Author Contributions Checklist

June 21, 2021

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

Publicly Available:

We analyzed a small number of datasets provided in the R-Survival package (Therneau and Lumley 2015). These include:

- Colon are data on trials of adjuvant chemotherapy for colon cancer, (Laurie et al. 1989).
- Lung are data extracted from the North Central Cancer Treatment Group on mortality for advanced lung cancer (Loprinzi et al. 1994).
- NAFLD is a large population-based study investigating non-alcoholic fatty liver disease (Allen et al. 2018).
- Heart investigates mortality in patients from the Stanford heart transplant program (Crowley and Hu 1977).
- PBCseq are follow-up laboratory data from the Mayo clinical trial on primary biliary cholangitis and D-penicillamine treatment (Therneau and Grambsch 2000).

The purpose of these examples is to evaluate the performance of our proposed method in comparison to the frequentist Cox model on real-world applications.

Non-publicly available:

The UK Biobank (UKB) (Sudlow et al. 2015) is a large-scale biomedical database established in 2006. In total there are 502628 participants, recruited between 2006 and 2010. All participants were between 40-69 years of age at their recruitment date. We study the association of standard risk factors and comorbidities, taken from the electronic health records, with the occurrence of myocardial infarction. For details of the analysis see Section 4 of the paper.

Availability

☒ Partial Data **are** publicly available.

☒ Partial Data **cannot be made** publicly available.

Publicly available data

☒ Data are available online at: https://github.com/alexwjung/ProbCox

☒ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

The UK-Biobank (UKB) data contains sensible information on individuals and cannot be openly shared for privacy considerations. However, researchers can apply for access to the UKB under: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access

The simulation study presented in the paper closely resembles the data structure encountered in electronic health records, similar to the data in the UKB, and can be used as guidance. Additionally, we provide a small simulation reflecting the analysis in the UKB.

Electronic health records and biobanks contain a vast array of information on individuals that open new possibilities to study disease and are therefore highly important in biomedical research. However, the wealth of information collected on individuals makes it difficult to obfuscate data in such a way as to preserve important structures in the data. This inevitably leads to issue for openly distributing the data.

Description

The data provided can be found in [Data]-(./data)

The fake data resembling the analysis in the UKB can be created with (./replication/UKB/00_fakedata.ipynb)

File format(s)

☒ CSV or other plain text.
☐ Software-specific binary format (.Rda, Python pickle, etc.): pkcle
☒ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (please specify):

Data dictionary

☐ Provided by authors in the following file(s):
☒ Data file(s) is(are) self-describing (e.g., netCDF files)
☐ Available at the following URL:

Additional Information (optional)

Part 2: Code

Abstract

We provide a python package (probcox) for the implemented method. This is a scalable version of the Cox model fitted via stochastic variational inference. The code is written entirely in pytorch/pyro and numpy. We provide an algorithm to efficiently simulate survival times with time-varying covariates that resemble disease association studies in electronic health records. The original scripts for all the analyses are provided. Additionally, there are jupyter notebooks that can be run in google colab to readily replicate all of the results presented (except for the UKB as we cannot provide the data as mentioned above). However, we do provide a similar replicated analysis for the UKB, based on a simulated dataset.

Description

Code format(s)

- ☐ Script files
  - ☒ R
  - ☒ Python
  - ☐ Matlab
  - ☐ Other:
- ☐ Package
  - ☒ R
  - ☒ Python
  - ☐ MATLAB toolbox
  - ☐ Other:
- ☐ Reproducible report
  - ☐ R Markdown
  - ☒ Jupyter notebook
  - ☒ Other: Google Colab
- ☒ Shell script
- ☐ Other (please specify):

Supporting software requirements

We provide replicable notebooks that can be used with Google Colab (requirement: Google account).

The notebooks can also be run via local jupyter notebooks.

