Network-based sparse modeling of breast invasive carcinoma survival data

André Veríssimo, Eunice Carrasquinha and Marta B. Lopes
IDMEC, Instituto Superior Técnico

Arlindo L. Oliveira
INESC-ID, Instituto Superior Técnico
Universidade de Lisboa

Marie-France Sagot
ERABLE, Inria
Université de Lyon

Susana Vinga
IDMEC, Instituto Superior Técnico
Universidade de Lisboa

***Abstract*—Learning survival models from oncological data has now become a major challenge due to the significant increase of molecular information. The inherent high-dimensionality of these datasets, where the number of features largely exceeds the number of observations, leads to ill-posed inverse problems and, consequently, to models that often lack interpretability.**

In order to tackle this problem, regularized optimization has emerged as a promising approach, allowing to impose constraints on the structure of the solutions, these include sparsity, for example, using LASSO, or other penalizing functions that use network-based information if the features have a graph-based configuration. We compared how different sparse methods perform when applied to a Breast Invasive Carcinoma dataset and how introducing network knowledge impact model prediction. These include Elastic Net and their coupling with DEGREECOX. The results regarding the concordance c-index show an improvement when network information is included, whereas the log-rank tests on the separation between high and low-risk patients exhibit a decrease in performance. It is expected that the obtained models can further support clinical decision and prognostic assessment of oncological patients.

I. INTRODUCTION

Statistical learning of oncology data constitutes a challenging task, requiring methods that should lead to accurate results but also to interpretable models. This is a key issue since they can be used in the decision process of clinicians when planning therapeutic strategies.

Breast Invasive Carcinoma (BRCA) is one of the most common types of cancer with more than $1,300,000$ new cases every year and $450,000$ deaths worldwide. Research on this type of cancer has produced very large datasets with an increasing amount of molecular data per patient. Such data typically exhibits highly correlated features, as the cancer cells disrupt regulatory networks that can lead to a high variability in the gene expression. This makes it difficult to identify the causal gene or mutation that started the process.

Restricting the solution space is a crucial step to tackle the optimization problem as the features greatly outnumber the observations. Some approaches that were developed in order to impose sparsity include LASSO[1], Elastic Net[2] and OSCAR[3]. A complementary approach to these sparse models is to use additional information that has less variability and can lead to a more robust and generalizable solution. Clustering or group organization has shown good results in the literature with Group LASSO[4] and Overlapping Group LASSO[5]. More recently other approaches have used functional, relationships or introduce network information as an external knowledge. Networks that have been used in cancer include protein-protein interactions, metabolic pathways, co-expression and functional maps. In particular, we have proposed DEGREECOX[6], based on the degree of each feature in the network. In this work, we apply existing methods in the literature to the BRCA dataset and perform a comparison of the prediction capabilities of the different models.

II. DATA DESCRIPTION

The dataset used for this study includes 1047 BRCA cases with transcriptional and clinical data, avaliable at The Cancer Genome Atlas (TCGA) data portal (https://gdc-portal.nci.nih.gov/). The transcriptomics data that were used in this study are: (1) clinical physiological information, such as age, gender, ethnicity, date of diagnosis, date of last follow-up, vital status, date of death; and (2) gene expression levels measured by fragments per kilobase per million normalized by the upper quartile (FPKM-UQ).

III. METHODS

In this context, we analysed the data using survival analysis, which involves modeling the time to an event of interest, by uncovering the relationships between the given covariates and time distributions[7]. The Cox proportional hazard model[7] is used to model these relationships and has been widely applied in oncology. This model has been extended to include sparse regularization. We have used R software packages for the analysis.

We applied six survival models with different types of regularization to the BRCA dataset: LASSO[1], Elastic Net[2] with three values for parameter λ (0.3, 0.5 and 0.7), and DEGREECOX[6]. The predictive performances of the models were compared using the Concordance index (c-index) and Kaplan-Meier (KM) model through log-rank tests.

IV. RESULTS

The models selected are able to greatly reduce the number of features from the original $55,682$ features to a subset of 7 to 40 with a good fit. From the full dataset, we used 889 cases to estimate the model's parameters using 10-fold cross-validation, while the remainder 158 cases were used as the test set with results being reported in this section. The KM model results were determined by using the estimated coefficients l to assert the prognostic risk of a patient. The patients are divided in two groups, high and low risk, and the log-rank test is performed to determine if the separation of the two groups is statistically significant. The results of the KM curves show that classical sparse models have better results that models that use network information. None of the network-based models were successful in separating patients in two categories, while the LASSO and two Elastic-net models ($\alpha = 0.5, 0.7$) have p-values below 0.05. The c-index is a relative measure that compares if the survival time of all permissible pairs of individuals follows their hazard relative risk. This measure showed opposite results from KM, as methods that used network information performed between 4.5% and 6.0% better than classical sparse models.

