The Journal of Open Source Software

aorsf: An R package for supervised learning using the oblique random survival forest

Byron C. Jaeger[1], **Sawyer Welden**[1], **Kristin Lenoir**[1], **and Nicholas M Pajewski**[1]

1 Wake Forest University School of Medicine

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The random survival forest (RSF) is a supervised learning method for right-censored time-to-event data that combines predictions from a large set of survival decision trees (Ishwaran, Kogalur, Blackstone, & Lauer, 2008). Similar to random forests for classification and regression (Breiman, 2001), trees in the RSF are de-correlated by using random subsets of training data to grow each tree and considering a random subset of predictor variables at each non-terminal node. Decision trees in the RSF may be axis based or oblique. Axis-based trees split non-terminal nodes using individual predictor variables, whereas oblique trees use a linear combination of variables. Although using oblique instead of axis based decision trees to grow the RSF improves its Brier score and concordance index in prediction tasks, the computational overhead of fitting an oblique RSF may be substantially higher than an axis based RSF (Jaeger et al., 2019), making it difficult to use the oblique RSF in applied settings.

`aorsf` is an R package designed to develop and compute predictions with oblique RSFs efficiently. Extensions of `aorsf`'s core features are supported by allowing users to supply their own function to identify a linear combination of inputs when growing oblique survival trees. The target audience includes both **analysts** aiming to develop an accurate risk prediction model (e.g., see Segar et al. (2021)) and **researchers** who want to conduct experiments comparing different techniques for identifying linear combinations of predictor variables (e.g., see Katuwal, Suganthan, & Zhang (2020)). Key features of `aorsf` include computational efficiency compared to existing software, extensive unit and integration testing that ensure cross-platform consistency and reproducibility, and user-friendly documentation paired with an application programming interface that facilitates proper usage of the core algorithms.

Existing software

The `obliqueRF` and `RLT` R packages support oblique classification and regression random forests but not oblique RSFs. The `ranger` and `randomForestSRC` packages support axis based RSFs but not oblique RSFs. The `obliqueRSF` R package fits oblique RSFs, but its computational inefficiency is a barrier in applied settings. `aorsf` is designed for computational efficiency and flexibility, allowing users to grow oblique RSFs using Newton Raphson scoring (the default), penalized Cox regression (the method used by `obliqueRSF`), or a user-defined function to identify linear combinations of variables.

Newton Raphson scoring

The default routine for creating linear combinations of predictor variables in `aorsf` applies Newton Raphson scoring to the partial likelihood function of the Cox regression model.

aorsf uses the same approach as the survival package to complete this estimation procedure efficiently. Full details on the steps involved have been made available by Therneau (2022). Briefly, a vector of estimated regression coefficients, $\hat{\beta}$, is updated in each step of the procedure based on its first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^{k} + U(\hat{\beta} = \hat{\beta}^{k})\, H^{-1}(\hat{\beta} = \hat{\beta}^{k})$$

While it is standard practice in statistical modeling to iteratively update $\hat{\beta}$ until a convergence threshold is met, the default approach in aorsf only completes one iteration (see ?orsf_control_cph). The rationale for this approach is based on two points. First, while completing more iterations reduces bias in the regression coefficients, general benchmarking experiments have found it does not improve the discrimination or Brier score of the oblique RSF (Jaeger et al., 2022). Second, computing U and H requires computation and exponentiation of the vector $X\hat{\beta}$, where X is the matrix of predictor values, but these steps can be skipped on the first iteration if an initial value of $\hat{\beta} = 0$ is chosen, allowing for a reduction in required computation.

Benchmarking

The increased efficiency of aorsf versus obliqueRSF results from improved memory management and Newton Raphson scoring. In general benchmarks of prediction accuracy and computational efficiency, aorsf has matched or exceeded the prediction accuracy of obliqueRSF while running two orders of magnitude faster (Jaeger et al., 2022).

Acknowledgements

The development of this software was supported by the Center for Biomedical Informatics at Wake Forest University School of Medicine and the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant Award Number UL1TR001420. Dr. Pajewski was additionally supported by grant award P30 AG021332 from the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH

References

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, *2*(3), 841–860.

Jaeger, B., Long, D. L., Long, D. M., Sims, M., Szychowski, J. M., Min, Y.-I., Mcclure, L. A., et al. (2019). Oblique random survival forests. *The Annals of Applied Statistics*, *13*(3), 1847–1883.

Jaeger, B., Welden, S., Lenoir, K., Speiser, J. L., Segar, M. W., Pandey, A., & Pajewski, N. M. (2022). Accelerated and interpretable oblique random survival forests. arXiv. doi:10.48550/ARXIV.2208.01129

Katuwal, R., Suganthan, P. N., & Zhang, L. (2020). Heterogeneous oblique random forest. *Pattern Recognition*, *99*, 107078.

Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., Butler, J., et al. (2021). Development and validation of machine learning–based race-specific models to predict 10-year risk of heart failure: A multicohort analysis. *Circulation*, *143*(24), 2370–2383.

Therneau, T. (2022, April). Survival package source code documentation. Retrieved from https://github.com/therneau/survival/blob/5440691d44abea537b08aeb60153a31654d66a9b/noweb