November 22, 2023

Summary of some simulation results

Anders Munch

1 Super learners

In all simulation studies, we compare five super learners, which are listed below. To evaluate performance super learner, we use an independent data set of 10.000 uncensored samples and calculate the integrated Brier score in this data set. All results are based on 500 simulated data sets.

1. The state learner (referred to as `statelearner`).
2. The super learner proposed in [Westling et al., 2021] (referred to as `survSL`).
3. A super learner based on the estimated integrated Brier score, where the censoring mechanism is estimated with the Kaplan-Meier estimator (referred to as `ipcw_km`).
4. A super learner based on the estimated integrated Brier score, where the censoring mechanism is estimated with a Cox model that includes all available covariates as main effects (referred to as `ipcw_cox`).
5. The (discrete) oracle super learner, which picks the model that minimizes the Brier score in the independent data of 10.000 uncensored samples (referred to as `oracle`).

Only the super learners 1., 2., and 3. provides estimates of the censoring distribution.

2 Zelefsky based simulation

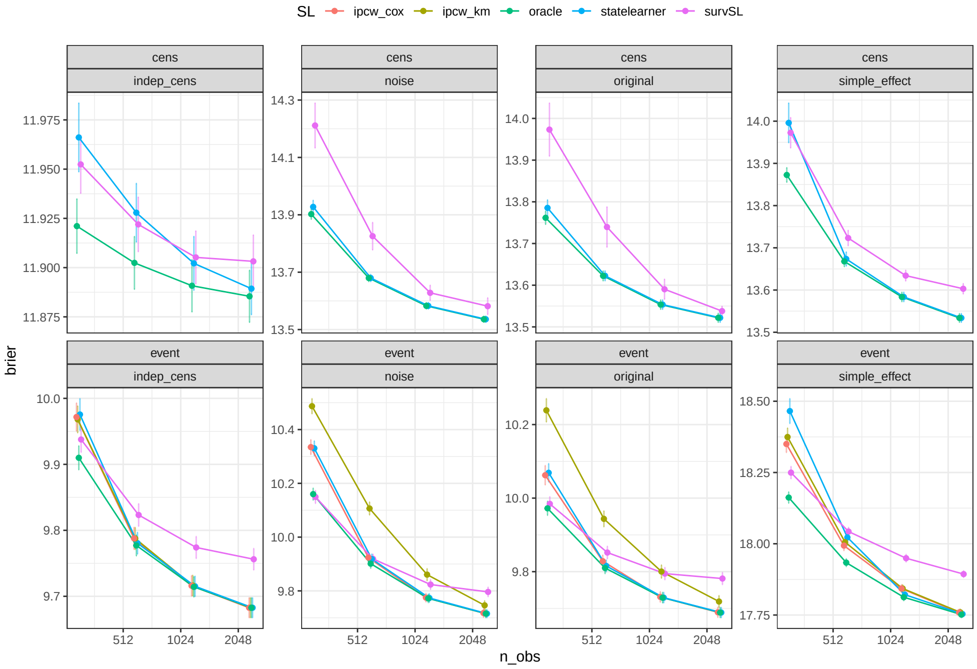
We generate data in four different ways:

1. Data as generated in [Gerds et al., 2013] (referred to as `original`).
2. As in 1., but where censoring in completely independent of covariates (referred to as `indep_cens`).
3. Same censoring mechanism as in 1., but where the outcome depend only on one of the covariates (referred to as `simple_effect`)
4. As in 1., but we add 5 independent standard Guassian covariates with no effect on neither outcome nor censoring (referred to as `noise`).

2.1 Kaplan-Meier, Cox, and random forest

In this setting, we include the following learners in all libraries:

- The Kaplan-Meier estimator
- A Cox model with main effects
- A random forest based on 50 trees

