`plsRcox`, Cox-Models in a high dimensional setting in R

Frédéric Bertrand[1,2,⋆], **Philippe Bastien**[3], **Nicolas Meyer**[1,4], **Myriam Bertrand**[1,2]

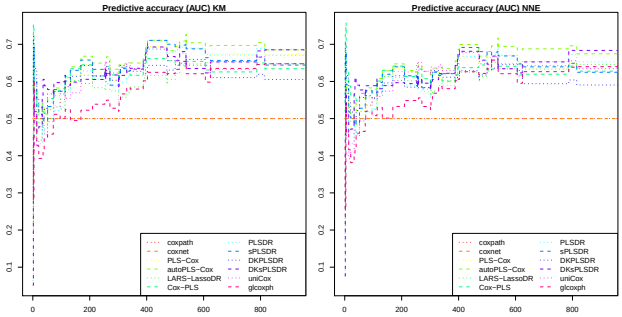
1. Université of Strasbourg
2. Centre national de la recherche scientifique
3. L'Oréal Recherche et Développement
4. Institut national de la santé et de la recherche médicale

⋆Contact author: fbertran@math.unistra.fr

Keywords: Partial least squares regression, Cox models, survival analysis, high dimensional data, R

A vast literature from the last decade is devoted to relating gene profiles and subject survival or time to cancer recurrence. Biomarker discovery from high-dimensional data, such as transcriptomic or SNP profiles, is a major challenge in the search for more precise diagnoses. The proportional hazard regression model suggested by Cox, [1], to study the relationship between the time to event and a set of covariates in the presence of censoring is the most commonly used model for the analysis of survival data. However, like multivariate regression, it supposes that more observations than variables, complete data, and not strongly correlated variables are available. In practice when dealing with high-dimensional data, these constraints are crippling. Collinearity gives rise to issues of overfitting and model mis-identification. Variable selection can improve the estimation accuracy by effectively identifying the subset of relevant predictors and enhance the model interpretability with parsimonious representation. In order to deal with both collinearity and variable selection issues, many methods based on Lasso penalized Cox proportional hazards have been proposed since the reference paper of Tibshirani, [3]. Regularization could also be performed using dimension reduction as is the case with PLS regression. We propose two original algorithms named sPLSDR and its non linear kernel counterpart DKsPLSDR, by using sparse PLS regression (sPLS) based on deviance residuals. We compared their predicting performance with state of the art algorithms based on reference benchmark datasets.

As sPLSDR and DKsPLSDR compare favorably with other methods in their computational time, prediction and selectivity, as indicated by results based on benchmark datasets, see Figure below, we view them as a useful addition to the toolbox of estimation and prediction methods for the widely used Cox's model in the high-dimensional and low sample size settings.



Model prediction accuracy comparision using iAUC, [2].

References

[1] Cox, D.R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society B*, **74**, 187–220.

[2] Heagerty, P.J., and Zheng, Y. (2005) Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, **61**(1), 92-=105.

[3] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.