

# Tensor multiblock logistic regression to classify liver tumors from MRI images

---

SELVESTREL Alexandre

November 28, 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<sup>th</sup> most widespread cancer and 4<sup>th</sup> mortality cause by cancer**

Classification:

- Hepatocellular Carcinoma (HCC): 75% of cases, resection often possible
- CCK = Cholangiocarcinoma (CCK): 6% of cases, resection difficult (possible in 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 levels of alpha-fetoprotein indicate HCC, but not always.
- Biopsy: invasive and potentially lethal (0.02% of patients)

But a lot of clues even without biopsy → Machine Learning

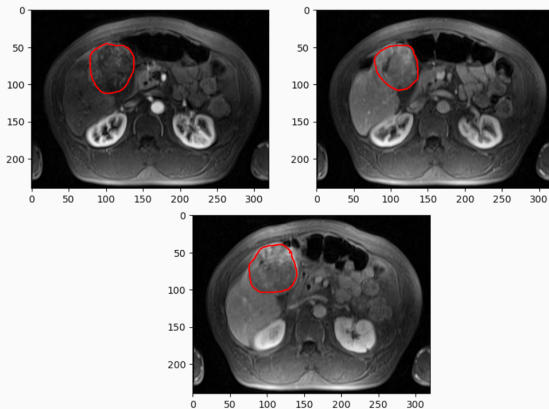
## Available data

- RMI images in 3D of liver tumors (arterial, portal, late)
- gender (63 men, 27 women)
- age at disease (average: 63 years old)

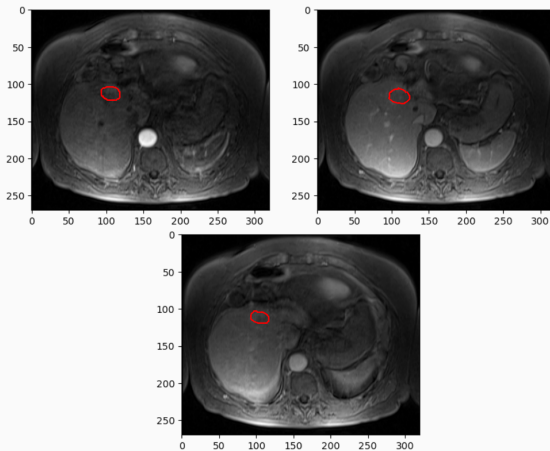
From Henri Mondor hospital: Sébastien Mulé

Same variables extracted from each RMI images 3 times (shape, texture, intensity) → specific structure

Need to adapt existing machine learning methods to this structure



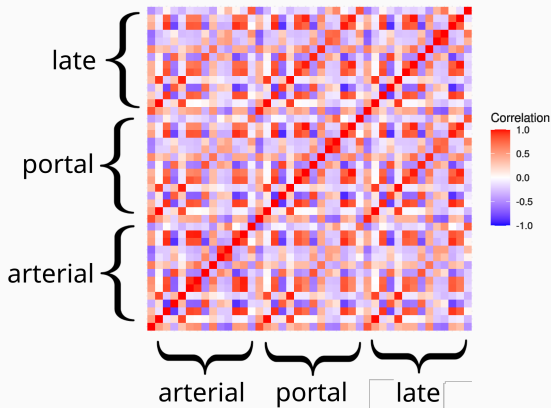
**Figure 1:** Example of RMI images of a HCC tumor (arterial, portal, late). More contrast in arterial



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

## Correlation matrix of features about texture (GLDM)

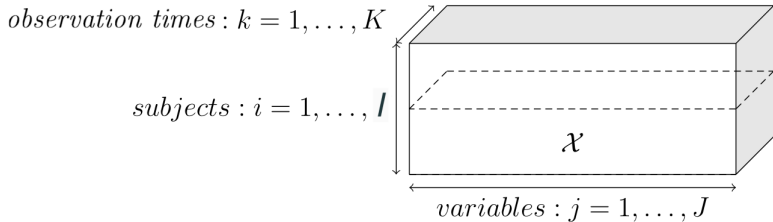
Strong correlations between imaging times for a given variable



**Figure 3:** Correlation matrix of the features relative to the Gray Level Dependence Matrix (GLDM)

# Tensor data

Finding the best algorithm considering the structure of the data.

