Multiway multiblock logistic regression to classify liver tumors from MRI images

SELVESTREL Alexandre

November 26, 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)

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

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 (shape, texture, intensity) $\rightarrow$ specific structure

Need to adapt existing machine learning methods to this structure

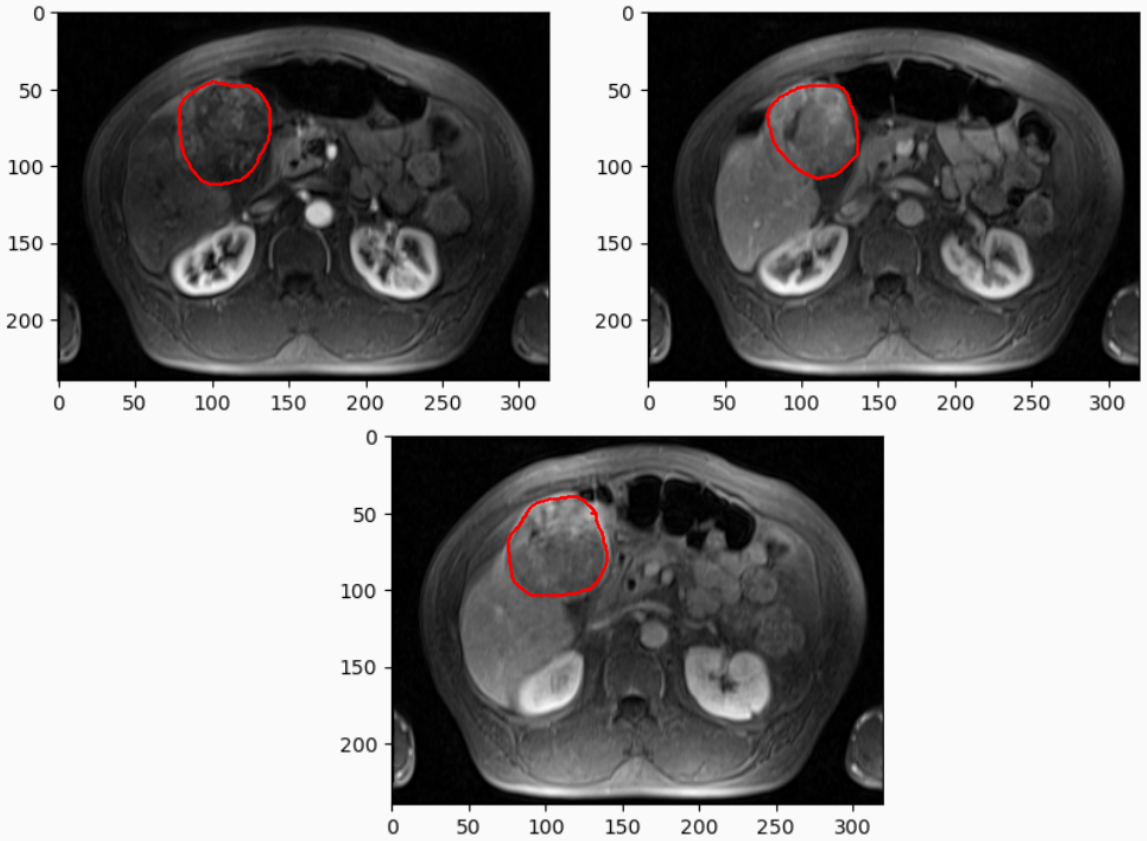


Figure 1: Example of RMI images of a HCC tumor (arterial, portal, late). More contrast in arterial