Version of primary software used Python version 3.7

R version 4.0.3

Libraries and dependencies used by the code We use conda [https://docs.conda.io/en/latest/] as a package manager and the full list of install packages can be found via [R environment]-(./docs/requirements_R.txt) and [Python environment]-(./docs/requirements_python.txt), respectively.

When probcox is installed via pip it installs most of the required packages automatically (assuming a basic python install: numpy, pandas, matplotlib, h5py, tqdm).

Supporting system/hardware requirements (optional)

We had access to a cluster to run the simulations. The main purpose for this was to run simulations simultaneously. The proposed package automatically adapts to the underlying hardware and parallelizes computation where possible.

Parallelization used

- ☐ No parallel code used
- ☒ Multi-core parallelization on a single machine/node
  - Number of cores used: varying - can be adjusted by user - will be clear in scripts.
- ☐ Multi-machine/multi-node parallelization
  - Number of nodes and cores used:

License

- ☒ MIT License (default)
- ☐ BSD
- ☐ GPL v3.0

☐ Creative Commons
☐ Other: (please specify below)

Additional information (optional)

The code for the model can be installed via:

```
$ pip install probcox
```

Scope

The provided workflow reproduces:

☒ Any numbers provided in text in the paper
☒ All tables and figures in the paper
- [] Selected tables and figures in the paper, as explained and justified below:

Workflow

Format(s)

☐ Single master code file
☐ Wrapper (shell) script(s)
☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
☐ Text file (e.g., a readme-style file) that documents workflow
☐ Makefile
☒ Other (more detail in *Instructions* below)

Instructions

We assume that individuals have a google account and can run the notebooks via google colabs. The replication can similarly be run on a local installation with jupyter notebooks.

The original scripts used to run the analyses, most of the raw outputs, and additional scripts for easier replication are provided. We mainly distinguish between application (similarly named as the used data) and simulations, where standard cases may be shortened to sim_sc and high dimensional cases to sim_hd.

- The accompanying data can be found in [Data]-(./data)
- Information on the installed packages, the Author Contributions Checklist, and other relevant documents are found in [Docs]-(./docs)
- Most of the outputs and raw files, like estimates, tables, figures, etc. are found in [Out]-(./out)
- A dedicated folder with the relevant outputs for the paper can be found in [Paper]-(./paper)
- The scripts used to run the analysis is found in [Scripts]-(./scripts)
- A dedicated folder for the analysis of the UKB data with subfolder [output]-(./ukb/out) and [scripts]-(./ukb/scripts) can be found in [UKB-Analysis]-(./ukb)

Notebooks for easy replication of the analysis can be found in [Replication]-(./replication) The figures and tables are listed in the order as they appear in the paper (supplementary materials at the end). Some of the additional results in the supplementary materials can be produced by the corresponding notebooks from the main paper by changing the settings (as descriped in the notebooks).

Link to the .ipynb files - link to a specific colab session

- Replication notebooks for the applications are in [Replicate Applications]-(./replication/application)
  - [Colon]-(./replication/application/colon.ipynb) - [Colab]-(https://colab.research.google.com/drive/1HifKMp2SjKB3NCnNe-vD1EiAf2bQQ7Rp?usp=sharing)

