Structural Limitations of PoSI in Cox Regression

Lena Schemet[1], Sarah Friedrich-Welz[1,2]

[1]Faculty of Mathematics, University of Augsburg, Germany

[2]Center for Advanced Analytics and Predictive Sciences (CAAPS), University of Augsburg, Germany

UNIVERSITÄT AUGSBURG
Mathematisch-Naturwissenschaftlich-Technische Fakultät

Motivation

- Post-selection inference (PoSI) is well established for linear regression, where it provides valid inference after arbitrary model selection by controlling coverage over a prespecified class of submodels.
- This perspective can be transferred to Cox regression, but its practical implications in the survival setting are not yet well understood.
- We study two Cox-PoSI constructions:
  - standardized max-$|Z|$,
  - score-based supremum.
- Main finding: both procedures are asymptotically valid, but both necessarily produce larger critical values than full-model simultaneous inference.

Cox-PoSI framework

Setup

- Survival data $(T_i, \Delta_i, X_i)_{i=1}^n$ with independent censoring.
- Candidate submodels

$$m \subseteq \{1, \dots, p\}, \qquad m \in \mathcal{M}_k.$$

- For each submodel m, let $\hat{\beta}_m$ denote the Cox partial likelihood estimator.
- Global index set

$$\mathcal{I} = \{(m, j) : m \in \mathcal{M}_k,\ j \in m\}.$$

Two Cox-PoSI statistics

- Standardized version:

$$T_n^Z = \max_{(m,j)\in\mathcal{I}} |Z_{m,j,n}|.$$

- Score-based version:

$$T_n^S = \max_{(m,j)\in\mathcal{I}} |S_{m,j,n}|.$$

- Both converge to maxima of centered Gaussian limits.
- Hence simultaneous confidence statements remain valid after arbitrary selection within $\mathcal{M}_k$.

Structural inflation

- If the full model is contained in $\mathcal{M}_k$, then its coordinates are part of the PoSI index set.
- Therefore,

$$T_n^Z \geq T_{\text{full}}^Z, \qquad T_n^S \geq T_{\text{full}}^S,$$

and thus

$$K_{1-\alpha}^Z \geq K_{1-\alpha,\text{full}}^Z, \qquad K_{1-\alpha}^S \geq K_{1-\alpha,\text{full}}^S.$$