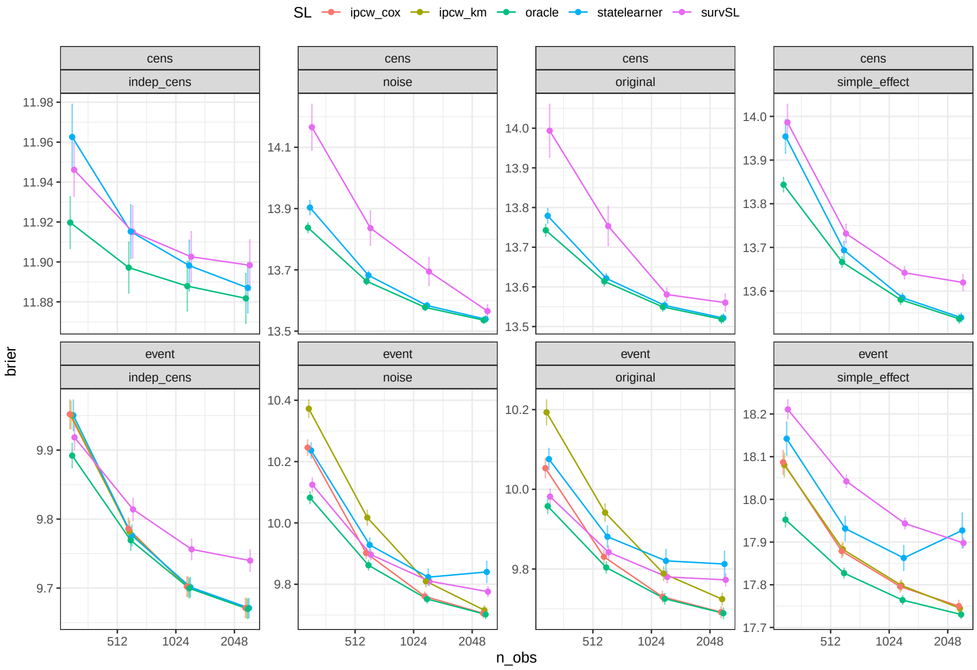


Zelefsky simulation setting using library consisting of Kaplan-Meier, Cox model, and random forests

2.2 Add LASSO

In this setting we add a learner to all libraries, so that all libraries include the learners:

- The Kaplan-Meier estimator
- A Cox model with main effects
- A random forest based on 50 trees
- A penalized Cox model with main effects, where the $\|\cdot\|_1$ penalty (LASSO) is used and the penalty parameter is selected using cross-validation based on Cox' partial likelihood



Zelefsky simulation setting including LASSO into the library

3 Effect of number of variables

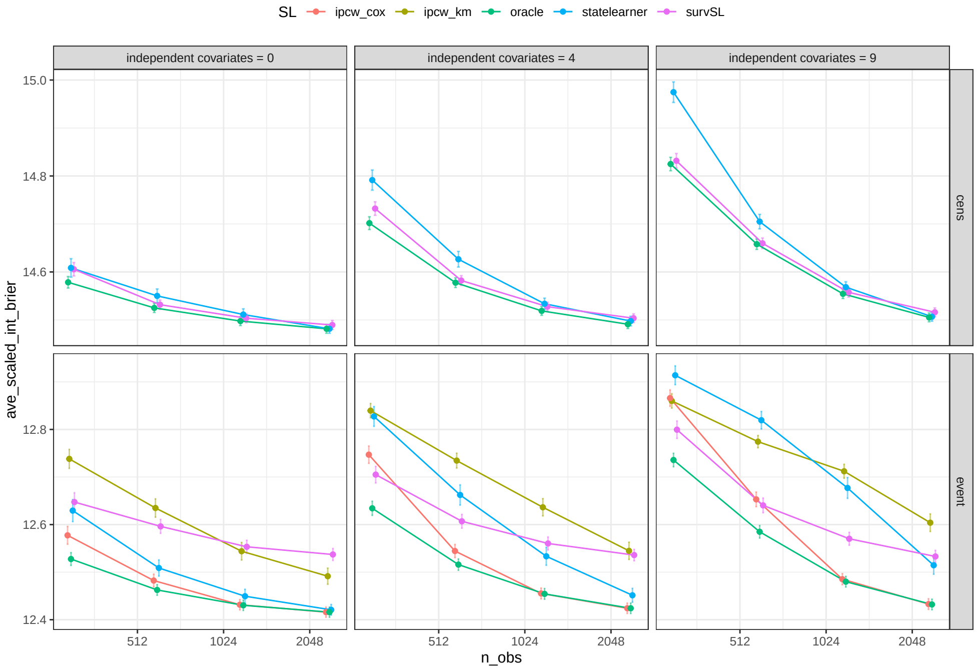
We generate data in three different way:

1. Outcome and censoring depends on one binary covariate (X_1). Another continuous covariate (X_2) that is correlated with X_1 is generated.
2. Same as in 1., but we also add 4 independent Gaussian covariates ($X_3, \dots X_6$).
3. Same as in 1., but we also add 9 independent Gaussian covariates ($X_3, \dots X_{11}$).

3.1 Kaplan-Meier, Cox, and random forest

In this setting, we include the following learners in all libraries:

- The Kaplan-Meier estimator
- A Cox model with main effects
- A random forest based on 50 trees

