Breast Cancer Survival Prediction

Statistical Methods for High Dimensional Data

Project report

Caltran Lorenzo
Careddu Luca
Francario Felice
Mariethoz Jean Zacharie

Department of Mathematics "Tullio Levi-Civita"
University of Padua
Padua, Italy

Academic year 2023/2024

Contents

1 Introduction

In this work, we exploit Survival analysis methods and Shrinkage methods (and a combination of both) to predict the survival probability of patients who underwent breast cancer surgery using their clinical and genetic information. The dataset is the ”Molecular Taxonomy of Breast Cancer International Consortium” (METABRIC) and contains 31 clinical variables, 489 gene variables, and 173 mutated gene variables (totaling 693 variables) of 1904 primary breast cancer samples. An explanation is provided for each clinical variable in the table 1.

Table 1: Clinical variables

Name	Type	Description
patient_id	object	Patient ID
age_at_diagnosis	float	Age of the patient at diagnosis time
type_of_breast_surgery	object	Breast cancer surgery type: 1- MASTECTOMY, which refers to a surgery to remove all breast tissue from a breast as a way to treat or prevent breast cancer. 2- BREAST CONSERVING, which refers to a surgery where only the part of the breast that has cancer is removed
cancer_type	object	Breast cancer types: 1- Breast Cancer or 2- Breast Sarcoma
cancer_type_detailed	object	Detailed Breast cancer types: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer

Continued on next page

Table 1: Clinical variables (Continued)

cellularity	object	Cancer cellularity post chemotherapy, which refers to the amount of tumor cells in the specimen and their arrangement into clusters
chemotherapy	int	Whether or not the patient had chemotherapy as a treatment (yes/no)
pam50_._claudin-low_subtype	object	Pam 50: is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell–cell adhesion genes, high expression of epithelial–mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns
cohort	float	Cohort is a group of subjects who share a defining characteristic (It takes a value from 1 to 5)

Continued on next page

Table 1: Clinical variables (Continued)

er_status_measured_by_ihc	float	To assess if estrogen receptors are expressed on cancer cells by using immunehistochemistry (a dye used in pathology that targets specific antigen, if it is there, it will give a color, it is not there, the tissue on the slide will be colored) (positive/negative)
er_status	object	Cancer cells are positive or negative for estrogen receptors
neoplasm_histologic_grade	int	Determined by pathology by looking the nature of the cells, do they look aggressive or not (It takes a value from 1 to 3)
her2_status_measured_by_snp6	object	To assess if the cancer positive for HER2 or not by using advanced molecular techniques (Type of next generation sequencing)
her2_status	object	Whether the cancer is positive or negative for HER2
tumor_other_histologic_subtype	object	Type of the cancer based on microscopic examination of the cancer tissue (It takes a value in 'Ductal/NST', 'Mixed', 'Lobular', 'Tubular/ cribriform', 'Mucinous', 'Medullary', 'Other', 'Metaplastic')
hormone_therapy	int	Whether or not the patient had hormonal as a treatment (yes/no)
inferred_menopausal_state	object	Whether the patient is post menopausal or not (post/pre)

Continued on next page

Table 1: Clinical variables (Continued)

integrative_cluster	object	Molecular subtype of the cancer based on some gene expression (It takes a value from ’4ER+’, ’3’, ’9’, ’7’, ’4ER-’, ’5’, ’8’, ’10’, ’1’, ’2’, ’6’)
primary_tumor_laterality	object	Whether it is involving the right breast or the left breast
lymph_nodes_examined_positive	float	To take samples of the lymph node during the surgery and see if there were involved by the cancer
mutation_count	float	Number of gene that has relevant mutations
nottingham_prognostic_index	float	It is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumour; the number of involved lymph nodes; and the grade of the tumour
oncotree_code	object	The OncoTree is an open-source ontology that was developed at Memorial Sloan Kettering Cancer Center (MSK) for standardizing cancer type diagnosis from a clinical perspective by assigning each diagnosis a unique OncoTree code
overall_survival_months	float	Duration from the time of the intervention to death
overall_survival	object	Target variable whether the patient is alive or dead
pr_status	object	Cancer cells are positive or negative for progesterone receptors

