Multiway multiblock logistic regression to classify liver tumors from MRI images

SELVESTREL Alexandre

November 25, 2024

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes

Supervisors : Arthur Tenenhaus, Laurent Lebrusquet

Medical partner : Henri Mondor hospital, radiologist: Sébastien Mulé

Liver tumors classification

6^{th} most widespread cancer and 4^{th} mortality cause by cancer

Classification:

- Hepatocellular Carcinoma (HCC): 75% of cases, resection often possible
- CCK = Cholangiocarcinoma (CCK): 6% of cases, resection difficult (possible 30% of cases)
- Others: benign (18% of cases) or Hepatoblastoma (1% of cases)

Difficulties for classification

- No perfect method using RMI images (contrast, shape, size, location): disagreement between radiologists
- High alpha-fetoprotein indicate HCC, but not always.
- Biopsy: invasive and potentially lethal (0.02% of patients)

Difficulties for classification

- No perfect method using RMI images (contrast, shape, size, location): disagreement between radiologists
- High alpha-fetoprotein indicate HCC, but not always.
- Biopsy: invasive and potentially lethal (0.02% of patients)

But a lot of clues even without Biopsy $\rightarrow$ Machine Learning

Specifically: adapting existing methods to the structure of the data (tensors)

Available data

- RMI images in 3D of liver tumors (arterial, portal, veinous -not used-, late)
- gender
- age at disease

Same variables extracted from each RMI images... 3 times

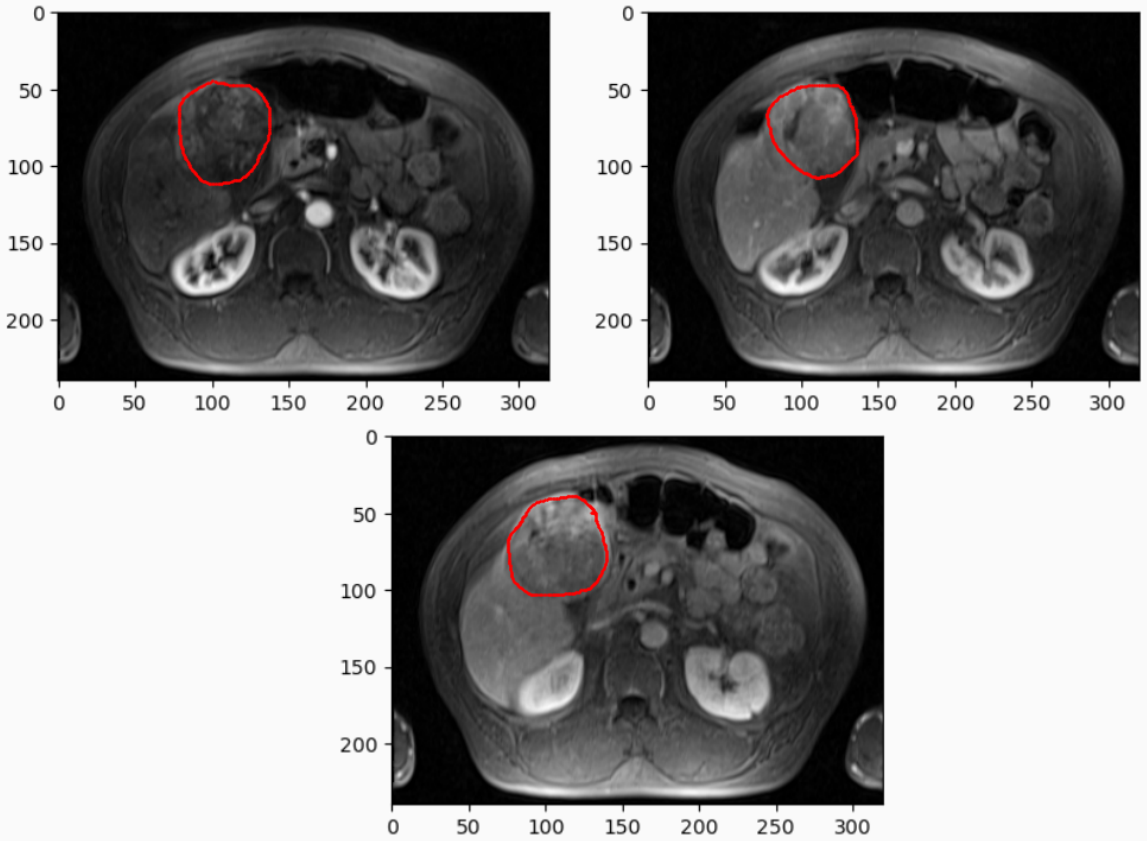


Figure 1: Example of RMI images of a HCC tumor (arterial, portal, late). More contrast in arterial

