HDcox: Unravelling the Invisible in High-Dimensional and Multi-Omic data sets in Survival Analyses

Genomics of Gene Expression Lab

Pedro Salguero[1], Ana Conesa[2] and Sonia Tarazona[1]

[1] Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Spain
[2] Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain

Objectives

The **main objective** of this thesis is to **develop** and validate **statistical methods** for **survival analysis** in **high-dimensional and multi-omic scenarios**, and create an R package for sharing these methodologies with the scientific community as an open-source tool.

The specific objectives are:

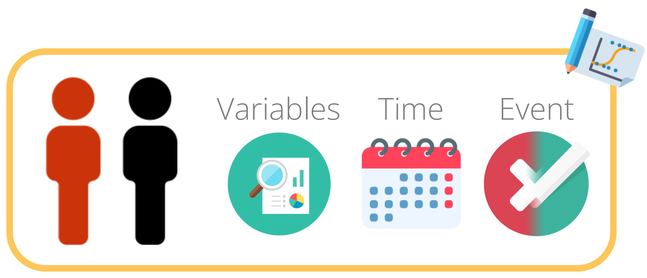
1. **Develop new algorithms for survival analysis on high dimensional data** based on Partial Least Squares[1] (PLS) and Cox Proportional Hazards[2] models.
2. **Enhance Interpretability**: Provide essential insights into the roles of predictors and how they contribute to survival outcomes instead of focusing only in prediction accuracy.
3. **Evaluate Predictive Performance** by using multiple metrics that assess the effectiveness of trained models on new observations, thereby providing an indication of the model's real-world applicability and generalizability.
4. **Develop Multi-Omic algorithms:** Adapt the developed methods for dealing with multi-omic data-sets in survival analysis.
5. **International Collaboration and Software Development:** Produce a series of open-source functions that the scientific community can freely use, thereby contributing to the broader field of survival analysis and precision medicine.

What is a Survival Analysis?

Reach the event of study

Do not reach the event of study

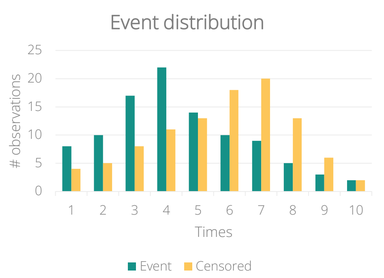
- Lost to follow-up
- Withdraw from the study
- ...

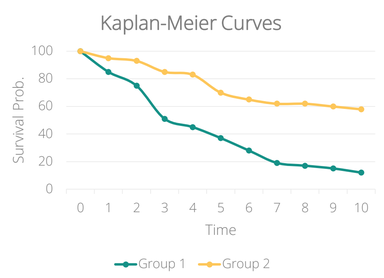


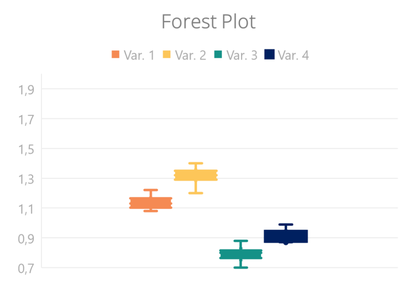
Assumptions | Significant Variables | Visualize Results

Classical Models are not able to work with high dimensional data (more variables than events) or suffer from multicollinearity problems.

