PyMSM: Python package for Competing Risks and Multi-state models for Survival Data

Hagai Rossman[1], Ayya Keshet[1*] and Malka Gorfine[2]*

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

[2] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel

** Equal contribution*

Background and Motivation

Multi-state data are common, and could be used to describe trajectories in diverse health applications; such as describing a patient's progression through disease stages or a patient's path through different hospitalization states. When faced with such data, a researcher or clinician might seek to characterize the possible transitions between states, their occurrence probabilities, or to predict the trajectory of future patients - all conditioned on various baseline and time-varying individual covariates. By fitting a multi-state model, we can learn the hazard for each specific transition, which would later be used to predict future paths. Predicting paths could be used at a single patient level, for example predict how long until a cancer patient will be relapse-free given his current health status, or at what probability will a patient end a trajectory at any of the possible states; and at the population level, for example predicting how many patients which arrive at the emergency-room will need to be admitted, given their covariates.

Capabilities

PyMSM is a Python package for fitting multi-state models, with a simple API which allows user-defined models, predictions at a single or population sample level, and statistical summaries and figures.

Features of this software include:

- Fitting a Competing risks Multistate model based on various types of survival analysis (time-to-event) such as Cox proportional hazards models or machine learning models, while taking into account right censoring, competing events, recurrent events, left truncation, and time-dependent covariates.
- Running Monte-carlo simulations (in parallel computation) for paths emitted by the trained model and extracting various summary statistics and plots.
- Loading or configuring a pre-defined model and generating simulated data in terms of random paths using model parameters, which could be highly useful as a research tool.
- Modularity and compatibility for different time-to-event models such as Survival Forests and other custom ML models provided by the user.

The package is designed to allow modular usage by both experienced researchers and non-expert users. In addition to fitting a multi-state model for a given data - PyMSM allows the user to simulate trajectories, thus creating a multi-state data-set, from a predefined model. This could be a valuable research tool - both for sharing sensitive simulated individual data and as a tool for any downstream task which needs individual trajectories.

To the authors best knowledge, this is the first open-source multi-state model tool that allows fitting of such models while also dealing with important concepts such as right censoring, competing events, recurrent events, left truncation, and time-dependent covariates.

Usage examples

This project is based on methods first introduced during 2020 for predicting national hospitalizations in Israel. Important health policy applications based on these methods were built and used by government policymakers throughout the pandemic. For example:

1. Help assess hospital resource utilization (Roimi et. al JAMIA 2021 - https://doi.org/10.1093/jamia/ocab005).
2. Associations between high hospital load and excess deaths (Rossman & Meir et. al. Nature Communications 2021 - https://www.nature.com/articles/s41467-021-22214-z).

A similar R version of this package is available in Roimi et al., yet this is the first Python version to be released as an open-source package containing extended features and use cases.

Other usage examples are provided in the software package docs such as breast cancer state transitions (Rotterdam dataset), AIDs competing risk data, disease stage data from the European Society for Blood and Marrow Transplantation (EBMT) and COVID-19 national hospitalizations.

Link to Software

A documentation website is available here: https://hrossman.github.io/pymsm/

Documentation includes:

- A quickstart for beginners and method explanations.
- Usage examples for preparing a dataset, initiating a multistate model, model fitting, path sampling, examining a model, outputting summary statistics and plots and other custom options.
- Four full example notebooks of applying this package to health related tasks and real-world data.

We are excited to share this project with the MLHC community and hope others find it useful for their own applications.