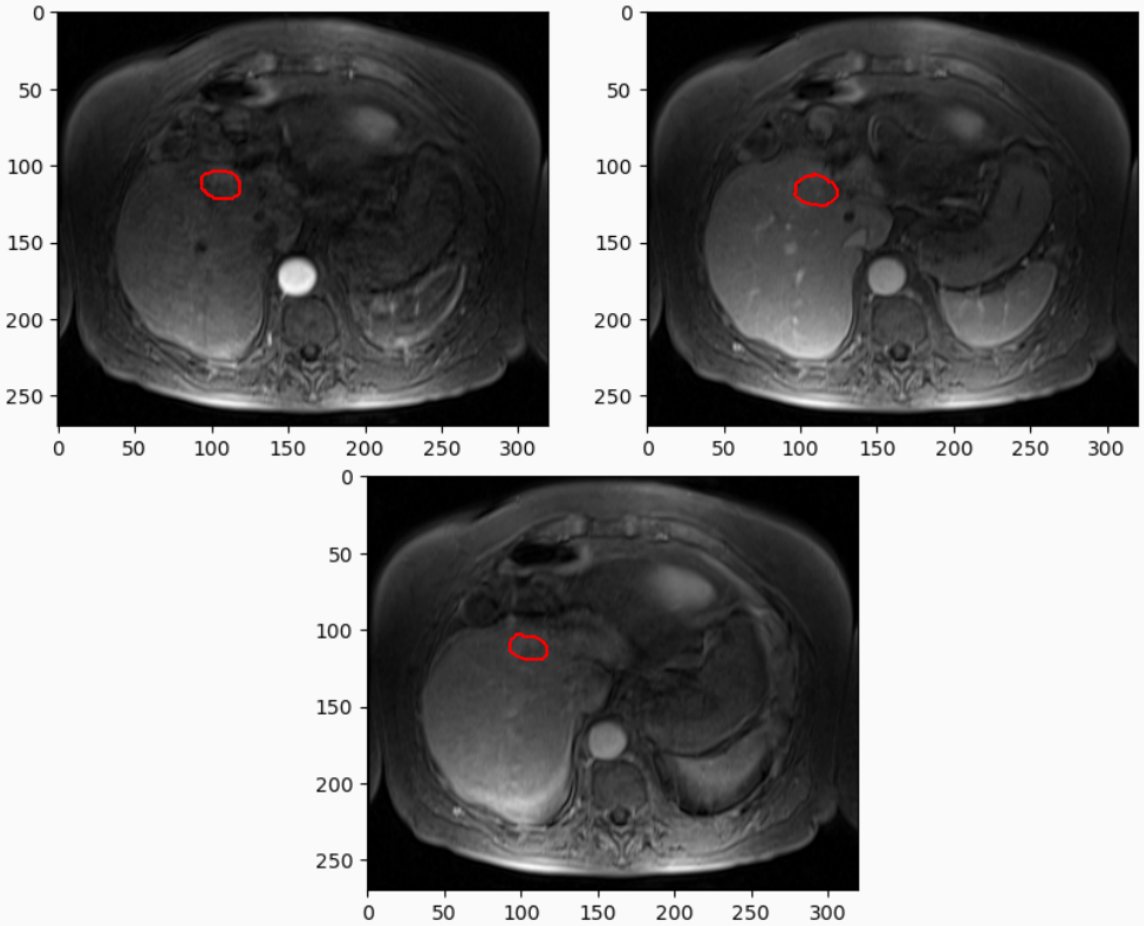


Figure 2: Example of RMI images of a CCK tumor (arterial, portal, late)

Correlation matrix of the texture (GLDM) coefficients

Strong correlations between the imaging times for a given variable

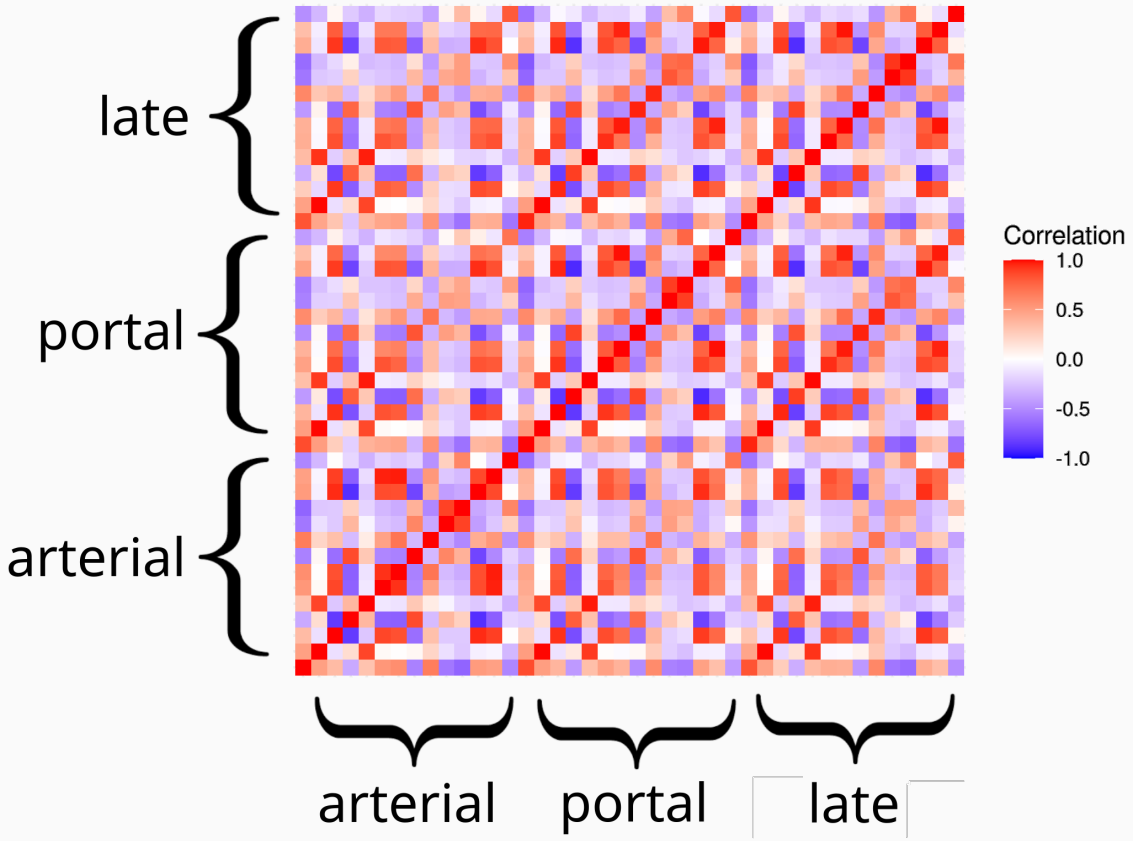


Figure 3: Correlation matrix of the texture coefficients elative to Gray Level Dependence Matrix (GLDM)

Tensor data

Finding the best algorithm considering the structure of the data.