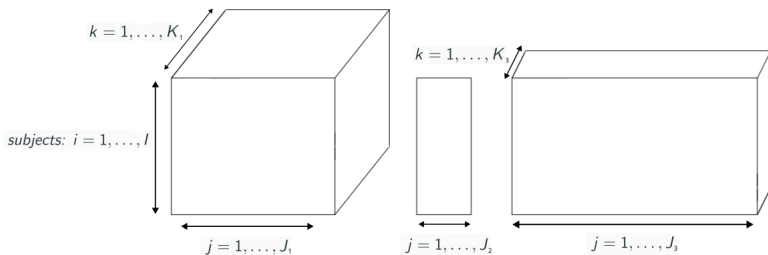


**Figure 4:** Type of data: tensorial



# Multibloc data

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



**Figure 5:** Type of data: multibloc

# Table of contents

Machine learning models

- tabular models

- tensor models

Simulations

Liver tumor data

- with pyradiomics

- latest data

Retrospective Analysis

# Machine learning models

---

# Logistic regression

Classical machine learning (works with few data and explainable)

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

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

To many features (vs  $I$ )  $\rightarrow$  need to limit variance of prediction.

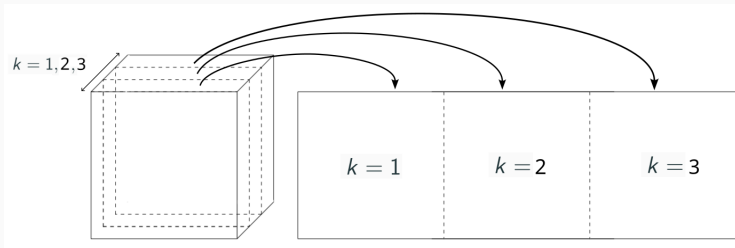
Penalization with  $\|\boldsymbol{\beta}\|_1$ : lasso

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

## Impact of the tensor nature of $\beta$

$\beta = (\beta_{j,k})_{j \in \llbracket 1, J \rrbracket, k \in \llbracket 1, K \rrbracket}$  so  $JK$  parameters to determine

$$x^T \beta \rightsquigarrow \sum_k \sum_j \beta_{j,k} x_{j,k} \quad \text{and} \quad \|\beta\|_1 = \sum_k \sum_j |\beta_{j,k}|$$



**Figure 6:** Unfolding a tensor

**Limitation of lasso:** Elimination of features without specific consideration for the same feature at other times/ other features at the same time

Common solution: grouping regression coefficients together in the penalization

$$\sum_{j,k} |\beta_{j,k}| \rightsquigarrow \sum_{g=1}^G \|\beta^g\|_2$$

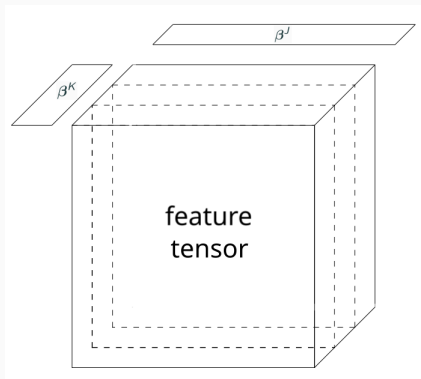
Tendency to set regression coefficients to zero by entire blocks.

**But:** grouping either by mode or by variable, not both

→ 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  $\beta$ ) [2].



**Figure 7:** Tensor structure of  $\beta$

# Tensor regression models

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

For  $J$  variables observed following  $K$  modalities (e.g. times)

$$\beta_{j,k} = \beta_j^J \beta_k^K$$

$\beta_j$  : impact of variable  $j$

$\beta_k$  : impact of modality  $k$

Only  $J + K$  parameters to determine (instead of  $JK$ )

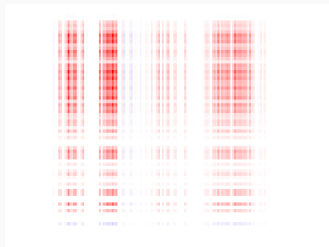


## Limits of rank 1

$\beta_{j,k} = \beta_j^J \beta_k^K$  implies that  $\beta$  looks like:



