The `MetabolicSurv` R package

Predicting Risk Groups for Survival of Cancer Patients Using a Robust Metabolomic Signature

Olajumoke Evangelina Owokotomo[1], Ziv Shkedy[1], Adetayo Kasim[2]

[1] *Interversity Institute for Biostatistics and Statistical Bioinformatics, CenStat, Universiteit Hasselt, Belgium*
[2] *Wolfson Research Institute, Durham University, University Boulevard, Thornaby, Stockton-on-Tees, United Kingdom*

Interuniversity Institute for Biostatistics and statistical Bioinformatics

INTRODUCTION

- Recent developments in technology has lead to more and more increase in the availability of high throughput medical and biological datasets. In clinical practice, availability of this type of datasets, has lead to identifying signatures which can serves as biomarkers.
- In contrast to the ordinary survival prediction and classification of cancer patients based on prognostic factors measured at baseline alone, recent literature now focuses on integrating high dimensional data in addition to improve prediction and classification into high and low risk-group for survival.
- While methods for the development of high dimensional signatures for survival of cancer patients exist in various context in the literature software to conduct this analysis is not available. The `MetabolicSurv` aimed to close this gap and provides a user friendly data analysis tool, both for modelling and visualization for this type of applications using metabolomics data.

DATA STRUCTURE

The i^{th} observation is denoted by $(T_i, \delta_i, X_{i1}, \ldots, X_{ip}, P_1, \ldots, P_m)$

$$\begin{bmatrix} T_1 & \delta_1 \\ T_2 & \delta_2 \\ T_3 & \delta_3 \\ \vdots & \vdots \\ \vdots & \vdots \\ T_N & \delta_N \end{bmatrix} \quad \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ X_{31} & X_{32} & \cdots & X_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Np} \end{bmatrix} \quad \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ P_{31} & P_{32} & \cdots & P_{3m} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ P_{N1} & P_{N2} & \cdots & P_{Nm} \end{bmatrix}$$

Survival | Metabolomics profile | Prognostic factors

`MetabolicSurv` PACKAGE

Category	Functions	Description
Basic	MSpecificCoxPh	Metabolite by metabolite Cox proportional hazard analysis.
	SurvPcaClass	Classifier based on first PCA.
	SurvPlsClass	Classifier based on first PLS.
	Majorityvotes	Classifiction for Majority Votes.
	Lasoelacox	Wrapper function for glmnet.
	MSData	Generate Artificial Metabolic Survival Data.
Advance	CVLasoelacox	Cross Validations for Lasso Elastic Net and Classification.
	CVSim	Cross-validation for Top $K_1, \ldots, K_n$ metabolites.
	CVPcaPls	Cross-validations for PCA and PLS based methods.
	CvMajorityvotes	Cross-validation for majority votes.
	MetFreq	Frequency of Selected Metabolites from MSpecificCoxPh.
	QuantileAnalysis	Sensitivity of the quantile used for classification.
	Icvlasoel	Inner and outer cross-validations for shrinkage methods.
	DistHR	Null distribution of the estimated HR.
	SIMet	Sequentially increase the number of top K metabolites.

METHODOLOGY

Estimation of the metabolic risk score in the `MetabolicSurv` package: Data reduction methods, Univariate analysis and Majority voting.

$$\begin{aligned} h(t) &= h_o(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \gamma_1 P_1 + \ldots + \gamma_m P_m), \\ &= h_o(t)\exp(\mu.), \\ &= \beta \mathbf{U}(\mathbf{X}) + \Sigma_{m=1}^{m} \gamma_m P_m \end{aligned}$$

In the `MetabolicSurv` package, $\mathbf{U}(X)$ is a linear combination of the metabolic profile obtained by;

- Scores of the first component of a supervised PCA.
- Scores of the first component of a supervised PLS analysis.
- Linear combination of metabolites with coefficients using LASSO or Elastic Net
- A single metabolite profile.

Data reduction/Variable selection

$$\mathbf{R}(x) = \begin{cases} U(X_r)\beta_{PCA1}, & \text{first PCA is used,} \\ U(X_r)\beta_{PLS1}, & \text{first PLS is used,} \\ U(X_k)\beta_j, & \text{metabolite j is used,} \\ \sum_{j=1}^{J} X_j \beta_j^s, & \text{shrinkage methods are used and } J \ll p. \end{cases}$$

Let τ be a pre defined threshold, for example the median of $R_i(x)$, then $\mathbf{R}_i(x) > \tau$ patient i is labelled as having high risk for mortality, otherwise having low risk for mortality. Let $\mathbf{Z}_i$ be the indicator variable which is given by;

$$\mathbf{Z}_i = \begin{cases} 1, & \text{low risk,} \\ 0, & \text{high risk.} \end{cases} \implies \implies \implies \implies h(t) = \Sigma_{m=1}^{m} \gamma_m P_m + \lambda Z_i.$$

METHODOLOGY

Cross-validation

- Assess how results of the statistical analysis can be generalize to an independent data set.
- A small proportion of the dataset is used to validate the result from the bigger proportion that was used to fit the model and label the patients.
- Distribution of the risk-group HR on the training set and testing set is recorded.