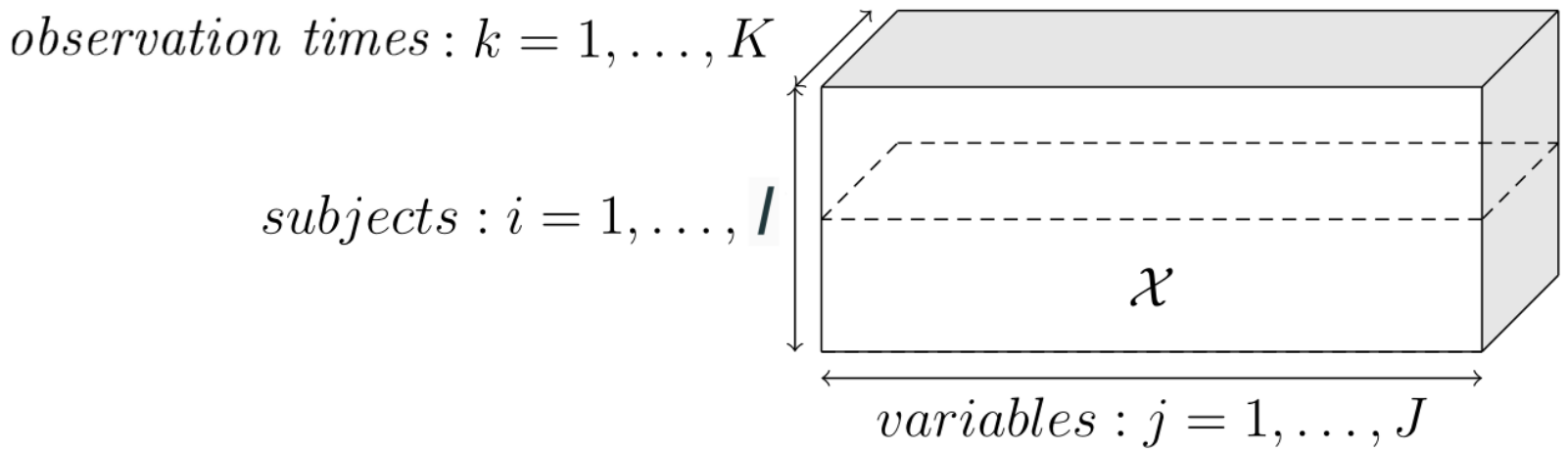


Figure 4: Type of data: tensorial

Multibloc data

Features about pixel/voxel intensities, shape and texture: different natures

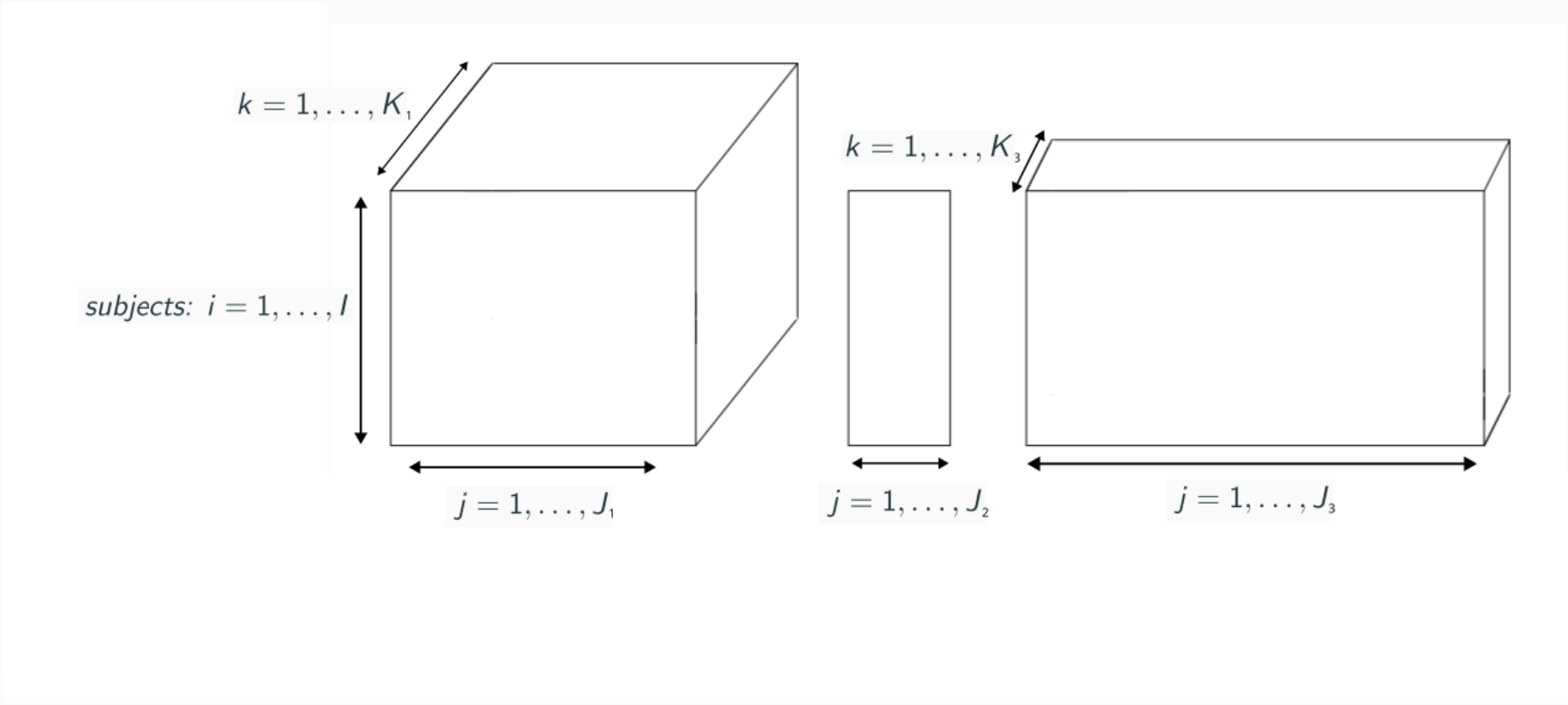


Figure 5: Type of data: multibloc

Table of contents

Machine learning models

Logistic regression

I

Classical machine learning (few data and explainable)

$$P(Y = 1|x) = \frac{\exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})}$$

Defines a likelihood function $\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = y_i|x_i)$

function to minimize : $-\log(\mathcal{L}(\boldsymbol{\beta})) + \text{penalization}$

Penalization to avoid overfitting. In our case L1 norm (lasso)

Limitations of the logistic regression lasso

• Each feature impact considered independently (no link between the times). Adapting penalty would change nothing on this

• Elimination of features without specific considerations for the same feature at other times/ other features at the same time

Limitations of the logistic regression lasso

• Each feature impact considered independently (no link between the times). Adapting penalty would change nothing on this

• Elimination of features without specific considerations for the same feature at other times/ other features at the same time

2 main solutions:

- Preprocessing the data so it becomes tabular (equivalent to the approach chosen with latest data). But using only PCA + clustering: bad results
- Adapting the model to the structure of the data (aim of the internship)

Tensor regression models

Idea: each variable and mode has its own influence on the prediction (i.e. on $\boldsymbol{\beta}$) [2].

For J variables observed following M modes (time, depth ...)

$$\beta_{j,k_1,\ldots,k_M} = \beta_j^J \beta_{k_1}^{K_1} \ldots \beta_{k_m}^{K_M}$$

Tensor regression models