Continued on next page

Table 1: Clinical variables (Continued)

radio_therapy	int	Whether or not the patient had radio as a treatment (yes/no)
3-gene_classifier_subtype	object	Three Gene classifier subtype It takes a value from 'ER-/HER2-', 'ER+/HER2- High Prolif', 'ER+/HER2- Low Prolif', 'HER2+'
tumor_size	float	Tumor size measured by imaging techniques
tumor_stage	float	Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread
death_from_cancer	object	Can take on the three values: 'Died from cancer', 'Died from other causes', 'Living'

2 Exploration Analysis and Data Preprocessing

2.1 Exploration Analysis

The mutated genes columns involved mostly zeros so we have taken advantage of this and stored the dataframe in a sparse matrix to later achieve faster cross-validations when training models with penalizations. Figure 1 gives an idea of the sparsity in the mutated genes columns.

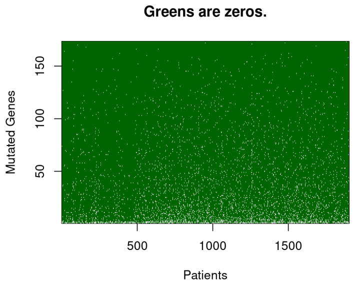
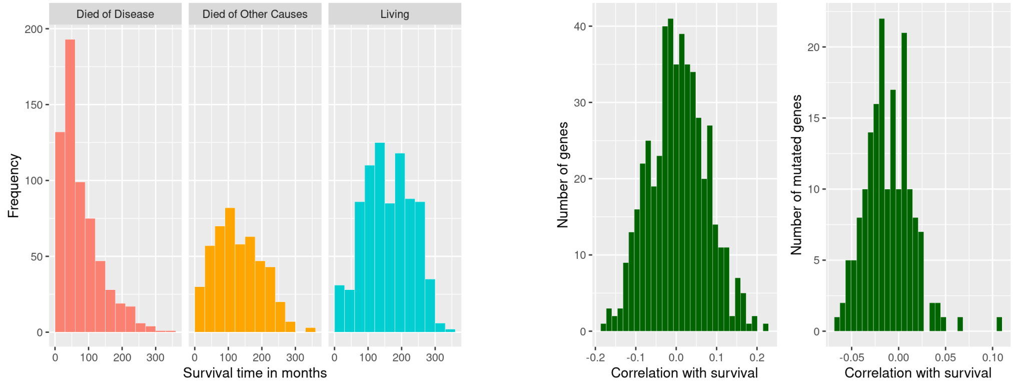


Figure 1: Sparsity in the mutated gene columns

For survival analysis (w.r.t. the overall_survival_months variable), an introductory test has been to study the frequencies of survival for each outcome of the death_from_cancer variable (which can take on three values as discussed before) and results are shown in figure 2(a) while for binary classification tasks (w.r.t. the overall_survival variable) the linear correlations between genes and response variables had to be verified and results are shown in figure 2(b).



(a) Survival Frequencies

(b) Correlation Frequencies

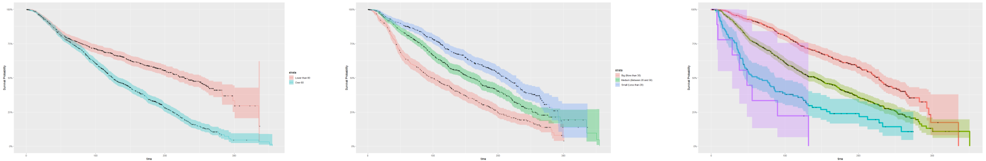
Figure 2: Frequencies

2.2 Data Preprocessing

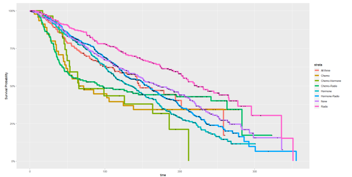
The dataset was already in good shape but for missing values in the numeric columns tumor_stage (501), neoplasm_histologic_grade (72), mutation_count (45), and tumor_size (20) which we decided to fill, for the sake of simplicity, with the means of the corresponding columns. Missing values found in some of the char columns were left unchanged and treated as factors during modeling. Minor fixes such as removing useless rows/columns were done without impacting the performance.