Figure 1: The ℓ-fold cross-validation

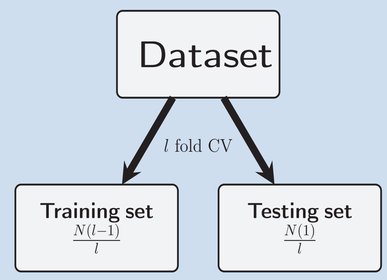
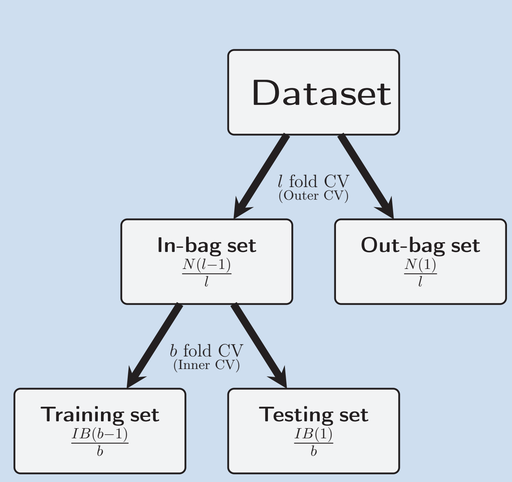


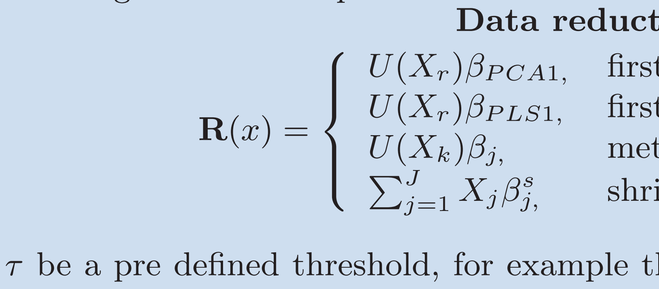
Figure 2: Inner & Outer cross-validation



SELECTED OUTPUT OF THE PACKAGE

PCA and PLS based method

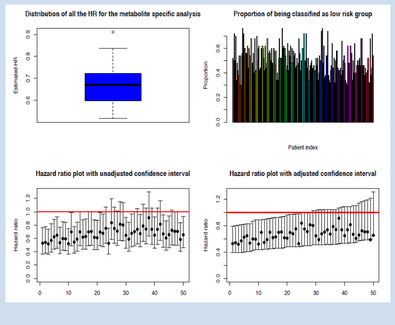
```
> Result <- SurvPcaClass(Survival = Survdata, Mdata = Metdata , Censor =
  Censordata, Reduce = TRUE, Select = 150, Prognostic = Progdata, Plots =
  TRUE, Quantile = 0.5)
> Result$KMplot;        Result$SurvBPlot
```



Feature Specific CoxPh Modelling

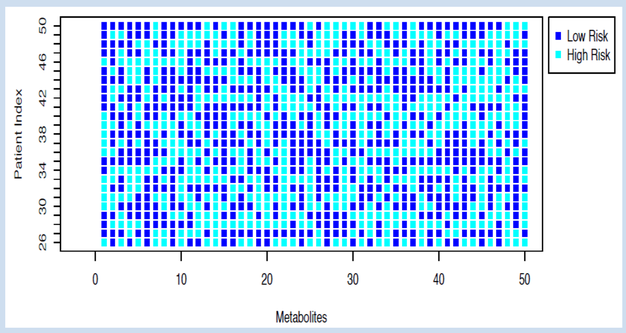
```
> Result1 <- MSpecificCoxPh(Survival = Survdata, Mdata = Metdata ,
  Censor = Censordata, Reduce = TRUE, Select = 50, Prognostic = Progdata,
  Quantile = 0.5)
> summary(Result1)
```

```
Number of Metabolites used:  50
Top 15 Metabolites out of  50
Estimated HR for the low risk group
Metnames        HR       LCI       UCI    FDRLCI    FDRUCI
1    Met386 0.5170718 0.3609665 0.7406870 0.3874103 0.7949484
2     Met76 0.5200284 0.3634770 0.7440074 0.3898663 0.7980241
3   Met2329 0.5261841 0.3676555 0.7530682 0.3938511 0.8067245
4    Met209 0.5283369 0.3677523 0.7590432 0.3937824 0.8127695
5    Met961 0.5357655 0.3744544 0.7665678 0.4003592 0.8195991
6   Met1055 0.5405980 0.3781031 0.7729272 0.4038715 0.8256034
7   Met2983 0.5453737 0.3818179 0.7789904 0.4074551 0.8312958
8   Met1034 0.5695780 0.3969927 0.8171917 0.4216815 0.8680124
9   Met2103 0.5845361 0.4090338 0.8353405 0.4332659 0.8848278
10  Met1286 0.5884136 0.4117891 0.8407958 0.4358754 0.8899754
11  Met2351 0.5886269 0.4116492 0.8416913 0.4357104 0.8908887
12  Met1913 0.5934799 0.4168376 0.8449775 0.4408139 0.8935802
13  Met2881 0.5981358 0.4201883 0.8514431 0.4439852 0.8996637
14  Met1464 0.6064816 0.4257121 0.8640112 0.4491543 0.9115886
15  Met2778 0.6093505 0.4276222 0.8683086 0.4509414 0.9156594
```