Idea: each variable and mode has its own influence on the prediction (i.e. on $\boldsymbol{\beta}$) [2].

For J variables observed following M modes (time, depth ...)

$$\beta_{j,k_1,\ldots,k_M} = \beta_j^J \beta_{k_1}^{K_1} \ldots \beta_{k_m}^{K_M}$$

reformulation with $\boldsymbol{\beta}$ as a vector:

$$\boldsymbol{\beta} = [\beta_{1,1,\ldots,1}, \beta_{2,1,\ldots,1} \ldots \beta_{1,2,\ldots,1} \ldots \beta_{J,K_1,\ldots,K_M}]^T \text{ (lexicographic order)}$$

$$= \boldsymbol{\beta}^{K_M} \otimes \ldots \otimes \boldsymbol{\beta}^{K_1} \otimes \boldsymbol{\beta}^J \text{ (Kronecker product)}$$

Concatenate non tensor coefficients at the end of the Kronecker structure

Limits of rank 1

$\beta^{K_M} \otimes \ldots \otimes \beta^{K_1} \otimes \beta^{J}$ is only rank 1

In 2D :

Figure 6: Example of rank 1 pictogram (only 0 and 1)

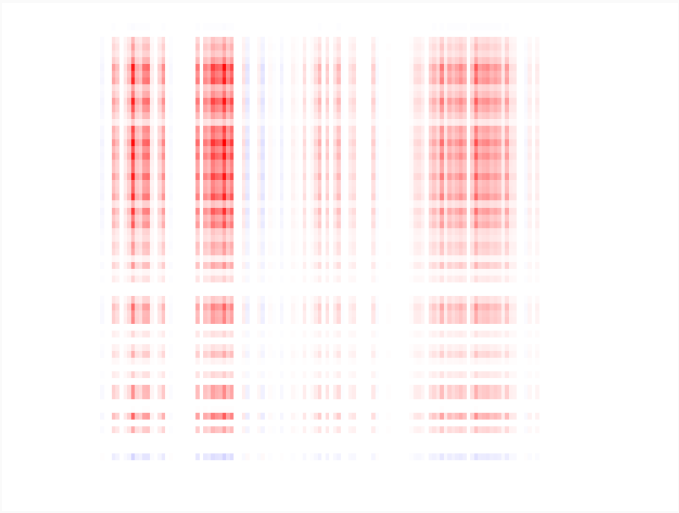


Figure 7: Example of rank 1 matrix (all values allowed)

Rank R multiway logistic regression

Allowing any rank R [1]

$$\boldsymbol{\beta} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^{K_M} \otimes \ldots \otimes \boldsymbol{\beta}_r^{K_1} \otimes \boldsymbol{\beta}_r^{J}$$

Chosen penalization: $\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^{K_M} \otimes \ldots \otimes \boldsymbol{\beta}_r^{K_1} \otimes \boldsymbol{\beta}_r^{J}\|_1$

- L1-type penlization: encourage sparsity of each $\boldsymbol{\beta}_r^{K_m}$ and $\boldsymbol{\beta}_r^{J}$
- Allows efficient optimization, considering the $\boldsymbol{\beta}$-structure

grouping by blocs

Problem: Several groups of variables of different natures (first order, shape, texture). But $\beta_r^{K_1} \ldots, \beta_r^J$ common to all groups.

Solution : Separating each group of variables in different blocks (tensors), with own $\beta_r^{K_1} \ldots, \beta_r^J$ for each block.