3 Survival Analysis

In the first part of the section we want investigate how some of the 29 clinical variables influence the survival probability.

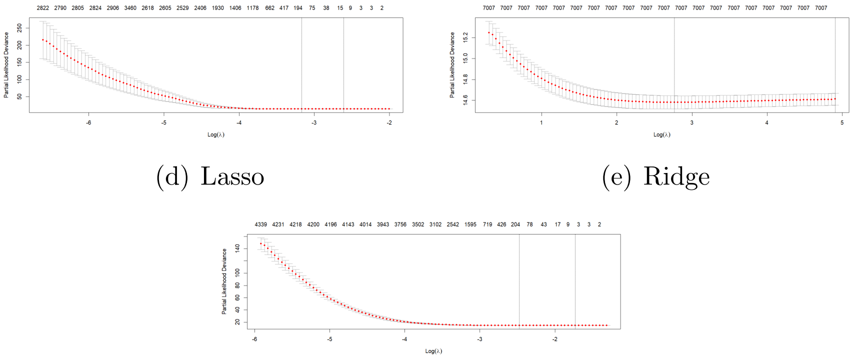


(a) KM plot for tumor stage (b) KM plot for tumor size (c) KM plot for tumor stage



From this last plot, it appears that Chemotherapy and the combination of Chemotherapy and Radiotherapy are the most effective forms of therapy. We also observe is that some of the more aggressive forms of therapy guarantee a lower survival probability. The reason why this happens is because they are used for older patients and with bigger sized tumors, all features that affect negatively the survival probability.

In next part of the project we extract the genomic part of the dataset to predict the variable ”overall survival months” using the Cox proportional Hazard model with three different types of regularization: Lasso, Ridge and Elastic Net.



(d) Lasso

(e) Ridge

(f) Elastic net (α=0.5)

As we observe, the Lasso and Elastic Net models yield an optimal value of λ very close to 0 (respectively 0.04216945 and 0.08433889), with 143 and 151 non-zero variables, so they are fairly regularized models. On the other hand, the Ridge regularized model gives a value λ=15.87653 with 7007 non-zero variables, so a much less regularized model. To select the best model we calculate the concordance indexes and they are 0.5974989 for the Lasso model, 0.6355165 for Ridge and 0.5992224 for Elastic Net, so according to this metric, the best model is the Cox Hazard proportional model with Ridge regularization.

4 Logistic Regression and Support Vector Machines (SVM)

In this section, we analyze different logistic regression and SVM models in predicting the `overall_survival` of a patient.

4.1 Logistic Ridge Regression

We choose the λ regression value by finding the model with the the minimum cross-validation error using cv.glmnet. The minimum cross-validation error is obtained with $\lambda = 0.191791$. Out of the 6906 estimated coefficients, 5894 are found to be significant using the bootstrap procedure. The overall accuracy of the best fitted model applied on a test set is 70.87%.

4.2 Logistic Lasso Regression

We find the best model by cross-validation (cv.glmnet) on the training set, and choosing $\lambda = 0.02772487$ using the 1 standard error rule corresponding to the simplest model. There are 38 non-zero coefficients with this model. Using the bootsrap procedure we find that only 3 significant($p < 0.05$) coefficients: `age_at_diagnosis`,`lymph_nodes_examined_positive`, `overall_survival_months`. The overall accuracy of the model on a test set is 75.85%.

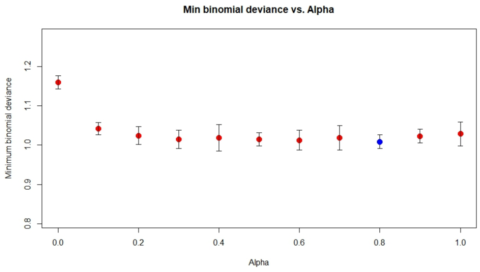


Figure 3: Minimum cross-validation errors of the elastic net model as a function of α.