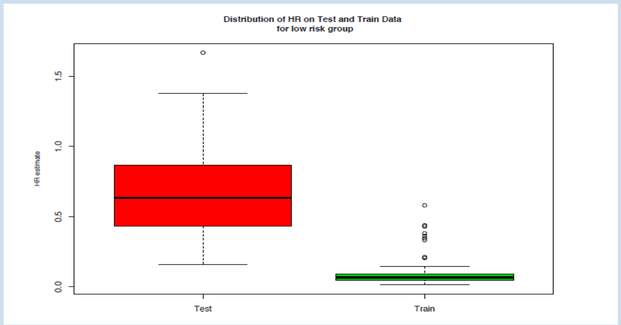
Majority Votes

```
> Result2=Majorityvotes(Result1,Progdata,Survdata, Censordata,J=2)
```



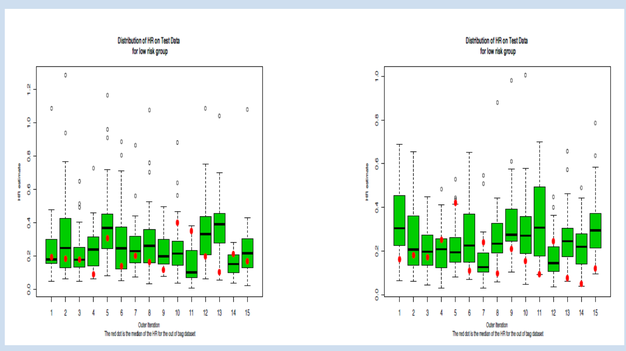
Cross validations for Lasso and ElasticNet

```
> Result3 = CVLasoelacox(Survdata,Censordata, Mdata = Metdata,Progdata,
  Quantile = 0.5, Metlist = NULL,Standardize = TRUE, Reduce=TRUE,
  Select=150, Alpha = 1,Fold = 4,Ncv = 100,nlambda = 100)
> summary(Result1)
```



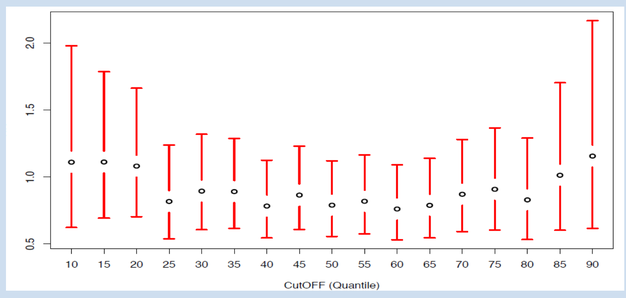
Inner & Outer Cross validations for Lasso

```
> Result4 = Icvlasoel(Survdata, Censordata, Progdata,Metdata, Fold =
  3,Ncv = 15, Nicv = 30, Alpha = 1, TopK = Mets, Weights = FALSE)
> Result5 = Icvlasoel(Survdata, Censordata, Progdata, Metdata, Fold =
  3,Ncv = 15, Nicv = 30, Alpha = 1, TopK = Mets, Weights = TRUE)
> plot(Result4, type = 1);        plot(Result5, type = 1)
```



Sensitivity Analysis of the Cutoff value

```
> Result6 <- QuantileAnalysis(Survdata,Metdata, Censordata,Reduce=TRUE,
  Select=150, Prognostic=Progdata,Plots = TRUE,DimMethod="PCA",Alpha=1)
```



DOWNLOAD

- `MetabolicSurv`: `https://cran.r-project.org/web/packages/MetabolicSurv/index.html`.
- More details available in `https://github.com/OlajumokeEvangelina/MetabolicSurv`.

REFERENCES

[1] O. E. Owokotomo and Z. Shkedy, *MetabolicSurv: A Biomarker Validation Approach for Classification and Predicting Survival Using Metabolomics Signature*, 2019. R package version 1.0.0.