  - [Lung]-(./replication/application/lung.ipynb) - [Colab]-(https://colab.research.google.com/drive/1IniSnT1bUINtUnu_owezJ0FWeKyXWgvu?usp=sharing)
  - [Heart]-(./replication/application/heart.ipynb) - [Colab]-(https://colab.research.google.com/drive/1bXWSxZA4KvRvxi5xZswDPbdIEaPTrljv?usp=sharing)
  - [Nafld]-(./replication/application/nafld.ipynb) - [Colab]-(https://colab.research.google.com/drive/13IJLUfXSqF_3U9dsEBuvo-Vy29r7WLzn?usp=sharing)
  - [PBCseq]-(./replication/application/pbcseq.ipynb) - [Colab]-(https://colab.research.google.com/drive/15Y9XK5YlldRgpha7D0aT9eMmpMN_JxLu?usp=sharing)
- Replication notebooks for the simulations are in [Replicate Simulations]-(./replication/simulations)
  - [Standard Case 1]-(./replication/simulation/standard_case1.ipynb) - [Colab]-(https://colab.research.google.com/drive/1iEoO9hHkgRWzaLhbU9VYhYk6U6V8nffG?usp=sharing)
  - [Standard Case 2]-(./replication/simulation/standard_case2.ipynb) - [Colab]-(https://colab.research.google.com/drive/1lIm7d866QtbIxqY6IRhIFrfTECLBWSDn?usp=sharing)
  - [High-dimensional Case]-(./replication/simulation/highdimensional_case.ipynb) - [Colab]-(https://colab.research.google.com/drive/1Db9x78fYhhj5yVTalMhKsP6wOm9tArKr?usp=sharing)
  - [Resources]-(./replication/simulation/resources.ipynb) - [Colab]-(https://colab.research.google.com/drive/1BWSuWMOFgxPveoWgb7AfX7DuPl6n1ZeV?usp=sharing)
- To replicate the tables presented in the paper go to [Replicate Tables]-(./replication/tables)
  - [Data Example]-(./replication/simulation/tables/data_example.ipynb) - [Colab]-(https://colab.research.google.com/drive/1yHM5iDRE0GqTsj7Jpql32PjpNJaopSJX?usp=sharing)
  - [Likelihood Approximation]-(./replication/simulation/tables/likelihood_approx.ipynb) - [Colab]-(https://colab.research.google.com/drive/1HJeGSiSX6_plwbgJleY4RjYFa13Gm2O-?usp=sharing)
  - [Standard Case 1]-(./replication/simulation/tables/standard_case1_table.ipynb) - [Colab]-(https://colab.research.google.com/drive/11XX0E36TUTNnTFhEeW-It7YIm-5vKc4q?usp=sharing)
  - [Standard Case 2]-(./replication/simulation/tables/standard_case2_table.ipynb) - [Colab]-(https://colab.research.google.com/drive/13Pt2tMoJAKkgpU-L9KmqWj-tgsgQNBaz?usp=sharing)
  - [High-dimensional Case]-(./replication/simulation/tables/highdimensional_case_table.ipynb) - [Colab]-(https://colab.research.google.com/drive/1Uj6lQaivKj7UaEgR-j5feZgXFhXke0R1?usp=sharing)
  - [Likelihood Approximation - large P]-(./replication/simulation/tables/likelihood_approx_additional1.ipynb) - [Colab]-(https://colab.research.google.com/drive/1USX1g8PmHkm6Di1WiwAV0u9nJdZ1JtPw?usp=sharing)
  - [Likelihood Approximation - predictor]-(./replication/simulation/tables/likelihood_approx_additional2.ipynb) - [Colab]-(https://colab.research.google.com/drive/1Kx2y_E4aSLx6AG0rlQd3pKDJ2F6HR-_f?usp=sharing)
- To replicate the figures presented in the paper go to [Replicate Figures]-(./replication/figures)
  - [Schematic]-(./replication/simulation/figures/schematic.ipynb) - [Colab]-(https://colab.research.google.com/drive/1Hz1IG6z4fOJBTNEIM6jSnyO6l586P3G1?usp=sharing)
  - [Likelihood Approximation]-(./replication/simulation/figures/likelihood_training.ipynb) - [Colab]-(https://colab.research.google.com/drive/1kz42UvTAag7XxEWCgMhw6GidP_fuwW4p?usp=sharing)
  - [High-dimensional]-(./replication/simulation/figures/hd.ipynb) - [Colab]-(https://colab.research.google.com/drive/1i_NbMRESZTNSHsqRlnRu0GuPA658UT9W?usp=sharing)

