The Journal of Open Source Software

surtvep: An R package for estimating time-varying effects

Lingfeng Luo[1], **Wenbo Wu**[2], **Jeremy Taylor**[1], **Jian Kang**[1], **and Kevin He**[1¶]

1 Department of Biostatistics, School of Public Health, University of Michigan **2** Departments of Population Health and Medicine, New York University Grossman School of Medicine ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Øystein Sørensen

Reviewers:

- @adibender
- @turgeonmaxime

Submitted: 05 July 2023
Published: unpublished

Summary

The surtvep package is an open-source software designed for estimating time-varying effects in survival analysis using the Cox non-proportional hazards model in R. With the rapid increase in large-scale time-to-event data from national disease registries, detecting and accounting for time-varying effects in medical studies have become crucial. Current software solutions often face computational issues such as memory limitations when handling large datasets. Furthermore, modeling time-varying effects for time-to-event data can be challenging due to small at-risk sets and numerical instability near the end of the follow-up period. surtvep addresses these challenges by implementing a computationally efficient Kronecker product-based proximal algorithm, supporting both unstratified and stratified models. The package also incorporates P-spline and smoothing spline penalties to improve estimation. Cross-validation and information criteria are available to determine the optimal tuning parameters. Parallel computation is enabled to further enhance computational efficiency. A variety of operating characteristics are provided, including estimated time-varying effects, confidence intervals, hypothesis testing, and estimated hazard functions and survival probabilities. The surtvep package thus offers a comprehensive and flexible solution to analyzing large-scale time-to-event data with dynamic effect trajectories.

Statement of Need

The Cox non-proportional hazards model is a flexible and powerful tool for modeling time-varying effects of covariates in survival analysis. However, as the size of a dataset increases, the computational costs of this model can become substantial. Current software solutions, which may be effective for smaller datasets, face challenges when handling larger datasets.

Numerous studies have demonstrated the widespread presence of time-varying effects. For instance, the scientific literature has shown that factors like age, sex, and race can have non-constant associations with survival in cases such as end-stage renal disease (He et al., 2017, 2022), and breast cancer patients receiving neo-adjuvant chemotherapy and head and neck cancer patients (Baulies et al., 2015; Brouwer et al., 2020). Ignoring the variations and relying solely on the Cox proportional hazards model can lead to inaccurate risk prediction and suboptimal treatment development.

With the rising need for modeling time-varying effects, researchers have developed methods to handle the data (Gray, 1992, 1994; Hastie & Tibshirani, 1993; Zucker & Karr, 1990). In terms of implementation, these methods expand the original data in a repeated measurement format (Therneau et al., 2017) using existing software such as the survival package (Therneau, 2023). Even with moderate sample sizes, this leads to a large and computationally burdensome working dataset. surtvep addresses this issue by implementing a computationally efficient

Kronecker product-based proximal algorithm (Perperoglou et al., 2006), which can handle time-varying effects in large-scale studies with improved efficiency and parallel computing capabilities. Compared with existing computational packages, such as the `coxph` function, `surtvep` demonstrates a much more efficient performance, with both runtime and memory consumption reduced considerably.

Another issue of numerical instability arises when analyzing data with binary covariates that have limited variation. `surtvep` implements a proximal Newton's method to improve the estimation. Additionally, adding a penalty can improve the estimation. `surtvep` also supports P-spline and smoothing spline (Eilers & Marx, 1996; Wood, 2017a, 2017b), to further improve estimation stability. The improved estimation performance of `surtvep` is demonstrated in our recent studies (Luo et al., 2023; Wu et al., 2022).

Finally, our method has several other features worth noting. First, `surtvep` supports the stratified model, which enables researchers to account for differences in baseline hazard functions across distinct clusters or other grouping factors. This is particularly useful when there are distinct subgroups within the data that may have different baseline hazards. Second, `surtvep` enables shared-memory parallel computation features, which can significantly improve the performance of the software when working with large datasets. Also, `surtvep` supports Breslow approximation (Breslow, 1974), which significantly improves the computational speed when a large number of times are present.