V. CONCLUSIONS

Sparse models perform very well in this dataset, being able to select a very small subset of features from the original $55,682$ features while still maintaining a good fit. The results show that these methods are highly competitive for the analysis of high-dimensional oncological data and that the network-based models would benefit from a network that is independent from the data.

REFERENCES

[1] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[2] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/abstract

[3] H. D. Bondell and B. J. Reich, "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, vol. 64, no. 1, pp. 115–123, Mar. 2008. [Online]. Available: http://doi.wiley.com/10.1111/j.1541-0420.2007.00843.x

[4] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00532.x/abstract

[5] L. Jacob, G. Obozinski, and J.-P. Vert, "Group Lasso with Overlap and Graph Lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 433–440. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553431

[6] A. Veríssimo, A. L. Oliveira, M.-F. Sagot, and S. Vinga, "DegreeCox a network-based regularization method for survival analysis," *BMC Bioinformatics*, vol. 17, no. S16, pp. 109–121, Dec. 2016. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1310-4

[7] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, Jan. 1972. [Online]. Available: http://www.jstor.org/stable/2985181

TABLE I: C-Index

	with Degree	*without Degree*
L1	0.785279	0.725440
0.5 L1 + 0.5 L2	0.788721	0.740584
0.3 L1 + 0.7 L2	0.800865	0.747124
0.7 L1 + 0.3 L2	0.788081	0.743977

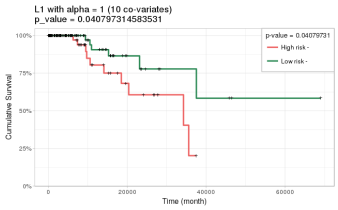

Fig. 1: Lasso (L1)

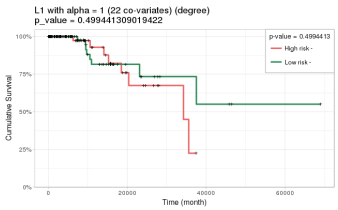

Fig. 2: Lasso with Degree Penalization (L1)

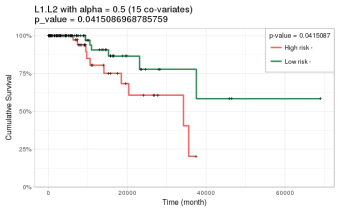

Fig. 3: Elastic Net (0.5 L1 + 0.5 L2)

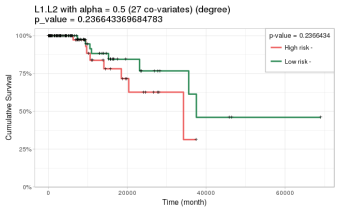

Fig. 4: Elastic Net with Degree Penalization (0.5 L1 + 0.5 L2)

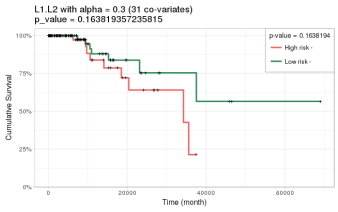

Fig. 5: Elastic Net (0.3 L1 + 0.7 L2)

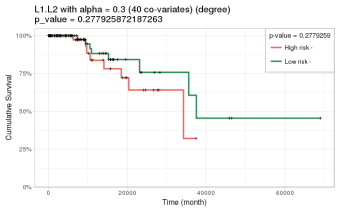

Fig. 6: Elastic Net with Degree Penalization (0.3 L1 + 0.7 L2)

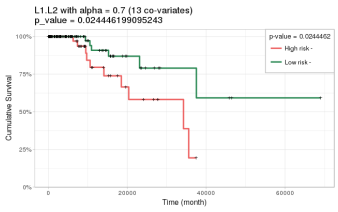

Fig. 7: Elastic Net (0.7 L1 + 0.3 L2)

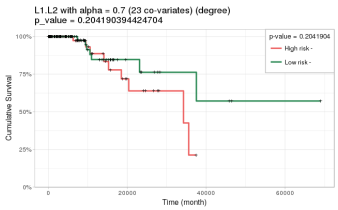

Fig. 8: Elastic Net with Degree Penalization (0.7 L1 + 0.3 L2)