**Figure 8:** Example of rank 1 pictogram (only 0 and 1)

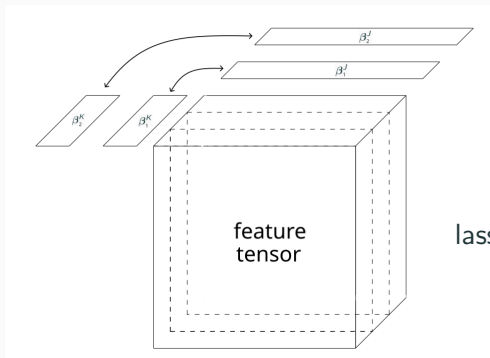


**Figure 9:** Example of rank 1 matrix (all values allowed)

This can be too simplistic

# Rank R tensor logistic regression [1]

Summing rank 1 together :  $\beta_{j,k} = \sum_{r=1}^R \beta_{j,r}^J \beta_{k,r}^K$

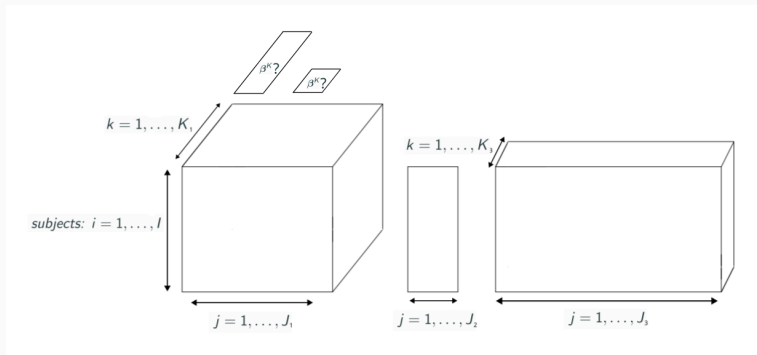


$$\text{lasso} \rightsquigarrow \sum_{r=1}^R \left( \|\beta_{(1,r)}^J\|_1 \|\beta_{(1,r)}^K\|_1 \right)$$

**Figure 10:** Tensor structure of  $\beta$

# Blocks of variables

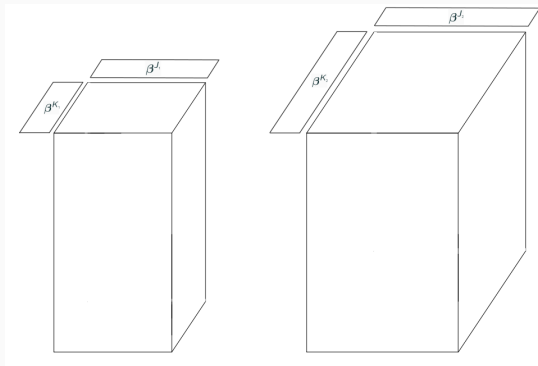
**Problem:** Several groups of variables of different natures (first order, shape, texture). But  $\beta_r^K$  and  $\beta_r^J$  common to all groups.  
 $K_1 = K_2 = K_3$  needed or else:



**Figure 11:** Problem if blocks have different orders or dimensions

# Tensor multiblock logistic regression

**Solution:** giving each block its own  $\beta^J$  and  $\beta^K$



**Figure 12:** Tensor multiblock model for rank 1

# Tensor multiblock logistic regression

Mathematically, this gives :

$$\mathbf{x}^T \boldsymbol{\beta} \rightsquigarrow = \sum_{l=1}^L \sum_{j,k} x_{j,k}^l \beta_{j,k}^l$$

With, for rank 1:  $\beta_{j,k}^l = \beta_j^{J_l} \beta_k^{K_l}$

But each  $\beta^l$  can have a different rank  $R_l$ , which gives:

