SaddlePoint Signature™ Fact Sheet

SaddlePoint-Signature is a multi-core software pipeline designed primarily for predictive multivariate medical data analytics in the regime of high-dimensional covariates and/or undersampling.

1. General description

Version	SaddlePoint-Signature 2.8.5
Platforms	most versions of Windows, Unix/Linux and MacOS (tested on Windows 10, Ubuntu and MacOS Sierra)
Data	n samples, representing d covariates $(x_1, \ldots, x_d)$ and corresponding clinical outcomes outcomes are (t, r) (event time and type), or y (ordinal real-valued or discrete outcomes)
Deliverables	optimal covariate selection for predictive regression without overfitting ranking of the covariates in the optimal set optimal multivariate predictive models quantification of prediction performance of optimal model on unseen data optimised personal risk scores and treatment response scores various statistical and visualization analyses of the data

2. Detailed functionality

Operation	multi-threading with user controlled maximum number of threads command-line user interface test mode available, with more extensive outputs missing data entries coded as NA or na
Data types	survival data (with multiple risks and or censoring) discrete ordinal outcome classes general non-discrete ordinal outcomes data controlled cohort heterogeneity (latent classes)

Synthetic data	different outcome types
	controlled covariate distributions, with or without correlations
	controlled levels of integrity (covariate missingness)
	controlled event time statistics
	controlled censoring profiles
	  cliff-edge censoring
	  exponential censoring
	  block censoring
	visualization of properties of synthetic data
	  rank correlation versus Pearson correlation of covariates with outcome
	  rank correlation of covariates with outcome versus univariate regression parameters
	  Pearson correlation of covariates with outcome versus univariate regression parameters
	  preselection ROC curves

Data visualization
- tables of descriptive statistics of covariates and outcomes
- histograms of values of covariates and outcomes
- Pearson and rank correlations across covariates
- Pearson and rank correlations between covariates and outcome
- covariate-conditioned outcome statistics (Kaplan-Meier curves, class fractions)

Data preprocessing
- covariate multiplexing (inclusion of covariate products, all or selected)
- preselection of covariates
  - user determined by hand
  - automated, based on Pearson or rank correlation with outcome
  - automated, based on univariate regression
  - automated, based on fraction of missing values
  - read from previously saved file
- fixing of covariates (not to be removed in optimization)
- preselection of samples
  - user determined by hand
  - automated (based on fraction of missing values)
- randomization of outcomes
- linear normalization of covariates (to zero average and unit variance)

Regression pipelines
- batch loop of bootstrapping proportional hazards regressions (Cox or ordinal class)
- batch loop of bootstrapping parametric regressions
- automated class balancing for ordinal outcome classes
- protocol for adjusting regression due to informative missingness of covariates
- Bayesian priors
  - L_1 or L_2 with fixed width
  - L_1 or L_2 with automatically adapted width
- automated iterative covariate set reduction
  - values of hazard ratios
  - z-scores of regression parameters
  - advanced probabilistic criterion
- cross-validation
  - LOOCV or 50/50 division into training and validation sets
  - controlled number of randomizations per cycle
- final covariate set selection
  - maximum accuracy of prediction on unseen data
  - balanced criterion involving also minimum overfitting gap

Outputs
- optimal non-overfitting covariate set, with ranking, saved as 'covariate mask'
- association parameters, hazard ratios, confidence intervals, z-scores, p-values
- training and validation curves
- classification confusion table (for optimal set)
- personalized risk score formulas (for optimal set)
- risk score statistics (histograms) for present or new data sets
- treatment response score formulas (in case covariates include intention to treat)
- record of batch loop bootstrapping outcomes (covariates included, associations)
- outcome statistics stratified by risk scores, for present or new data sets