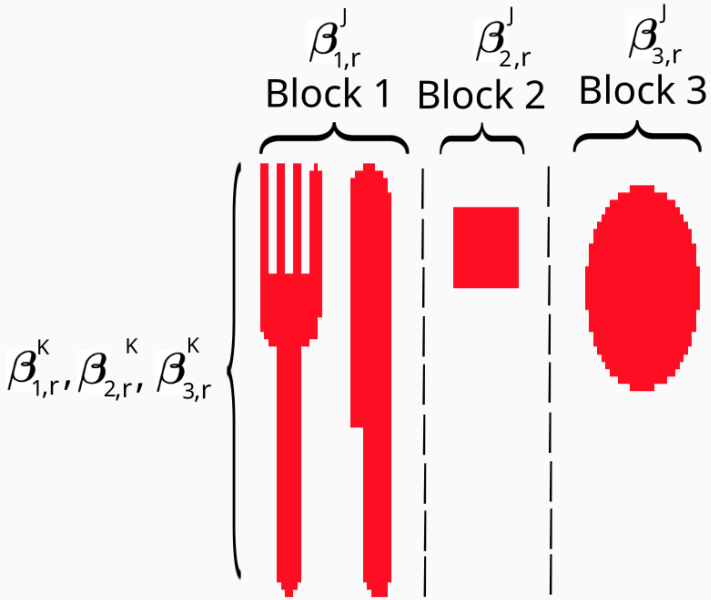


Figure 8: Separation by blocks of a pictogram

Multiway Multiblock Logistic Regression (MMLR)

For L blocks, when β of order 2 (extension rank R straightforward):

$$\beta = \left[\sum_{r=1}^{R_1} \beta^K_{(1,r)} \otimes \beta^J_{(1,r)}; \quad \dots \quad ; \sum_{r=1}^{R_L} \beta^K_{(L,r)} \otimes \beta^J_{(L,r)} \right]$$

Allows having different orders for 2 different blocks, e.g.: averaging over time only 1 block

Same penalization as multiway model but summed by block

Simulations

Parameters to control:

- Difficulty of the classification (overlap between classes, distance between means of classes etc ...)
- Balance between classes
- Structure of the regression coefficients β (several blocks)
- Quality of the classification (AUC)
- Quality of the reconstruction of β (pictograms)

Data generation

Chose the $\boldsymbol{\beta}$ to be reconstructed (pictograms)

Generate the $(\mathbf{x}_i)_{i \in [\![1,I]\!]}$ with 2 multivariate normal laws of means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common covariance matrix $\boldsymbol{\Sigma}$ such that:

- $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ colinear to $\boldsymbol{\beta}$
- One of the principal axis of $\boldsymbol{\Sigma}$ colinear to $\boldsymbol{\beta}$

Separation of classes linked with eigenvalues of $\boldsymbol{\Sigma}$ (to be compared with $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$)

