Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations

Simon Bussy*[1,2], Van Tuan Nguyen[2,3], Antoine Barbieri[4], Sarah Zohar[1], and Anne-Sophie Jannot[1,5]

[1]*INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France*
[2]*LOPF, Califrais' Machine Learning Lab, Paris, France*
[3]*LPSM, UMR 8001, CNRS, Sorbonne University, Paris, France*
[4]*INSERM, UMR 1219, Bordeaux Population Health Research Center, Univ. Bordeaux, France*
[5]*Biomedical Informatics and Public Health Department, EGPH, APHP, Paris, France*

Abstract

This paper introduces a prognostic method called *lights* to deal with the problem of joint modeling of longitudinal data and censored durations, where a large number of both longitudinal and time-independent features are available. In the literature, standard joint models are either of type shared random-effect or joint latent class ones ; where the association structure between the longitudinal and the time-to-event submodels takes respectively the form of either shared association features learned from the longitudinal processes and included as potential risk factor in the survival model, or latent classes modeling population heterogeneity. We pick modeling ideas from both worlds and use appropriate penalties during inference for being able to learn from a high-dimensional context. The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on a publicly available dataset. Our proposed method significantly outperforms the state-of-the-art joint models regarding risk prediction in terms of C-index in a so-called real-time prediction paradigm, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant features being relevant from a practical perspective. Thus, we propose a powerfull tool with the ability of automatically determining significant prognostic longitudinal features, which is of increasing importance in many areas: for instance personalized medicine, or churn prediction in a customer profile and activity monitoring setting, to name but a few.

Keywords. High-dimensional estimation; Joint modeling; Multivariate longitudinal data; Survival analysis

1 section 1

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

*Corresponding author: simon.bussy@gmail.com

Galen Andrew and Jianfeng Gao. Scalable training of l1-regularized log-linear models. In *International Conference on Machine Learning*, pages 33–40. ACM, 2007.

Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Johanna JM Takkenberg, and Emmanuel Lesaffre. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical methods in medical research*, 26(4):1787–1801, 2017.

Peter C Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*, 31(29):3946–3958, 2012.

Peter C Austin. Correction:'generating survival times to simulate cox proportional hazards models with time-varying covariates'. *Statistics in Medicine*, 32(6):1078–1078, 2013.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *arXiv preprint arXiv:1108.0775*, 2011.

Emmanuel Bacry, Martin Bompaire, Philip Deegan, Stéphane Gaïffas, and Søren V Poulsen. Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *The Journal of Machine Learning Research*, 18 (1):7937–7941, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.

Paul Blanche, Cécile Proust-Lima, Lucie Loubère, Claudine Berr, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Norman E Breslow. Contribution to discussion of paper by dr cox. *JJournal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34:216–217, 1972.

Simon Bussy, Agathe Guilloux, Stéphane Gaïffas, and Anne-Sophie Jannot. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical methods in medical research*, 28(5):1523–1539, 2019a.

Simon Bussy, Raphaël Veil, Vincent Looten, Anita Burgun, Stéphane Gaïffas, Agathe Guilloux, Brigitte Ranque, and Anne-Sophie Jannot. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *BMC medical research methodology*, 19(1):50, 2019b.

Yueh-Yun Chi and Joseph G Ibrahim. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432–445, 2006.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Gideon Dresdner Fabian Pedregosa, Geoffrey Negiar. copt: composite optimization in python. 2020. doi: 10.5281/zenodo.1283339. URL http://openopt.github.io/copt/.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Benjamin Gompertz. Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, (115):513–583, 1825.

Frank E Harrell, Kerry L Lee, and Daniel B Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

Laura A Hatfield, Mark E Boye, and Bradley P Carlin. Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of Biopharmaceutical Statistics*, 21(5):971–991, 2011.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, 16(1):117, 2016.

Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC medical research methodology*, 18(1):50, 2018.

Miran A Jaffa, Mulugeta Gebregziabher, and Ayad A Jaffa. A joint modeling approach for right censored high dimensional multivariate longitudinal data. *Journal of biometrics & biostatistics*, 5(4), 2014.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806, 1992.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

Haiqun Lin, Charles E McCulloch, and Susan T Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16):2369–2382, 2002a.

Haiqun Lin, Bruce W Turnbull, Charles E McCulloch, and Elizabeth H Slate. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65, 2002b.

Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255: 2897–2899, 1962.

B. N. Mukherjee and S. S. Maiti. On some properties of positive definite toeplitz matrices and their possible applications. *Linear algebra and its applications*, 102:211–240, 1988.

Paul A Murtaugh, E Rolland Dickson, Gooitzen M Van Dam, Michael Malinchoc, Patricia M Grambsch, Alice L Langworthy, and Chris H Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1): 126–134, 1994.

Cécile Proust-Lima, Mbéry Séne, Jeremy MG Taylor, and Hélène Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research*, 23(1):74–90, 2014.

Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: The r package lcmm. *Journal of Statistical Software, Articles*, 78(2):1–56, 2017. ISSN 1548-7660. doi: 10.18637/jss.v078.i02. URL https://www.jstatsoft.org/v078/i02.

Dimitris Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software, Articles*, 72(7):1–46, 2016. ISSN 1548-7660. doi: 10.18637/jss.v072.i07. URL https://www.jstatsoft.org/v072/i07.

Dimitris Rizopoulos and Pulak Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12): 1366–1380, 2011.

R. T. Rockafellar. *Conjugate duality and optimization*. Regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1974. ISBN 0-89871-013-8. URL http://opac.inria.fr/record=b1083670. Essentially to be regarded as supplement to the book Convex analysis.

Michael Schemper and Robin Henderson. Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255, 2000.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Glenn T Sueyoshi. Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of econometrics*, 51(1-2):25–58, 1992.

Terry M Therneau and Patricia M Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.

Ryan Tibshirani. Proximal gradient descent and acceleration. *Lecture Notes*, 2010.

Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Jeroen K Vermunt and Jay Magidson. Latent class models for classification. *Computational Statistics & Data Analysis*, 41(3-4):531–537, 2003.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Ping Wang, Wei Shen, and Mark Ernest Boye. Joint modeling of longitudinal outcomes and survival using latent growth modeling approach in a mesothelioma trial. *Health Services and Outcomes Research Methodology*, 12(2-3):182–199, 2012.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

Menggang Yu, Ngayee J Law, Jeremy MG Taylor, and Howard M Sandler. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, pages 835–862, 2004.

Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. *Advances in neural information processing systems*, 24:352–360, 2011.

Zhongheng Zhang, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse, and Catharina GM Groothuis-Oudshoorn. Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7), 2018.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.