Functions

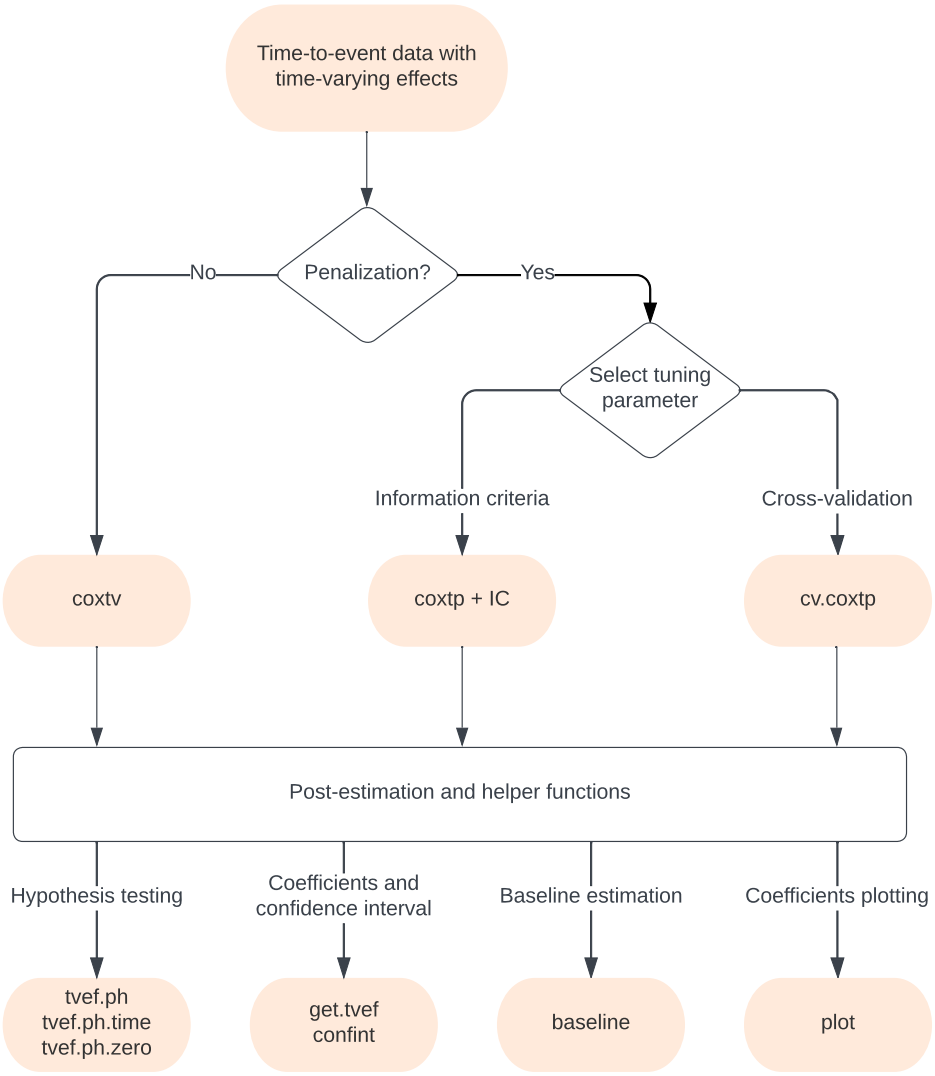

Figure 1: Flowchart for functions in the `surtvep` package. `coxtv` utilizes proximal Newton's method to estimate the time-varying coefficients. `coxtp` combines the Newton's approach with penalization. `IC` calculates different information criteria to select the best tuning parameter in front of the penalty term. `cv.coxtp` uses cross-validation for tuning parameter selection. `tvef.ph`, `tvef.ph.time` and `tvef.ph.zero` provide hypothesis testing for the fitted model. `get.tvef` retrieves the time-varying coefficients for the fitted model. `confint` provides confidence intervals for these coefficients. `baseline` offers the baseline hazard estimations. `plot` visualizes the estimated time-varying coefficients.