Example in 2D

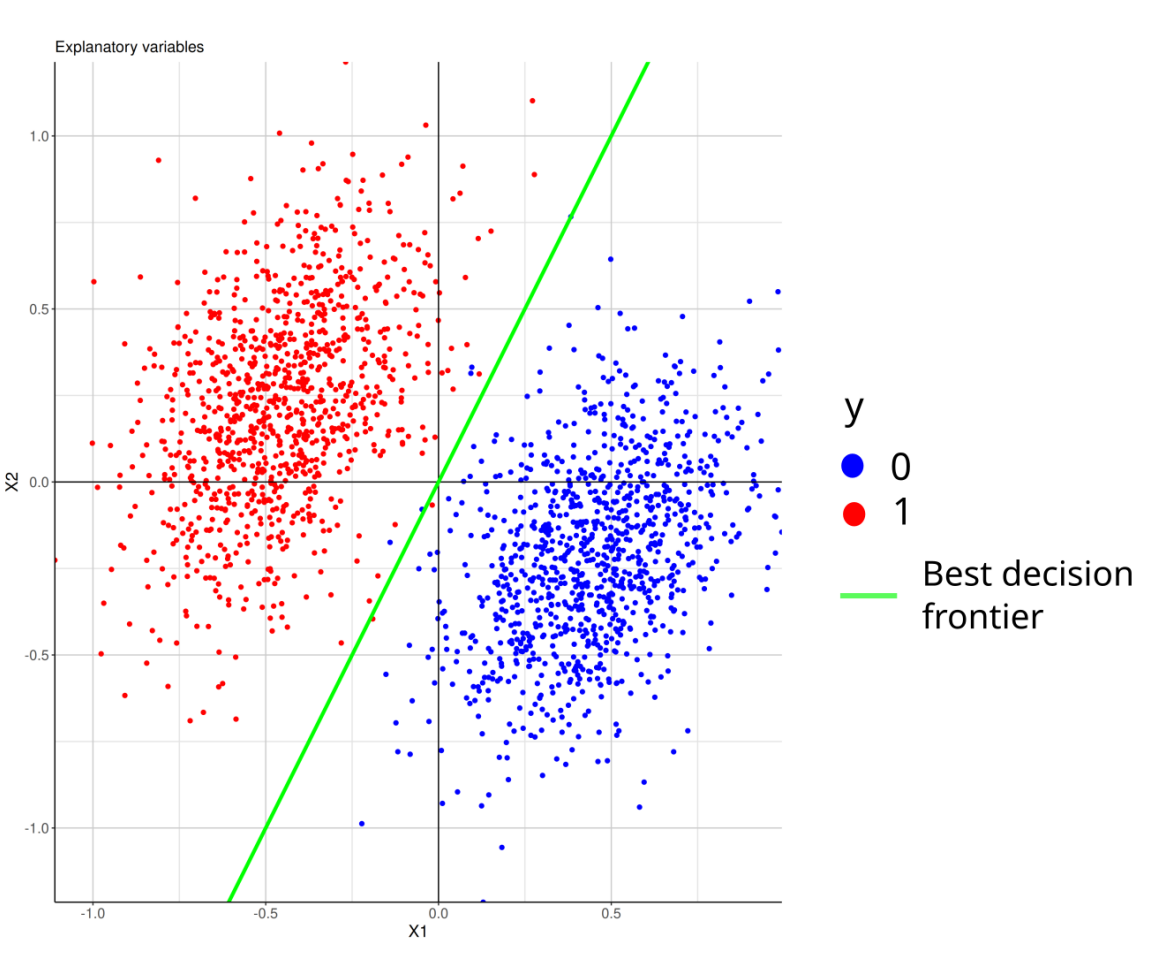


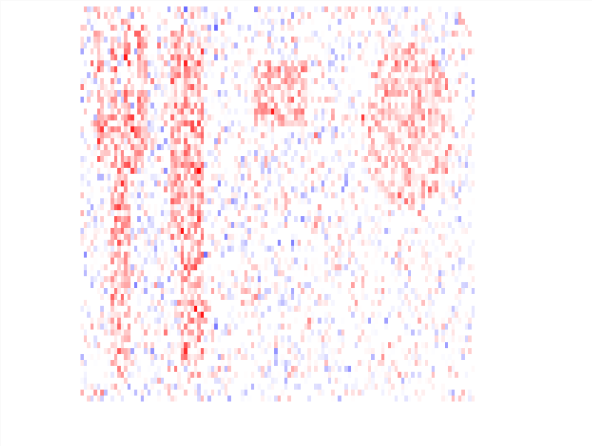
Figure 9: Example of explanatory variables for $\beta = (-2, 1)$

AUC simulated data

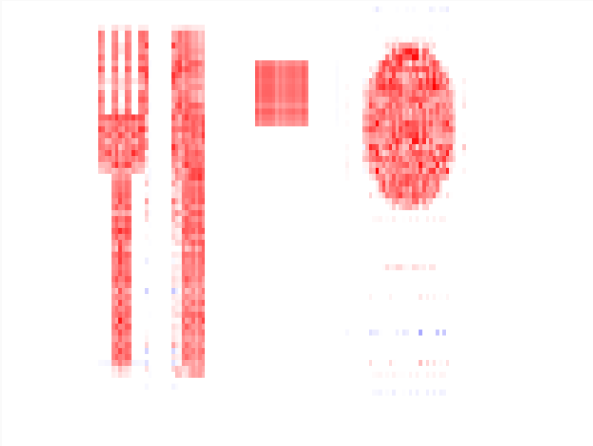
Table 1: AUC for each model on simulated data for 3000 individuals

$(\sigma_\beta, \sigma_{\text{noise}})$	lasso	g. l (blocs)	g.l (mode)	g.l (var)	tensor	tensor blocks
(0.1,0.5)	0.83	0.86	0.94	0.94	0.99	0.99
(0.1,0.8)	0.63	0.64	0.68	0.68	0.93	0.99

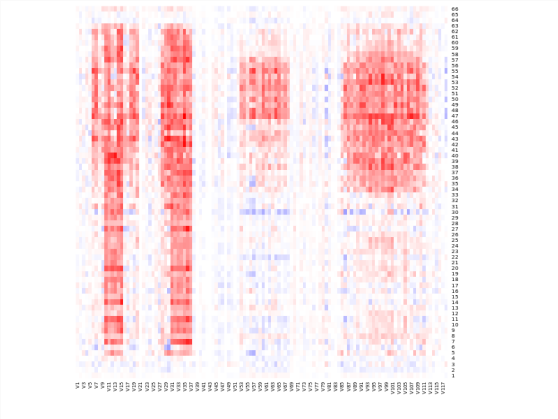
Reconstructed β



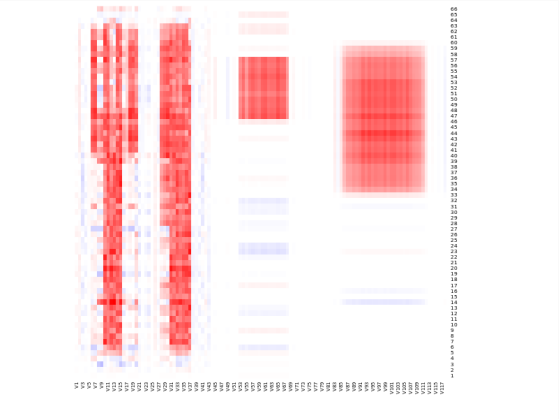
(a) lasso
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.5)$



(b) multiway multiblock
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.5)$



(c) multiway
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$



(d) multiway multiblock
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$

real data

Features extraction with pyradiomics

Extraction of $\simeq 100$ features (about intensities, shape, texture) for each 2D or 3D image.

Difficulties: Which spacing? Which slices if 2D?

Features extraction with pyradiomics

Extraction of $\simeq 100$ features (about intensities, shape, texture) for each 2D or 3D image.

Difficulties: Which spacing? Which slices if 2D?