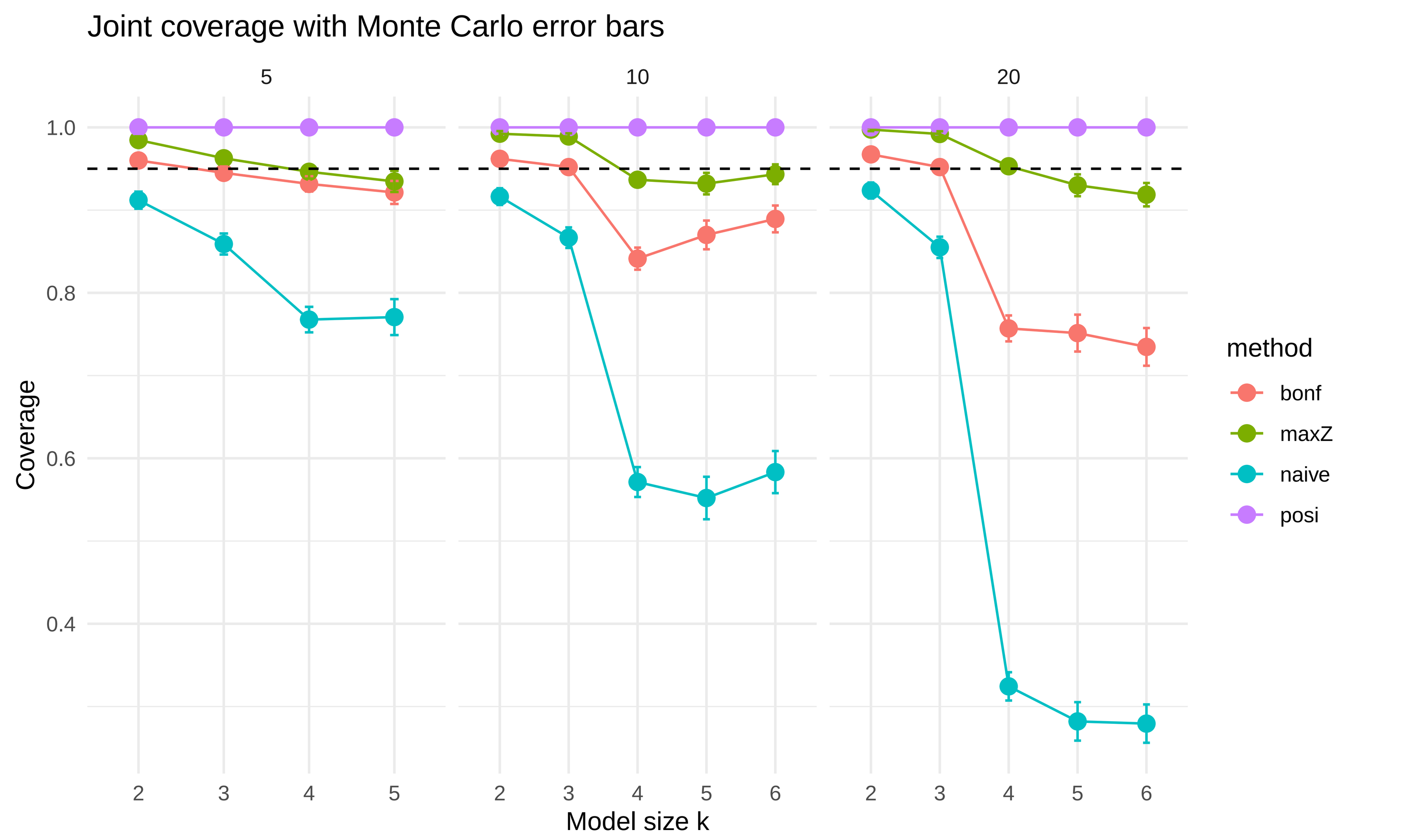
- Inflation is structural:
  - not a numerical artifact,
  - not a tuning issue,
  - caused by maximizing over a larger index set.
- Empirically, the growth is close to $\sqrt{2\log|\mathcal{I}|}$.

Key message:

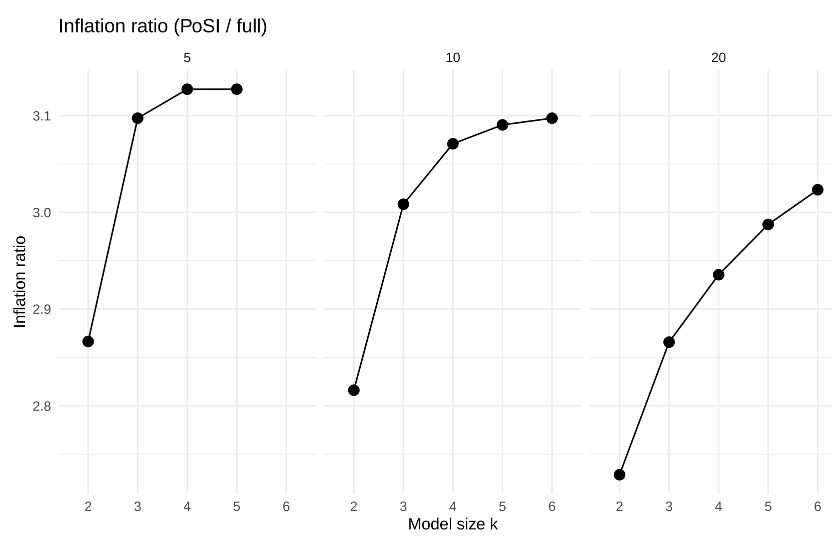
PoSI controls inference over the whole candidate model class, not only over the selected model.