`surtvep` is a powerful statistical software package designed for analyzing time-varying effects of time-to-event data. The software offers two main functions for estimating time-varying coefficients in survival analysis.

To model time-varying coefficients in `surtvep`, we first define the time-varying coefficients as $\beta(t)$, which represents the effects of predictor variables on the outcome at different points in time. We then use a set of B-spline basis functions to span the $\beta(t)$, which provides a flexible

and accurate way to capture the time-dependent effects of the predictors. These B-spline basis functions are generated using the `splines` R package with a fixed number of basis functions.

Once we have established the basis functions for the time-varying coefficients, `coxtv` employs a proximal Newton's approach to estimate the coefficients in front of the B-spline basis functions. This approach iteratively updates the coefficients until a maximum of the log-partial likelihood is reached. Backtracking line search is utilized to improve the estimation. We have also implemented a shared-memory parallelization to enable faster convergence.

`coxtp` is the second main function, adding a penalty term to the original objective function. This approach iteratively updates the coefficients until a maximum of the penalized log-partial likelihood is reached. `coxtp` provides two options for penalized regression: P-spline and smoothing spline.

- P-spline stands for penalized B-spline. It combines the B-spline basis with a discrete quadratic penalty on the difference of basis coefficients between adjacent knots. When the penalty term goes to infinity, the time-varying effects are reduced to be constant.
- Smoothing spline is a derivative-based penalty combined with B-spline. When the cubic B-spline is used for constructing the basis functions, the smoothing spline penalizes the second-order derivative, which reduces the time-varying effect to a linear term when the penalty term goes to infinity. When the quadratic B-spline is used for constructing the basis functions, the smoothing spline penalizes the first-order derivative, which reduces the time-varying effect to a constant when the penalty term goes to infinity. See Wood (2017b) for details.

`surtvep` also provides a function `IC` to select the best tuning parameter in front of the penalty term. `IC` can be used to calculate the modified Akaike information criterion (mAIC), the Takeuchi information criterion (TIC) and the generalized information criterion (GIC) (Akaike, 1998; Luo et al., 2023; Takeuchi, 1976). Generally, mAIC, TIC and GIC have relatively similar performance. Using one of these criteria to select tuning parameters is considerably faster than using cross-validation, which is also provided in `surtvep` via function `cv.coxtp`.

Finally, `surtvep` offers a comprehensive suite of hypothesis testing capabilities, allowing researchers to assess the validity and significance of their models. Specifically, `surtvep` can perform the following hypothesis tests: (1) testing the proportional hazards assumption to verify the model's suitability for the given data; and (2) examining the pointwise significance of covariates effects at different event times to assess the impact of each covariate on the outcome of interest. To conduct these hypothesis tests, `surtvep` employs the Wald test statistic, a widely-used method for inference.

Quick Start

The purpose of this section is to introduce the basics of `surtvep`. Interested users are referred to the online tutorial at *https://um-kevinhe.github.io/surtvep/index.html* for detailed instructions.

`surtvep` can be easily installed by launching an R prompt and running the following commands:

```
install.packages('surtvep')
```

Next, we load an example data set that includes two columns `z` of continuous covariates, a column `time` indicating the time to an event, and a column "event" of event indicators.

```
data("ExampleData")
z     <- ExampleData$z
time  <- ExampleData$time
event <- ExampleData$event
```

We can fit the Newton's method without penalization using the most basic call to `coxtv`. For the Newton's method with penalization, we call the `coxtp` function.

```
fit.tv <- coxtv(z = z, event = event, time = time)
fit.penalize <- coxtp(z = z, event = event, time = time)
```

We use `IC` to calculate the information criteria and select the best tuning parameter:

```
fit.ic <- ic(fit.penalize)
```

`fit.tv` is an object of class `coxtv` that contains all the relevant information of the fitted model for further use. `fit.ic` contains three objects of class `coxtp`, corresponding to the selected model using mAIC, TIC and GIC. Various methods are provided for the objects such as plotting and hypothesis testing.