$$\beta_{j,k}^l = \sum_{r=1}^{R_l} (\beta_r^{J_l})_j (\beta_r^{K_l})_k$$

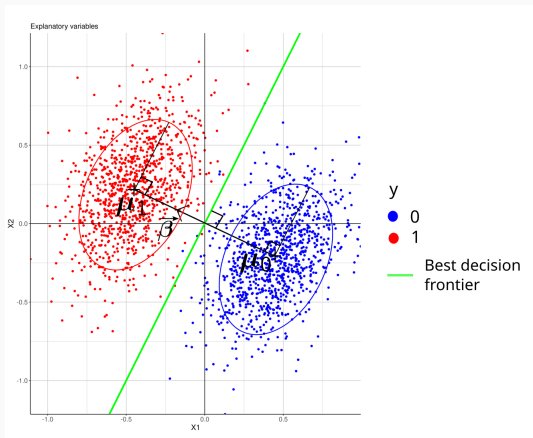
# Simulations

---

## Parameters to control:

- Difficulty of the classification (overlap between classes, distance between means of classes etc ...)
- Balance between classes
- Structure of the regression parameter  $\beta$  (several blocks)
- Quality of the classification (AUC)
- Quality of the reconstruction of  $\beta$

## Illustration in 2D



**Figure 13:** Example of explanatory variables for  $\beta = (-2, 1)$



Chose the  $\beta$  to be reconstructed (pictograms)

Generate the  $(\mathbf{x}_i)_{i \in \llbracket 1, I \rrbracket}$  with 2 multivariate normal laws of means  $\mu_0$  and  $\mu_1$  and common covariance matrix  $\Sigma$  such that:

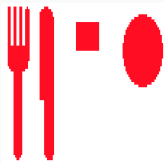
- $\mu_1 - \mu_0$  colinear to  $\beta$
- One of the principal axis of  $\Sigma$  colinear to  $\beta$

Separation of classes linked with eigenvalues of  $\Sigma$  (to be compared with  $\|\mu_1 - \mu_0\|$ )

# AUC simulated data

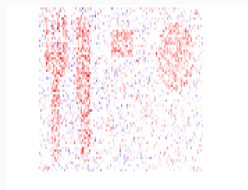
**Table 1:** Cross validated AUC for each model on simulated data for 3000 individuals

$(\sigma_{\beta}, \sigma_{\text{noise}})$	lasso	g.l. (blocks)	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



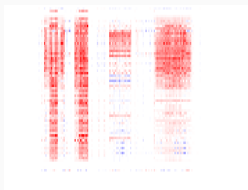
**Figure 14:** Pictogram of shape  $66 \times 117$

# Reconstructed $\beta$



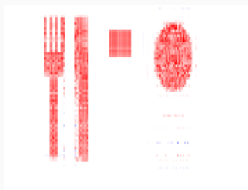
**(a)** lasso

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



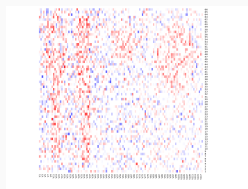
**(b)** tensor  $R : 10$

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



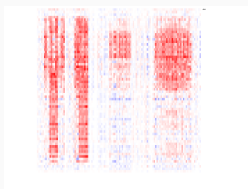
**(c)** T.M.  $R : (12, 1, 10)$

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



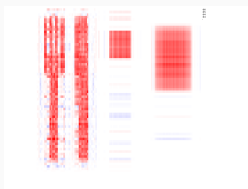
**(d)** lasso

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



**(e)** tensor  $R : 10$

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



**(f)** T.M.  $R : (6, 1, 1)$

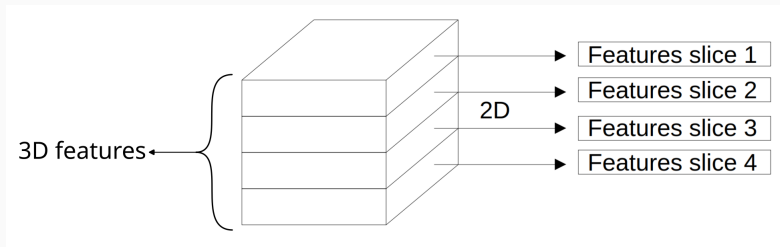
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$

## Liver tumor data

---

## Feature extraction with pyradiomics [3]

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



**Figure 16:** Feature extraction with pyradiomics for an RMI image composed of 4 slices

## Feature extraction in 3D

Each radio  $\rightarrow$  1 particular spacing along  $(x, y, z)$

**But** : Calculations of pyradiomics only use voxels (= 3D pixels).

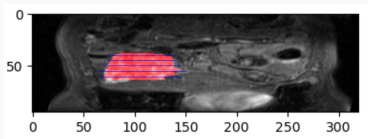
Not always meaningful if the scale changes at each radio (e.g. for Gray Level Run Length Matrix, based on number alignments of pixels of same intensity)