4.3 Elastic Net

We apply an elastic net model with α=0.8 obtained by finding the minimum cross-validation error over a sequence of 10 values of between 0 and 1. We choose lambda λ= 0.02848036 by the 1 standard error rule. There are 64 non-zero coefficients. The overall accuracy of the model applied on a test set is 76.38%.

4.4 Sparse SVM

We find the best model by cross-validation on the training set, and choosing λ= 0.1044995 using the 1 standard error rule corresponding to the simplest model. There are 17 non-zero coefficients. Appyling the bootstrap approach we find only 3 significant coefficients: `ccnd2,tgfbr2,lama2`. The overall accuracy of the model applied on a test set is 69.55%.

4.5 Squared Hinge Loss

We find the best model by cross-validation on the training set, and choosing λ=0.1408054 using the 1 standard error rule corresponding to the simplest model. There are 21 non-zero coefficients. Applying the bootstrap approach with 500 bootstraps approach we find only 7 statistically significant coefficients with standard deviation different from 0: `(Intercept),age_at_diagnosis,cohort,lymph_nodes_examined_positive,mutation_count,overall_survival_months and tumor_size`. The overall accuracy of the model applied on a test set is 74.80%.

5 Group Lasso

In this section, we apply group lasso to predict the overall survival of a patient. To do so, we use the gglasso and the cv.gglasso functions from the ononymous package.
First of all we built a model on the groups of the categorical variable. That is, for each factor we considered its set of indicator variables as a group.
Then, starting from the same groups we considered two others groupings. For the second model, we considered the genetic attributes as a single group. Instead, for the last one we joined each gene with the same gene, mutated.

In the following plots, we can observe how the coefficients and the cross-validation errors changes with the increase of λ.

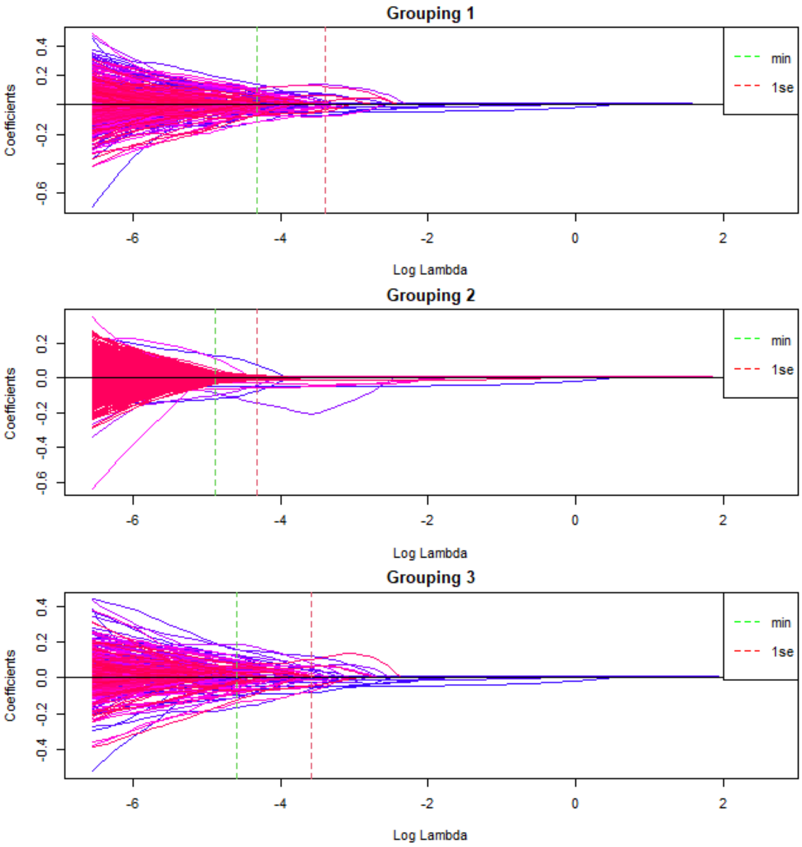


Figure 4: Coefficients as a function of λ. The green line indicates the lambda with the minimum cross-validation error, while the red line indicates the largest lambda with cross-validation error within one standard error of the minimum.