We can visualize the time-varying coefficients through the plot method:

```
plot(fit.tv, ylim = c(-3,10))
plot(fit.ic$mAIC, ylim = c(-3,10))
```

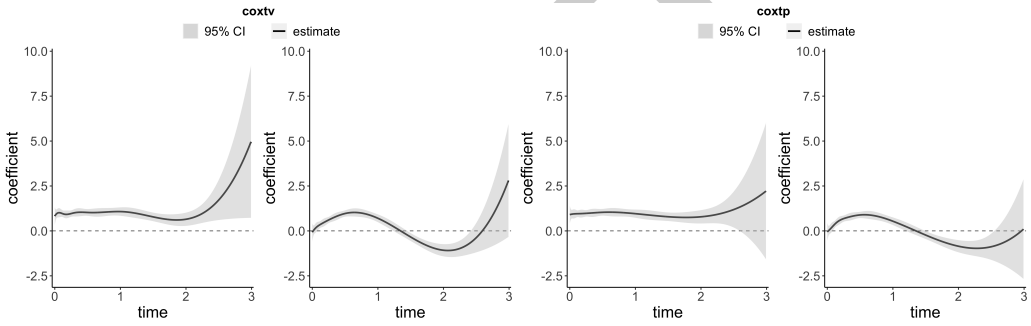

Figure 2: The estimated time-varying coefficients (log hazard ratio) from `coxtv` and `coxtp`. The tuning parameter for `coxtp` is selected using mAIC.

In utilizing the `coxtv` and `coxtp` functions, users have the flexibility to choose based on their dataset's specifications. Numerical instabilities are commonly encountered when analyzing survival data of a small sample size or when the data includes some binary covariates with proportions that approach either zero or one. In these scenarios, the second-order information matrix can become ill-conditioned. See discussion in (Luo et al., 2023; Wu et al., 2022). To address this issue, the employment of the penalized method `coxtp` is recommended. When determining the number of basis functions, a typical range is between 5-10. Though the choice is somewhat flexible, it has limited impact on results unless set too small (Gray, 1992). Users might consider increasing this number when applying the penalized method.

Data Example

We demonstrate the effectiveness of `surtvep` by applying it to a real-world dataset from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program (National Cancer Institute, 2019). We estimate the hazard ratios of the cancer stage of kidney, lung, and breast, as shown in Figure 3. Our analysis highlights the dynamic nature of hazard ratios for cancer death among patients with metastatic stage compared to those with localized stage.

In the first year after diagnosis, the hazard ratio is strikingly high, indicating a significant difference in survival outcomes between metastatic and localized stage patients. However, this disparity shrinks considerably by the eighth year, reflecting the diminishing relevance of the initial cancer stage in the prognosis of long-term survivors. This example illustrates the importance of accounting for time-varying effects, which has been effectively addressed by `surtvep` through its flexible and efficient approach to modeling these dynamics. By providing

accurate and efficient modeling of time-varying effects in large-scale datasets, `surtvep` serves as a valuable tool for researchers working with complex survival data.

