The Journal of Open Source Software

survPen: an **R** package for hazard and excess hazard modelling with multidimensional penalized splines

Mathieu Fauvernier[1, 2], **Laurent Remontet**[1, 2], **Zoé Uhry**[1, 2, 3], **Nadine Bossard**[1, 2], **and Laurent Roche**[1, 2]

1 Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique - Bioinformatique, Lyon, France **2** Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France **3** Département des Maladies Non-Transmissibles et des Traumatismes, Santé Publique France, Saint-Maurice, France

DOI: 10.21105/joss.01434

Software

- Review
- Repository
- Archive

Submitted: 03 May 2019
Published: 23 August 2019

Background

Survival analysis deals with studying the elapsed time until an event occurs. When the event of interest is death, it aims at describing the survival probability and its corresponding mortality hazard. In epidemiology, as patients may die from their disease or from other causes, it is relevant to study the mortality due to their disease; also called "excess mortality". This excess mortality is useful to make comparisons between different countries and time periods (Allemani et al., 2018; Uhry et al., 2017) and is directly linked to the concept of net survival (Perme, Stare, & Estève, 2012), i.e. the survival that would be observed if patients could only die from their disease.

`survPen` is an R package that implements flexible regression models for (net) survival analysis. Model specification is carried out on the logarithm of the (excess) hazard scale. `survPen` provides an efficient procedure to estimate the model parameters, and tools for (excess) hazard and (net) survival predictions with associated confidence intervals.

In survival and net survival analysis, in addition to modelling the effect of time (via the baseline hazard), one has often to deal with several continuous covariates and model their functional forms, their time-dependent effects, and their interactions. Model specification becomes therefore a complex problem and penalized regression splines (Ruppert, Wand, & Carroll, 2003; Wood, 2017) represent an appealing solution to that problem as splines offer the required flexibility while penalization limits overfitting issues.

Current implementations of penalized survival models can be slow or unstable and sometimes lack some key features like taking into account expected mortality to provide net survival and excess hazard estimates. In contrast, `survPen` provides an automated, fast, and stable implementation (thanks to explicit calculation of the derivatives of the likelihood) and offers a unified framework for multidimensional penalized hazard and excess hazard models.

Summary

`survPen` is an implementation of multidimensional penalized hazard and excess hazard models for time-to-event data in **R** (R Core Team, 2018). It implements the method detailed in Fauvernier et al. (2019) which is itself included in the framework for general smooth models proposed by Wood, Pya, & Säfken (2016). Other **R** packages propose to fit flexible survival models via penalized regression splines (`rstpm2`, `bamlss`, `R2BayesX`, etc). However, the way they estimate the smoothing parameters is not optimal as they rely on either

derivative-free optimization (`rstpm2`) or MCMC (`bamlss`, `R2BayesX`), leading to possibly unstable or time-consuming analyses. The main objective of the `survPen` package is to offer a fully automatic, fast, stable and convergent procedure in order to model simultaneously non-proportional, non-linear effects of covariates and interactions between them. A second objective is to extend the approach to excess hazard modelling (J. Estève, Benhamou, Croasdale, & Raymond, 1990; L. Remontet, Bossard, Belot, Estève, & French Network of Cancer Registries, 2007). `survPen` is a free and open-source **R** package, available via GitHub at https://github.com/fauvernierma/survPen or via the CRAN repository at https://CRAN.R-project.org/package=survPen. The major features of `survPen` are documented in a walkthrough vignette that is included with the package (https://htmlpreview.github.io/?https://github.com/fauvernierma/survPen/blob/master/inst/doc/survival_analysis_with_survPen.html)

Those features include:

- Univariate penalized splines for the baseline hazard as well as any other continuous covariate.
- Penalized tensor product splines for time-dependent effects and interactions between several continuous covariates.
- Interactions between penalized splines and unpenalized continuous or categorical variables.
- Automatic smoothing parameter estimation by either optimizing the Laplace approximate marginal likelihood (LAML; Wood et al., 2016) or likelihood cross-validation criterion (LCV; O'Sullivan, 1988).
- Excess hazard modelling by specifying expected mortality rates.

`survPen` may be of interest to those who 1) analyse any kind of time-to-event data: mortality, disease relapse, machinery breakdown, unemployment, etc 2) wish to describe the associated hazard and to understand which predictors impact its dynamics.

Using the `survPen` package for time-to-event data analyses will help choose the appropriate degree of complexity in survival and net survival contexts while simplifying the model building process.

Multidimensional splines with `survPen` are currently being used in three major ongoing projects:

- Modelling the effects of time since diagnosis, age at diagnosis and year of diagnosis on the mortality due to cancer using French cancer registries data (FRANCIM network, around 1,200,000 tumours diagnosed between 1989 and 2015). This study will provide the new national estimates of cancer survival in France and its results will be used in the evaluation of the French "Plan Cancer" at the end of 2019.
- Modelling the effect of the European Deprivation Index (EDI) on the mortality due to cancer in France, using data from the FRANCIM network; this is the first time that EDI is available in all FRANCIM registries (around 210,000 tumours, diagnosed between 2006 and 2009 in 18 registries).
- For the first time modelling the effects of time since onset, age at onset, current age, year of onset and sex on the mortality due to multiple sclerosis in the biggest cohort of multiple sclerosis patients in France (37,524 patients diagnosed over the period 1960-2014 in 18 OFSEP centres).

Acknowledgements

This research was conducted as part of the first author's PhD thesis supported by the French Ministère de l Enseignement supérieur, de la Recherche et de l Innovation. The authors thank

the ANR (Agence Nationale de la Recherche) for supporting this study of the CENSUR group (ANR grant number ANR-12-BSV1-0028). This research was also carried out within the context of a four-institute cancer surveillance program partnership involving the Institut National du Cancer (INCa), Santé Publique France (SPF), the French network of cancer registries (FRANCIM), and Hospices Civils de Lyon (HCL) through a grant from INCa (attributive decision N° 2016-131). The authors are grateful to Jacques Estève for his valuable advice.

References

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Niksic, M., Bonaventure, A., et al. (2018). Global surveillance of trends in cancer survival 2000-14 (concord-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*, *391*(10125), 1023–1075. doi:10.1016/S0140-6736(17)33326-3

Estève, J., Benhamou, E., Croasdale, M., & Raymond, L. (1990). Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine*, *9*(5), 529–38. doi:10.1002/sim.4780090506

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., & the Challenges in the Estimation of Net Survival Working Survival Group. (2019). Multidimensional penalized hazard model with continuous covariates: Applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. doi:10.1111/rssc.12368

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, *9*(2), 363–379. doi:10.1137/0909024

Perme, M. P., Stare, J., & Estève, J. (2012). On estimation in relative survival. *Biometrics*, *68*(1), 113–120. doi:10.1111/j.1541-0420.2011.01640.x

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Remontet, L., Bossard, N., Belot, A., Estève, J., & French Network of Cancer Registries. (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, *26*(10), 2214–2228. doi:10.1002/sim.2656

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press. doi:10.1017/CBO9780511755453

Uhry, Z., Bossard, N., Remontet, L., Iwaz, J., Roche, L., Grell Eurocare-5 Working Group, & Censur Working Survival Group. (2017). New insights into survival trend analyses in cancer population-based studies: The SUDCAN methodology. *European Journal of Cancer Prevention*, *26 Trends in cancer net survival in six European Latin Countries: the SUDCAN study*, S9–S15. doi:10.1097/CEJ.0000000000000301

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Second Edition, Chapman; Hall/CRC.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. doi:10.1080/01621459.2016.1180986