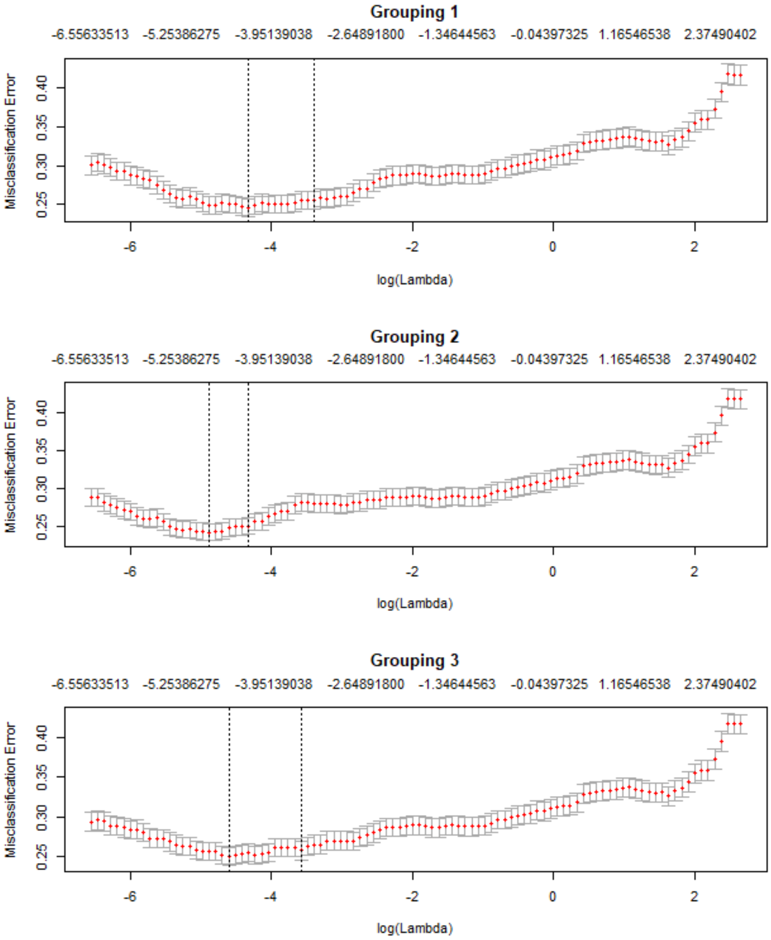


Figure 5: Cross-validation errors based on misclassification.

By using the lambdas that minimize the cross-validation error, we fitted the best model for each grouping and presented the results obtained in the following table. We can notice how the second grouping, the one where we consider all the genetic attributes, is the best in terms of performance.

Accuracy	Sensitivity	Specificity	N. Coef	Lambda	C-V Error
74.54%	66.87%	80.47%	82	0.01325308	0.2458909
76.64%	68.67%	82.79%	502	0.007583899	0.2419461
75.59%	69.88%	80.00%	92	0.01002547	0.2504931

Table 2: The rows represents the models in the same order in which we presented them. *N. coef* stands for *number of non-zero coefficients*

6 Adaptive Lasso

We tried to fit some models with adaptive lasso using the *two-stage approach*. We used the weights $w_j = |\tilde{\beta}_j|^{-\gamma}$, where $\tilde{\beta}$ are the initial estimates obtained using Ridge, Lasso and Elastic net with $\alpha = 0.5$, and γ is chosen in such a way as to minimize the cross-validation error (up to the second decimal).

We present in the following table the results, we can notice that using Lasso or Elastic net as initial estimates leds to a lower cross-validation error as compared to just applying Lasso, however the accuracies calculated on the test set are comparable, or even lower.

Initial Est.	Accuracy	Sensitivity	Specificity	N. Coef	C-V Error
—	76.12%	66.87%	83.26%	646	0.2465483
Ridge	63.25%	40.36%	80.93%	4137	0.2991453
Lasso	74.54%	68.67%	79.07%	245	0.1900066
Elastic Net	76.64%	71.69%	80.47%	77	0.2163051

Table 3: The first line represent the model obtained by applying just Lasso. *N. coef* stands for *number of non-zero coefficients*, the cross-validation error is calculated using misclassification.