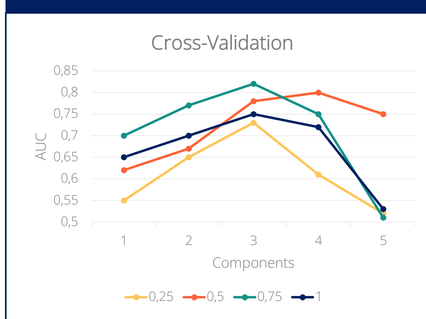
Classical Data Visualization

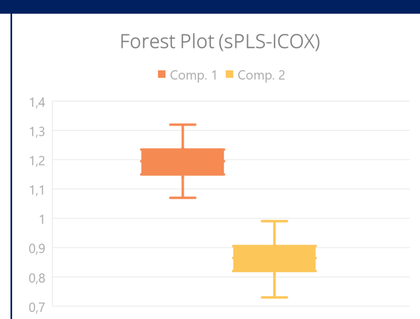


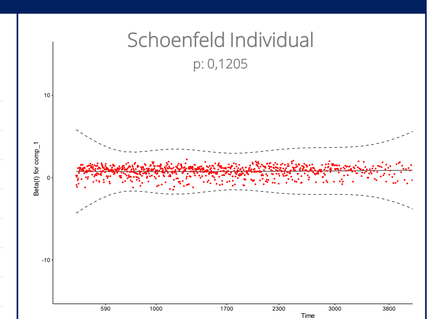


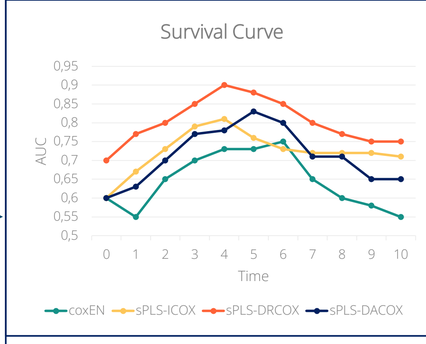


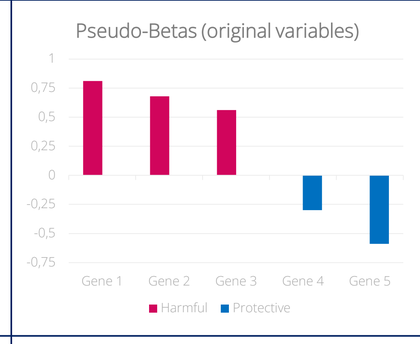
Understanding the survival models

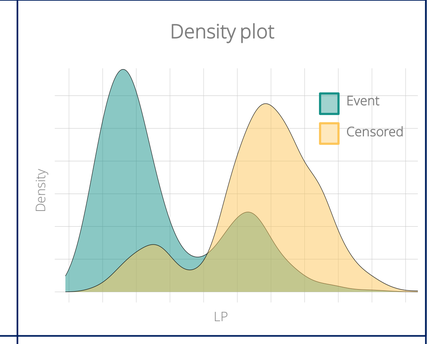


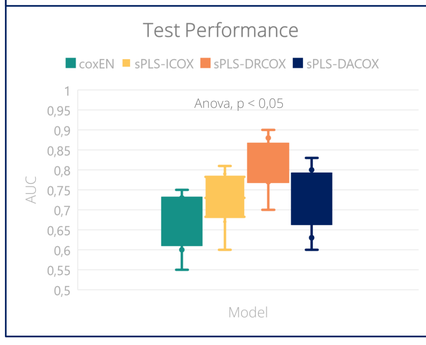


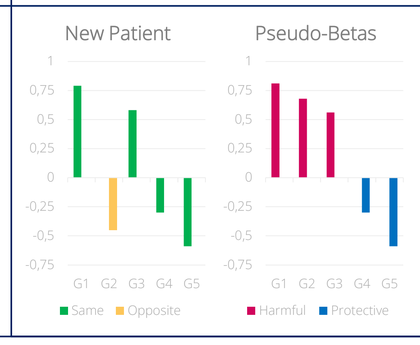


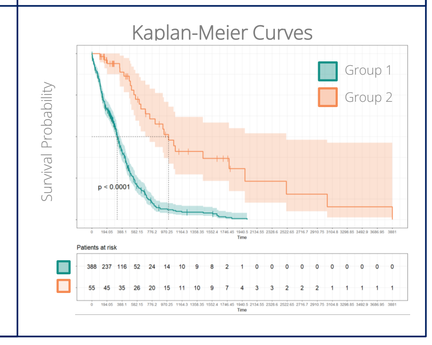










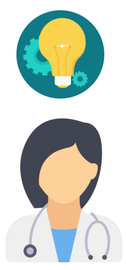


What is HDcox?

HDcox

- From high dimensional or multi-omic data

- It allows survival analysis based on PLS-COX



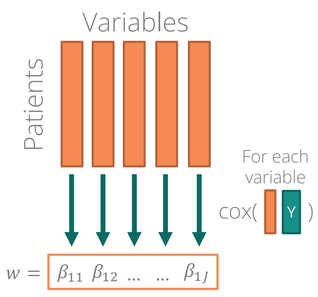
- Understanding the results

Analyses

(s)PLS Approaches

(s)PLS-ICOX[3]

- Change the internal PLS regression by multiple survival Cox regressions.



(s)PLS-DRCOX[4]

- Use Deviance Residuals from a NULL Cox model to reduce dimensionality.

DR = DR(cox(NULL, Y))

PLS(X , Y)

(s)PLS-DACOX

- Change the internal PLS regression by multiple survival cox regressions.

Y = Y censored/event

PLSDA(X , Y)

PLS-COX Models reduce variable dimensionality into PLS components. After computing the model, a survival Cox model is computed using the new matrix as input.

Multi-Omic Approaches

Single Block Approach	• Performs individual analysis per omic and then integrates PLS components.	(s)PLS-ICOX and (s)PLS-DRCOX
Multiple Block Approach	• Performs MB.(s)PLS-COX models (all omics simultaneously).	(s)PLS-DRCOX and (s)PLS-DACOX

Core Principles of HDcox

✓ **Interpretability:** Unlike many Machine Learning techniques that focus purely on prediction accuracy, HDcox emphasizes comprehensibility. The package is designed to provide insight into the underlying structure of the data, interpreting results in the context of original predictors. This enhances understanding of how the model functions and the specific role of individual variables in predicting outcomes.

✓ **Robustness:** HDcox is equipped with a wide array of features for model optimization, including an extensive cross-validation function that facilitates the accurate selection of PLS components and variable selection penalty. To generate reliable results, the package ensures adheres to key survival analysis principles such as the proportional hazard and event per variable (EPV).

✓ **Predictive Performance:** Multiple evaluation metrics such as C-Index, Brier Score, and AUC have been implemented for effective model comparison. Additionally, HDcox can analyze unique observations to identify which variables might exhibit a harmful or preventive influence. It also computes the optimal cutpoint for data segregation, a feature that can be visualized via Kaplan-Meier plots and applied to new data.

This focus on predictive accuracy, combined with a commitment to interpretability and robustness, makes HDcox a comprehensive tool for survival analysis in high-dimensional and multi-omic data.

Available at:
github.com/ConesaLab/HDcox

1. Cox, D.R. (1972), Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34: 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

2. Höskuldsson, A. (1988), PLS regression methods. J. Chemometrics, 2: 211-228. https://doi.org/10.1002/cem.1180020306

3. Bastien, P., Vinzi V. E. and Tenenhaus M. (2005) PLS generalised linear regression. Computational Statistics & Data Analysis, Volume 48: 17-46. https://doi.org/10.1016/j.csda.2004.02.005

4. Bastien, P., Bertrand F. Meyer N. and Maumy-Bertrand M. (2015) Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics, Volume 31: 397-404. https://doi.org/10.1093/bioinformatics/btu660