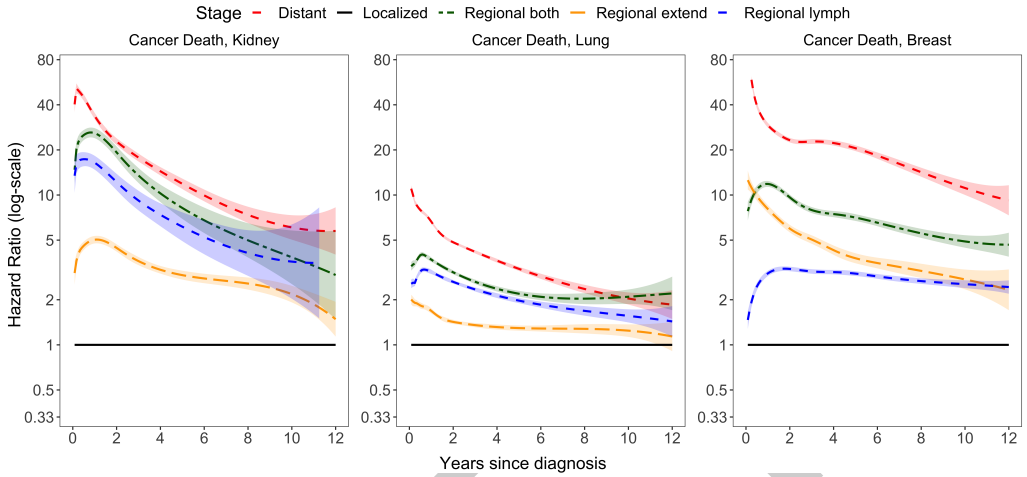

Figure 3: Time-varying effects of cancer stage in SEER data.

Availability

Stable releases of the `surtvep` package will be made available via the Comprehensive R Archive Network. Alternatively, the `surtvep` package is available on GitHub (*https://github.com/UM-KevinHe/surtvep*). Use of the `surtvep` package has been extensively documented in the package documentation and on the tutorial website (*https://um-kevinhe.github.io/surtvep/index.html*).

Funding

This project was partially supported by the US National Cancer Institute (R01CA-129102) and National Institute of Diabetes and Digestive and Kidney Diseases (R01DK-129539).

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotogu Akaike* (pp. 199–213). Springer.

Baulies, S., Belin, L., Mallon, P., Senechal, C., Pierga, J., Cottu, P., Sablin, M., Sastre, X., Asselain, B., Rouzier, R., & others. (2015). Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. *British Journal of Cancer*, *113*(1), 30–36.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99.

Brouwer, A. F., He, K., Chinn, S. B., Mondul, A. M., Chapman, C. H., Ryser, M. D., Banerjee, M., Eisenberg, M. C., Meza, R., & Taylor, J. M. (2020). Time-varying survival effects for squamous cell carcinomas at oropharyngeal and nonoropharyngeal head and neck sites in the United States, 1973-2015. *Cancer*, *126*(23), 5137–5146.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, *87*(420),

942–951.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics*, *50*(3), 640–652.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779.

He, K., Yang, Y., Li, Y., Zhu, J., & Li, Y. (2017). Modeling time-varying effects with large-scale survival data: An efficient quasi-Newton approach. *Journal of Computational and Graphical Statistics*, *26*(3), 635–645.

He, K., Zhu, J., Kang, J., & Li, Y. (2022). Stratified Cox models with time-varying effects for national kidney transplant patients: A new blockwise steepest ascent method. *Biometrics*, *78*(3), 1221–1232.

Luo, L., He, K., Wu, W., & Taylor, J. M. (2023). Using information criteria to select smoothing parameters when analyzing survival data with time-varying coefficient hazard models. *Statistical Methods in Medical Research*, *in press*.

National Cancer Institute. (2019). *Surveillance, Epidemiology, and End Results (SEER) Program. SEER*Stat Database*. https://www.seer.cancer.gov.

Perperoglou, A., Cessie, S. le, & Houwelingen, H. C. van. (2006). A fast routine for fitting Cox models with time varying effects of the covariates. *Computer Methods and Programs in Biomedicine*, *81*(2), 154–161.

Takeuchi, K. (1976). Distribution of an information statistic and the criterion for the optimal model. *Mathematical Science*, *153*, 12–18.

Therneau, T. (2023). *A package for survival analysis in R*. https://CRAN.R-project.org/package=survival

Therneau, T., Crowson, C., & Atkinson, E. (2017). *Using time dependent covariates and time dependent coefficients in the Cox model*. Survival Vignettes. https://stat.ethz.ch/R-manual/R-patched/library/survival/doc/timedep.pdf

Wood, S. N. (2017a). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.

Wood, S. N. (2017b). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing*, *27*, 985–989.

Wu, W., Taylor, J. M., Brouwer, A. F., Luo, L., Kang, J., Jiang, H., & He, K. (2022). Scalable proximal methods for cause-specific hazard modeling with time-varying coefficients. *Lifetime Data Analysis*, *28*(2), 194–218.

Zucker, D. M., & Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics*, *18*(1), 329–353.