In 3D:

- Common spacing
- Average shape over time

Features extraction in 2D

Slices along z axis $\rightarrow$ same spacing along (x, y)

But: No common spacing along z

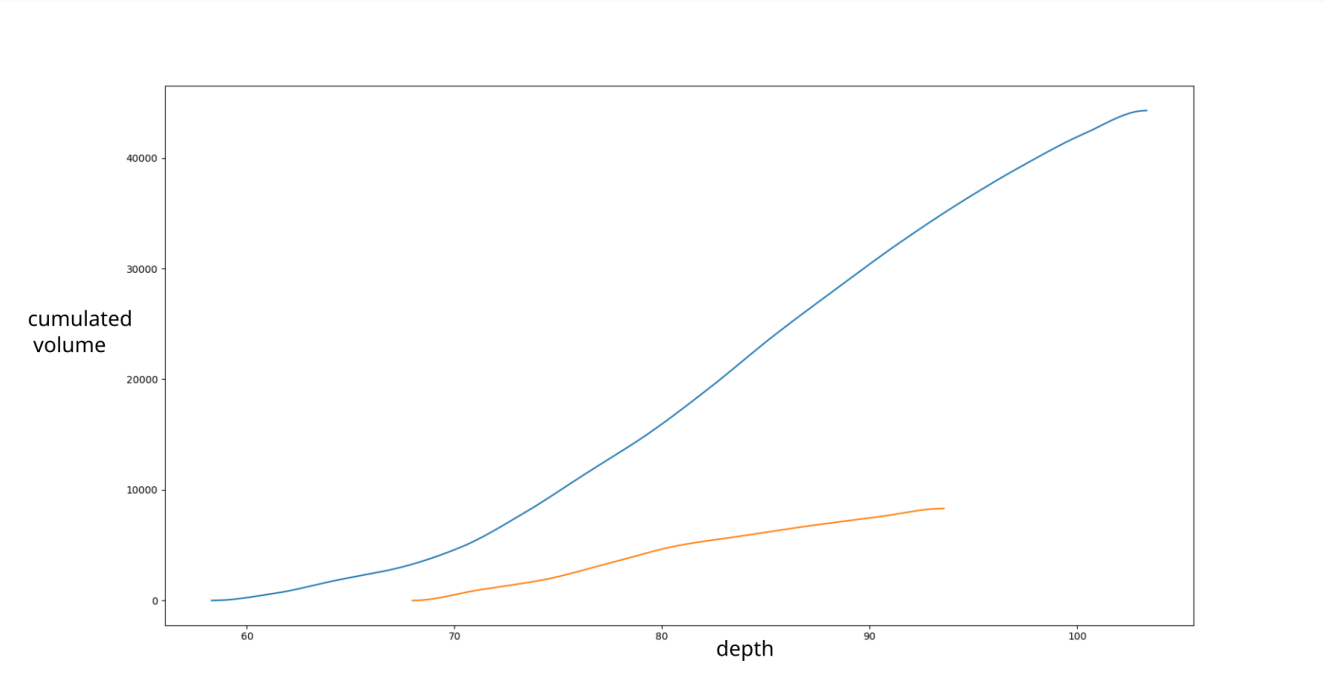


Figure 11: Curves of the cumulated volume of tumor (in mm^3) as a function of the depth travelled in the liver (in mm) for 2 tumors during arterial phase

Features extraction in 2D

Solution: Selecting 5 slices equally spaced along the cumulated volume axis

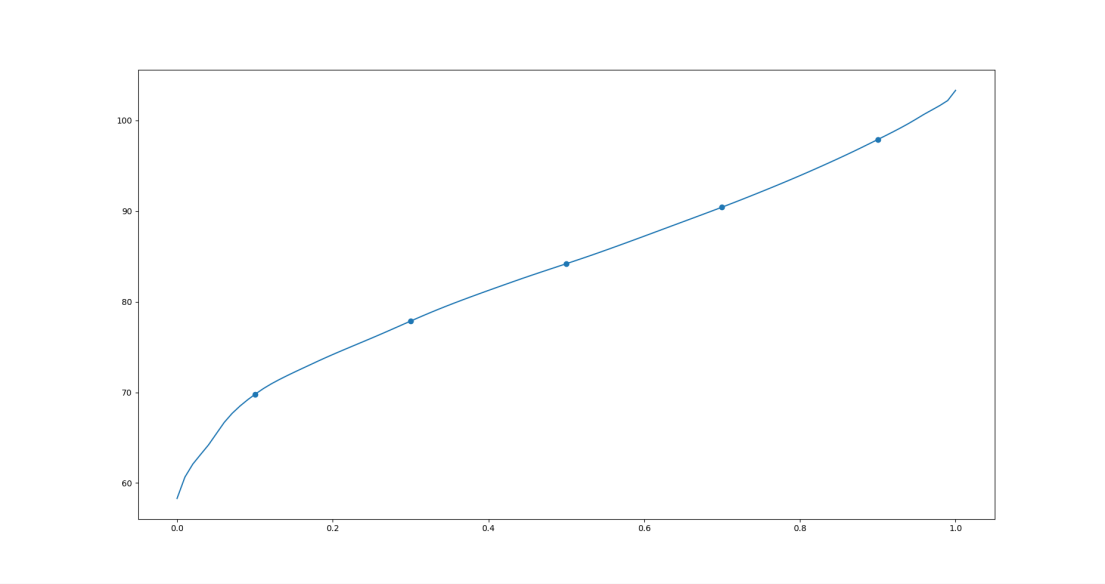


Figure 12: Curve of the depth travelled in the liver (in *mm*) as a function of the standardized cumulated volume of the tumor. The points represent the selected slices.