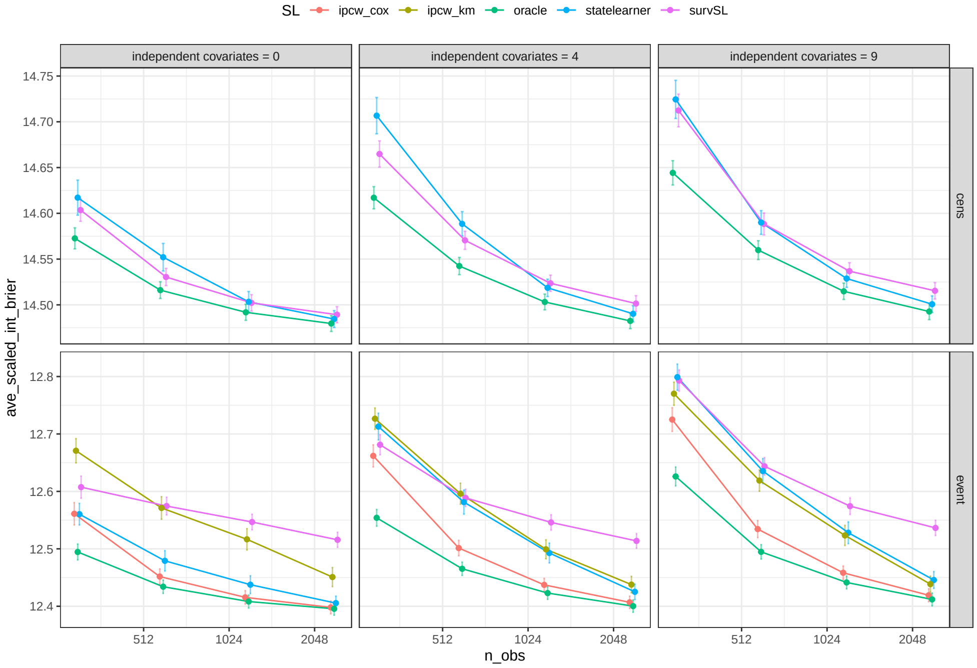


Zelefsky simulation setting using library consisting of Kaplan-Meier, Cox model, and random forests

3.2 Add LASSO

In this setting we add a learner to all libraries, so that all libraries include the learners:

- The Kaplan-Meier estimator
- A Cox model with main effects
- A random forest based on 50 trees
- A penalized Cox model with main effects, where the $\|\cdot\|_1$ penalty (LASSO) is used and the penalty parameter is selected using cross-validation based on Cox' partial likelihood



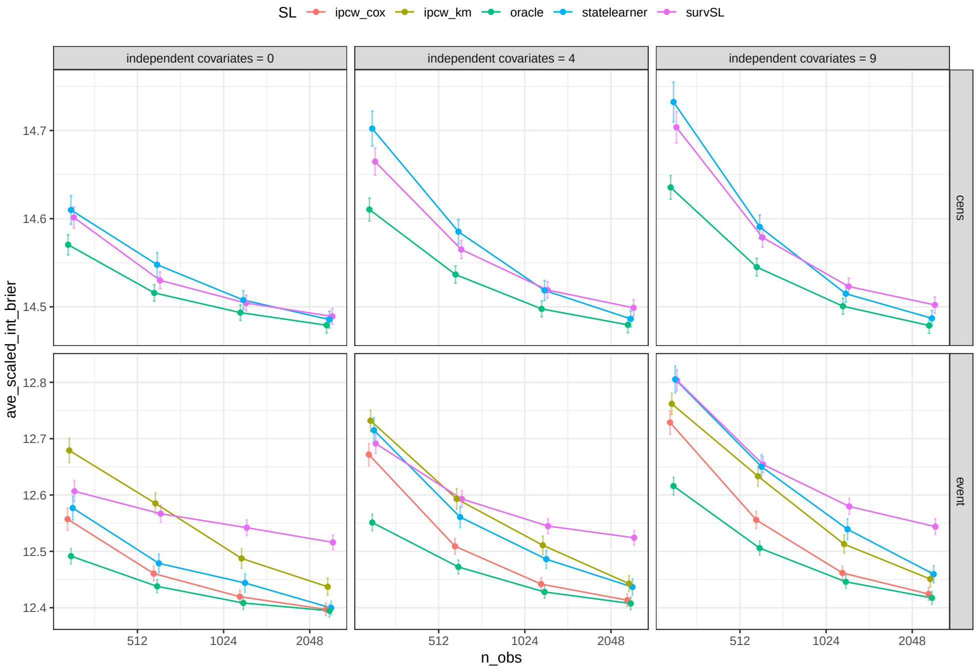
Zelefsky simulation setting including LASSO into the library

3.3 Add LASSO and elastic net

In this setting we add a learner to all libraries, so that all libraries include the learners:

- The Kaplan-Meier estimator
- A Cox model with main effects
- A random forest based on 50 trees
- A penalized Cox model with main effects, where the $\|\cdot\|_1$ penalty (LASSO) is used and the penalty parameter is selected using cross-validation based on Cox' partial likelihood

- A penalized Cox model with main effects, where the $\|\cdot\|_1 + \|\cdot\|_2$ penalty (elastic net) is used and the penalty parameter is selected using cross-validation based on Cox' partial likelihood



Zelefsky simulation setting including LASSO into the library

4 References

T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013.

T. Westling, A. Luedtke, P. Gilbert, and M. Carone. Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*, 2021.