**Solution:** Standardize the spacing along  $(x, y, z)$ . Allowed by resampling (interpolation) of the image.

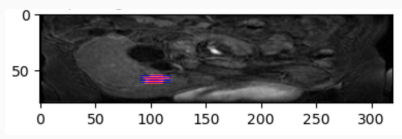
## Feature extraction in 2D

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

**Difficulty** Tumors are of variable locations, sizes and shapes in the frontal plane (front - back plane)



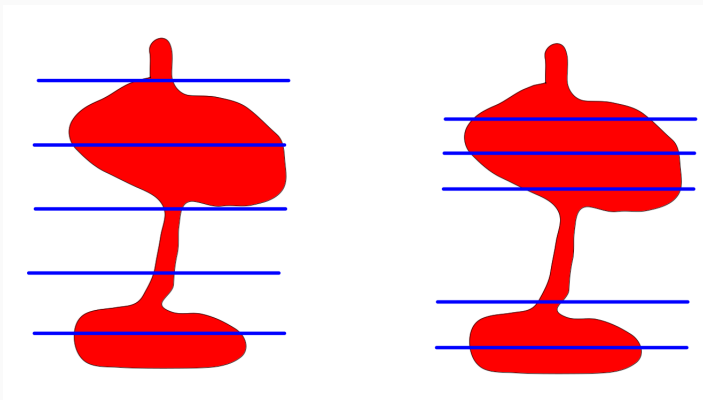
**Figure 17:** Extracting 5 slices in a big tumor



**Figure 18:** Extracting 5 slices in a small tumor

## Feature extraction in 2D

Every slice not equally informative:



**Figure 19:** Slicing relative to the depth vs. relative to the volume travelled in the tumor



# Results

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

Cross validated 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	$0.73 \pm 0.03$	$0.71 \pm 0.03$	$0.70 \pm 0.04$	$0.71 \pm 0.03$	$0.71 \pm 0.03$	$0.66 \pm 0.04$	$0.71 \pm 0.03$

Cross validated area under curve (AUC) on 2D real data

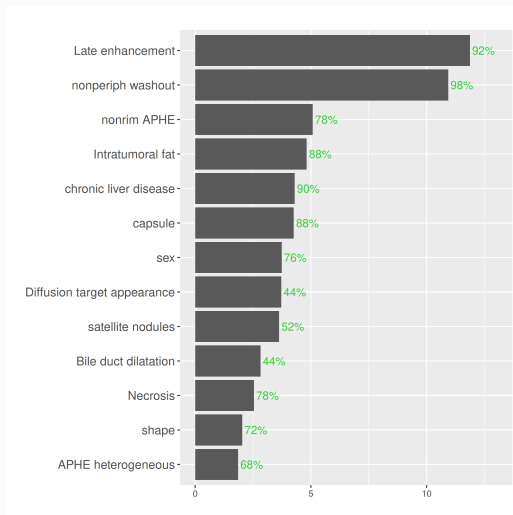
12 binary features determined by radiologists (late enhancement, non peripheral washout etc...) + sex and existence of chronical disease

With lasso model:

- AUC:  $0.97 \pm 0.02$
- balanced accuracy:  $0.88 \pm 0.05$

Would be interesting to test other models...

# Feature importance



**Figure 20:** Feature 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...). Comparing it to CNN.

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 for increased speed (currently quite slow in R)

# Retrospective Analysis

---

## Personal learnings

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

Soft skills in machine learning: becoming 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 written documents)

Research in machine learning: an accessible world

# Conclusion

A promising framework for the diagnosis of liver tumors

The simulation part of an article on the multiblock tensor model

Ethically positive impact (controllable deployment, a precise need, no replacement of human beings...)

## **On a personal level:**

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

Supportive, available and calm supervision (even 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.

 Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts.

**Computational Radiomics System to Decode the Radiographic Phenotype.**

*Cancer Research*, 77(21):e104–e107, 10 2017.

 Hua Zhou, Lexin Li, and Hongtu Zhu.

**Tensor regression with applications in neuroimaging data analysis.**

*Journal of the American Statistical Association*, 108:540–552, 06 2013.