  - [Resource Comparison]-(./replication/simulation/figures/resource.ipynb) - [Colab]-(https://colab.research.google.com/drive/1MAf9qRDnYtG9XnW-GzVzyldVtMk-qlC2?usp=sharing)
  - [Forest Plot]-(./replication/simulation/figures/forest_plot.ipynb) - [Colab]-(https://colab.research.google.com/drive/1sXzVkaF6_X4wSx_WgtR5PCCAFw8XMoqQ?usp=sharing)
  - [Additional Predictor]-(./replication/simulation/figures/lp.ipynb) - [Colab]-(https://colab.research.google.com/drive/1pfteqvgAbetdgRIWjoExQRYpYfT4x-q4?usp=sharing)
  - [Baseline Hazard]-(./replication/simulation/figures/baseline_hazard.ipynb) - [Colab]-(https://colab.research.google.com/drive/1PDp2G-ob1tjIlnh03j9TyoH7QlxDuGYM?usp=sharing)
- To replicate a similar analysis as in the UKB go to [Replicate Fake UKB]-(./replication/ukb)
  - [Fake Data Generation]-(./replication/ukb/00_fakedata.ipynb) - [Colab]-(https://colab.research.google.com/drive/1wT4pw2WEk6npzx7lrSaOjo5JUwTEfVXr?usp=sharing)
  - [Analysis]-(./replication/ukb/01_fakeanalysis.ipynb) - [Colab]-(https://colab.research.google.com/drive/1dP4TCF12Nx50bgn7GA2YkBNo9fAFbD2M?usp=sharing)

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ $<$ 1 minute
☐ 1-10 minutes
☐ 10-60 minutes
☒ 1-8 hours
☐ $>$ 8 hours
☐ Not feasible to run on a desktop machine, as described here:

Additional information (optional)

- The replciation information is also available on https://github.com/alexwjung/ProbCox/tree/main/paper/ProbCox (including hyperrefs and links for easier access to the corresponding folders and colabs)
- Rerunning all the simulations on a single desktop machine will take a considered amount of time. Therefore, we provide individual simulation runs (chosen by demand) that can be checked/compared to the results provided on https://github.com/alexwjung/ProbCox.
- The simulation results for the high-dimensional case can suffer from numerical instabilities, this happens for the particular prior specification of student(nu=1, s=0.001). With s $>$ 0.01 we find the result to stabilize much better, however, there is also a stronger regularization applied. Our replication results are not exact, however, differences are marginal and the overall results are the same.
- The fake simulation for the UKB data needs to write ~2GB of data. In the colab notebooks this would need to be written to the google drive.

Notes (optional)

References

Allen, Alina M, Terry M Therneau, Joseph J Larson, Alexandra Coward, Virend K Somers, and Patrick S Kamath. 2018. "Nonalcoholic Fatty Liver Disease Incidence and Impact on Metabolic Burden and Death: A 20 Year-Community Study." *Hepatology* 67 (5): 1726–36.

Crowley, John, and Marie Hu. 1977. "Covariance Analysis of Heart Transplant Survival Data." *Journal of the American Statistical Association* 72 (357): 27–36.

Laurie, John A, Charles G Moertel, Thomas R Fleming, Harry S Wieand, John E Leigh, Jebal Rubin, Greg W McCormack, James B Gerstner, James E Krook, and James Malliard. 1989. "Surgical Adjuvant Therapy of Large-Bowel Carcinoma: An Evaluation of Levamisole and the Combination of Levamisole and Fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic." *Journal of Clinical Oncology* 7 (10): 1447–56.

Loprinzi, Charles Lawrence, John A Laurie, H Sam Wieand, James E Krook, Paul J Novotny, John W Kugler, Joan Bartel, Marlys Law, Marilyn Bateman, and Nancy E Klatt. 1994. "Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group." *Journal of Clinical Oncology* 12 (3): 601–7.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *Plos Med* 12 (3): e1001779.

Therneau, Terry M, and Patricia M Grambsch. 2000. "The Cox Model." In *Modeling Survival Data: Extending the Cox Model*, 39–77. Springer.

Therneau, Terry M, and Thomas Lumley. 2015. "Package 'Survival'." *R Top Doc* 128 (10): 28–33.