Results

Type of data	lasso	g.l. (block)	g.l. (time)	g.l. (var)	tensor	tensor blocks
3D	0.74 ± 0.04	0.78 ± 0.03	0.76 ± 0.03	0.73 ± 0.03	0.77 ± 0.03	0.77 ± 0.03

Area under curve (AUC) on 3D real data

Type of data	lasso	g.l. (block)	g.l. (slice)	g.l. (time)	g.l. (var)	tensor	tensor blocks
2D	0.73 ± 0.03	0.71 ± 0.03	0.70 ± 0.04	0.71 ± 0.03	0.71 ± 0.03	0.66 ± 0.04	0.71 ± 0.03

Area under curve (AUC) on 2D real data

Latest data

13 binary features determined by radiologists (late enhancement, non peripheral washout etc...) + sex

With lasso model:

- AUC: 0.97 ± 0.02
- balanced accurcy: 0.88 ± 0.05

Would be interesting to test other models...

features importance

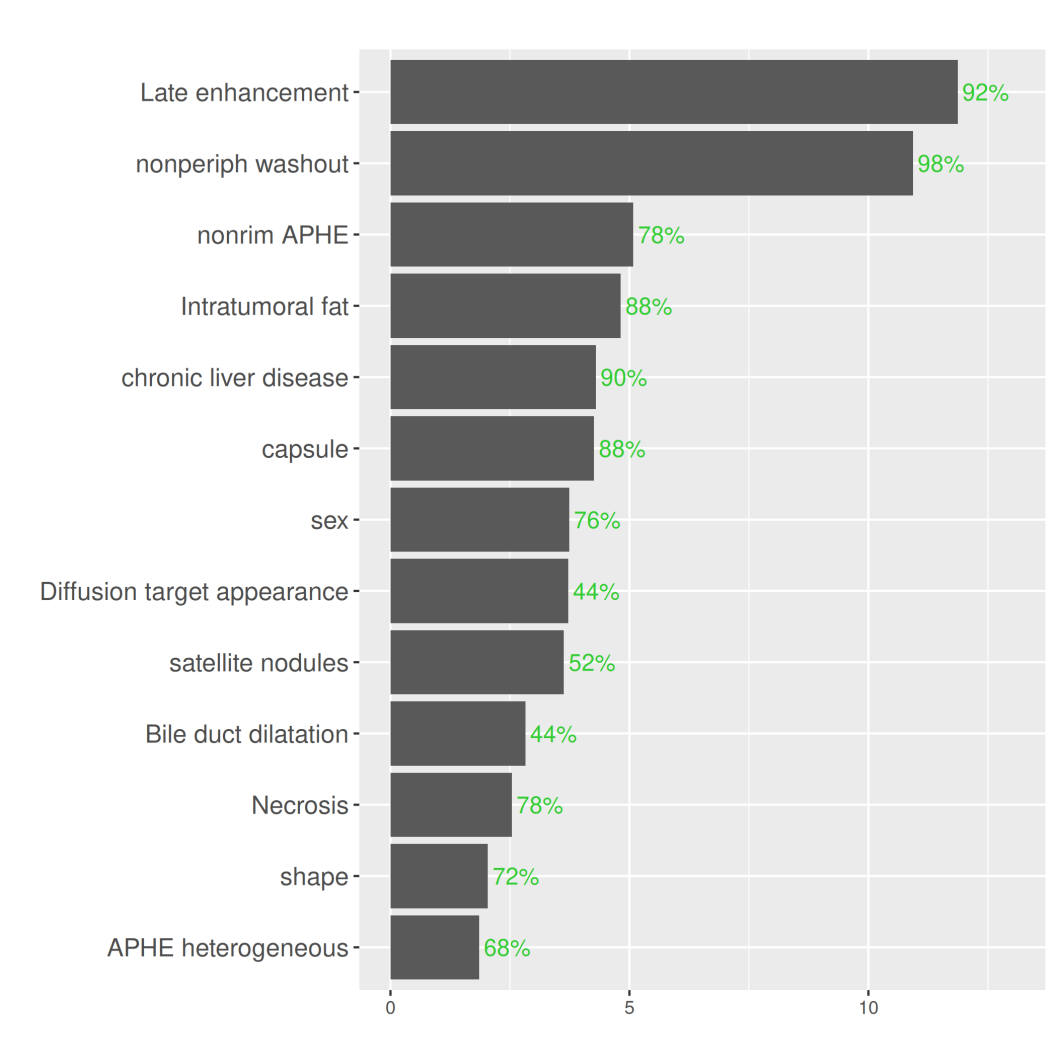


Figure 13: Features importance with lasso (in green percentage of runs with non null coefficient)

Possible extensions

Testing other penalizations (group lasso, elastic net)

Extending the multiblock approach to other classical machine learning algorithms (other GLMs, SVM etc...)

Testing other models on the latest data (in order to obtain a model that can be deployed in the hospital)

Implementing the multiblock code in C forincreased speed (currently quite slowin R)

Retrospective Analysis

Impact of the internship on me

Direct impact: continuing in thesis (increase in motivation for research activities)

Soft skills in machine learning: become more critical vs results, searching for other data whenever possible

Being part of a team in a scientific context (not only 1 supervisor): importance of communication and reporting (even when no writen documents)

The reasearch in machine learning: an accessible world

Consequences of the internship

A promising framework for the diagnosis of liver tumors

The simulation part of an article on the multiblock multiway logistic regression

More information about the correct context for using that kind of models

An ethically positive impact (controllable deployment, a precise need, no replacement of humans...)

Conclusion

A good representation of a research work (and its challenges)

Supportive, available and Calm supervision (evn as deadlines approach)

Looking forward to continuing in this direction

bibliography


Fabien Girka, Pierrick Chevaillier, Arnaud Gloaguen, Giulia Gennari, Ghislaine Dehaene-Lambertz, Laurent Le Brusquet, and Arthur Tenenhaus.

Rank-R Multiway Logistic Regression.

In *52èmes Journées de Statistique*, Nice, France, 2021.

les 52èmes journées de Statistique 2020 sont reportées ! Elles auront lieu du 7 au 11 Juin 2021.


Laurent Le Brusquet, Gisela Lechuga, and Arthur Tenenhaus.

Régression Logistique Multivoie.

In *JdS 2014*, page 6 pages, Rennes, France, June 2014.