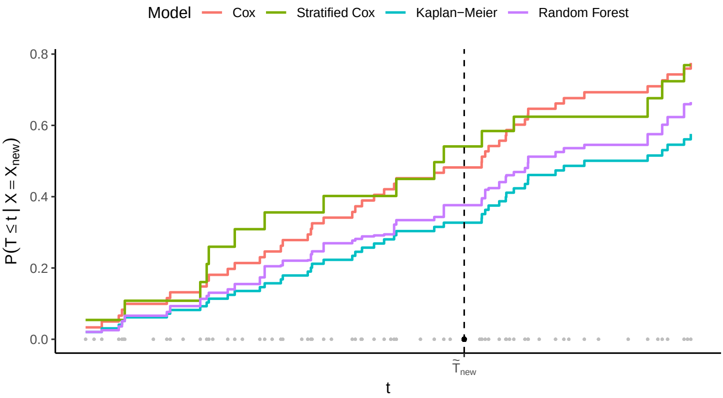


Survival curve estimators evaluated on hold-out samples

Consider the problem of selecting a hyper-parameter or model using cross-validation. We split the data $\mathcal{D}_n = \{O_1, \ldots, O_n\}$ in two, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$.

On split $\mathcal{D}_n^1$ Fit models $\{\hat{f}_\lambda : \lambda \in \Lambda\}$ or $\{\hat{f}_1, \hat{f}_2, \ldots \hat{f}_k\}$.

On split $\mathcal{D}_n^2$ Evalute the performance using a loss function L.



Taking the censoring distribution into account

To alliviate these problems problems we can reweight the observed outcome or the loss function to account/adjust for the censoring:

- Inverse probability of censoring weighted loss functions [van der Laan and Dudoit, 2003]. For instance, weighted negative log-likelihood or (integrated) Brier score.
- Pseudo-values [Andersen et al., 2003, Mogensen and Gerds, 2013].
- Censoring unbiased transformations [Fan and Gijbels, 1996, Steingrimsson et al., 2019].

These approaches are particularly attractive when we are willing to assume that the censoring does not depend on the baseline covariates.

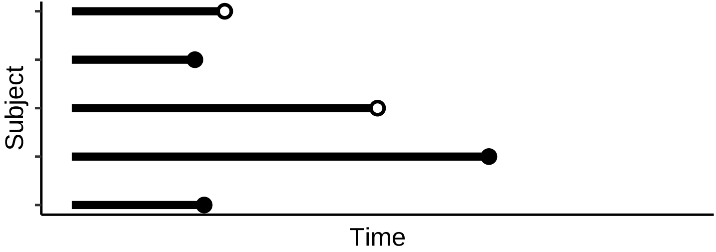
Modeling the censoring distribution

If we are not sure that the censoring is independent we need to model the dependence on the covariates.

Modeling the censoring distribution

If we are not sure that the censoring is independent we need to model the dependence on the covariates.

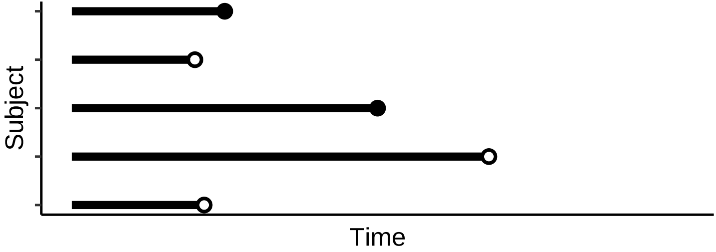
An (infinite?) loop



Modeling the censoring distribution

If we are not sure that the censoring is independent we need to model the dependence on the covariates.

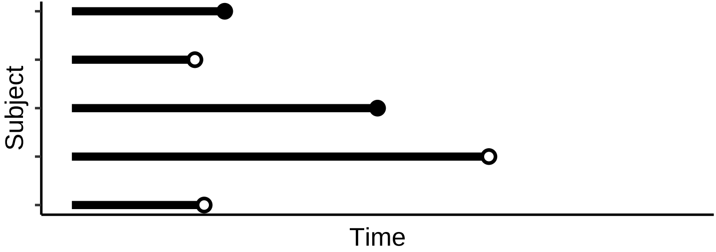
An (infinite?) loop



Modeling the censoring distribution

If we are not sure that the censoring is independent we need to model the dependence on the covariates.

An (infinite?) loop



Inverse-weighted survival games [Han et al., 2021]: Iterate the estimation until convergence (hopefully).

Conclusion

How should we do cross-validation for general survival models?

- Using the negative partial log-likelihood is problematic
  - $\rightarrow$ The least false model is not well-defined (without reference to the censoring regime)
  - $\rightarrow$ For many standard survival estimators, we cannot use it on hold-out samples
- Using loss functions designed to measure the loss for the model of interest is challenging in the presence of complicated censoring
  - $\rightarrow$ We need a model for the censoring ...
    - $\rightarrow$ We need a model for the outcome ...

Conclusion

How should we do cross-validation for general survival models?

- Using the negative partial log-likelihood is problematic
  - $\rightarrow$ The least false model is not well-defined (without reference to the censoring regime)
  - $\rightarrow$ For many standard survival estimators, we cannot use it on hold-out samples
- Using loss functions designed to measure the loss for the model of interest is challenging in the presence of complicated censoring
  - $\rightarrow$ We need a model for the censoring ...
    - $\rightarrow$ We need a model for the outcome ...

Questions, comments, suggestions?

Thank you!

References

P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Routledge, 1996.

J. Fine. Comparing nonnested cox models. *Biometrika*, 89(3):635–648, 2002.

R. D. Gill, M. J. Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

X. Han, M. Goldstein, A. Puli, T. Wies, A. Perotte, and R. Ranganath. Inverse-weighted survival games. *Advances in Neural Information Processing Systems*, 34, 2021.

N. L. Hjort. On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, pages 355–387, 1992.

U. B. Mogensen and T. A. Gerds. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*, 32(18):3102–3114, 2013.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383, 2019.

C. A. Struthers and J. D. Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73 (2):363–369, 1986.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.