Simulation study

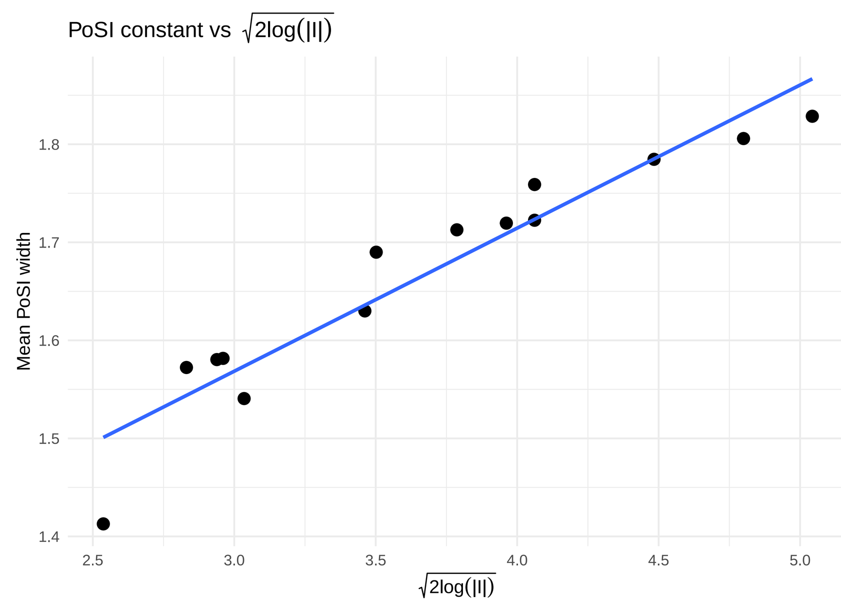
- Right-censored data with $n = 200$, $p = 10$, three non-zero coefficients, and independent censoring.
- Candidate model class $\mathcal{M}_k$: all submodels up to size k.
- Naive post-selection intervals under-cover and deteriorate as k increases.
- Cox-PoSI maintains joint coverage, but at the price of substantially larger critical values.



Joint coverage over $\mathcal{M}_k$.



$K_{\text{PoSI}}/K_{\text{full}}$ vs. model size.



PoSI constant vs. $\sqrt{2\log|\mathcal{I}|}$.

Real data example: TCGA-BRCA

Data and selected model

- Clinical survival data from the TCGA Breast Invasive Carcinoma cohort.
- Candidate covariates: age and stage-derived indicators.
- Candidate model class: all submodels up to size $k = 3$.
- AIC selected a model with age, age^2, and stage IV.

Interpretation

- Naive and full-model simultaneous intervals remain relatively informative.
- PoSI intervals are much wider.
- The application confirms the structural cost already seen in simulation.

Critical constants

Method	Constant
Naive	1.96
Full max-$\lvert Z\rvert$	2.84
PoSI	11.44

Stage IV hazard ratio interval

Method	95% interval
Naive	[6.05, 19.31]
Full max-$\lvert Z\rvert$	[4.65, 25.09]
PoSI	[0.36, 319.94]

Main practical message

PoSI is valid, but expensive.

Even for a moderate model class, the PoSI critical constant is much larger than the naive and full-model alternatives.
This produces intervals that are theoretically valid but may become practically too wide to be useful.
For stage IV, the PoSI interval expands to

$$[0.36,\ 319.94],$$

illustrating the loss of informativeness very clearly.

- Moderate model class
- Strong inflation of critical constants
- Strong widening of intervals
- Same qualitative pattern as in simulation

Take-home messages

- Cox-PoSI yields valid simultaneous inference over finite model classes.
- This gives valid post-selection inference after arbitrary selection within $\mathcal{M}_k$.
- The price is unavoidable structural inflation of critical values.
- In practice, Cox-PoSI is best viewed as a benchmark for post-selection uncertainty rather than a routine default.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, **10**, 1100–1120.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, **41**, 802–837.

Kuchibhotla, A.K., Brown, L.D., Buja, A., George, E.I. and Zhao, L. (2018). Valid post-selection inference in assumption-lean linear regression. *arXiv:1806.04119*.

lena.schemet@uni-a.de