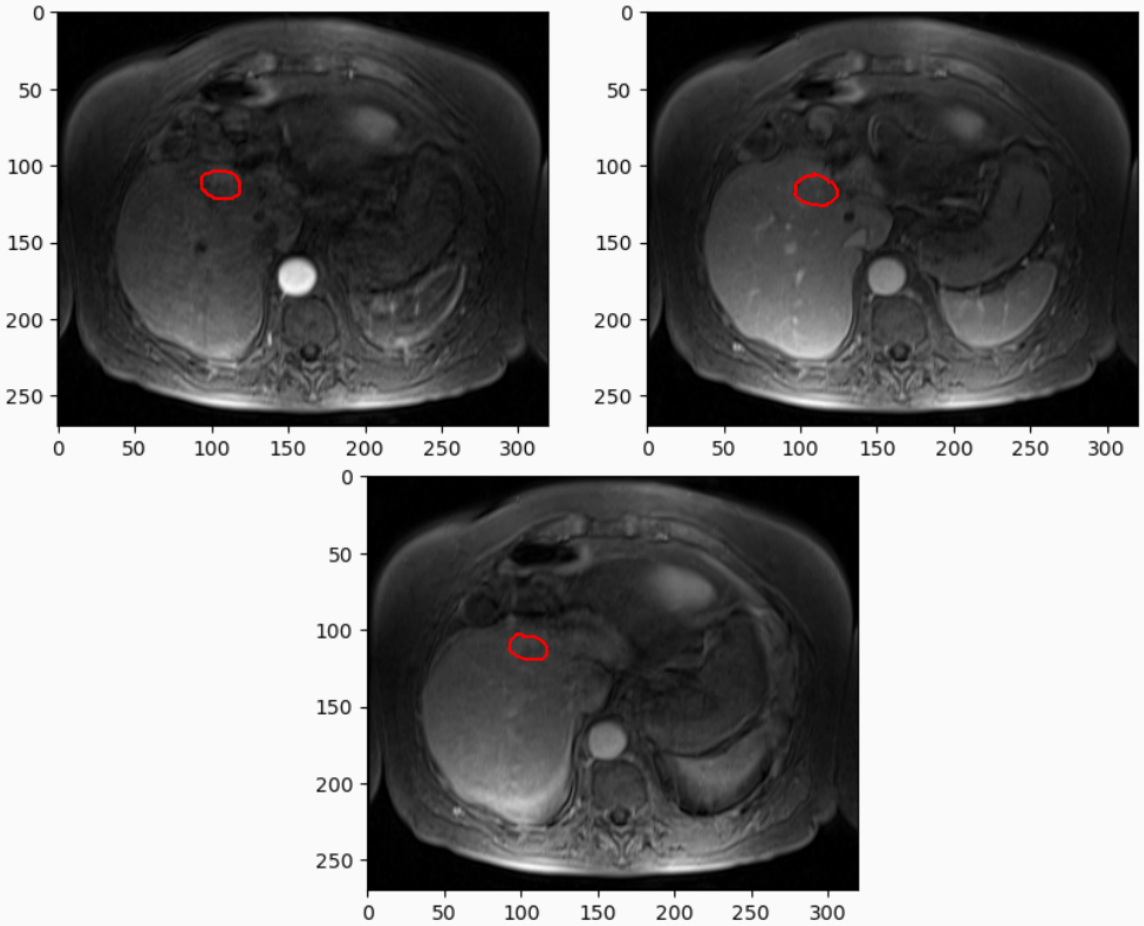


Figure 2: Example of RMI images of a CCK tumor (arterial, portal, late)

Correlation matrix of the texture (GLDM) coefficients

Strong correlations between the imaging times for a given variable

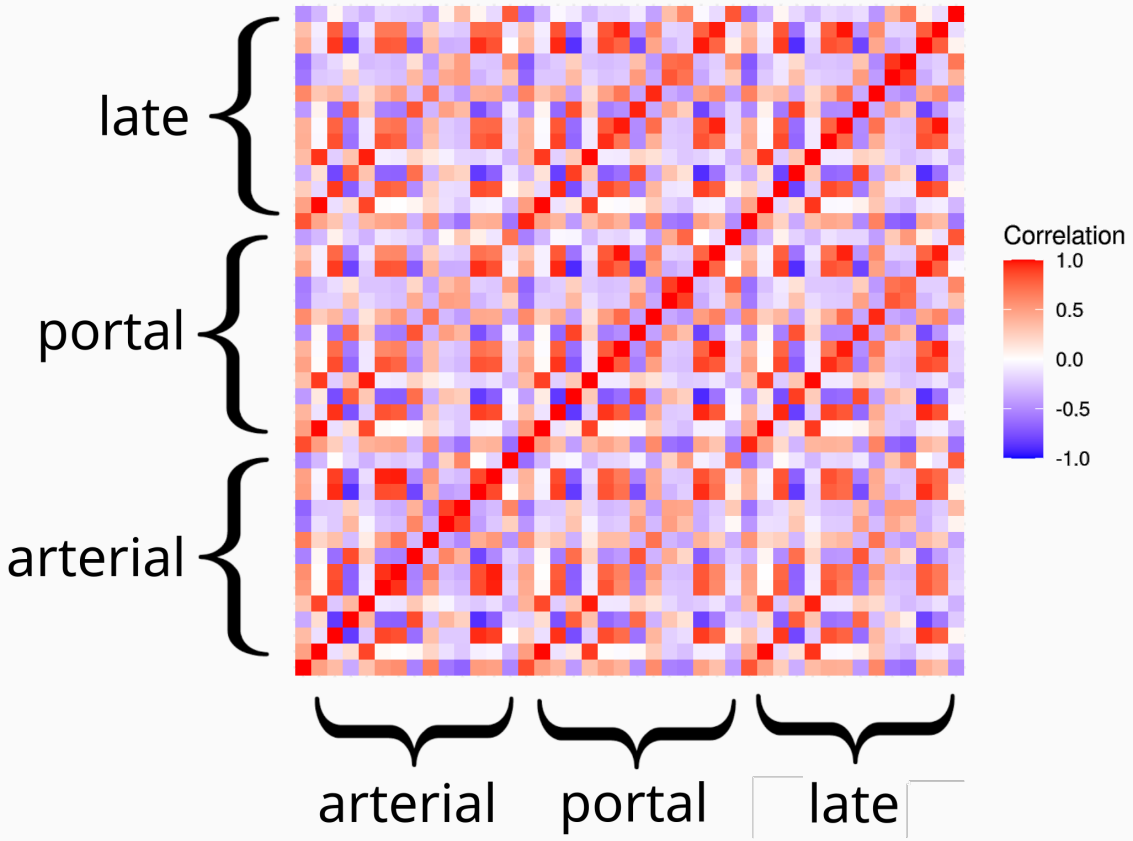


Figure 3: Correlation matrix of the texture coefficients elative to Gray Level Dependence Matrix (GLDM)

Tensor data

Finding the best algorithm considering the structure of the data.

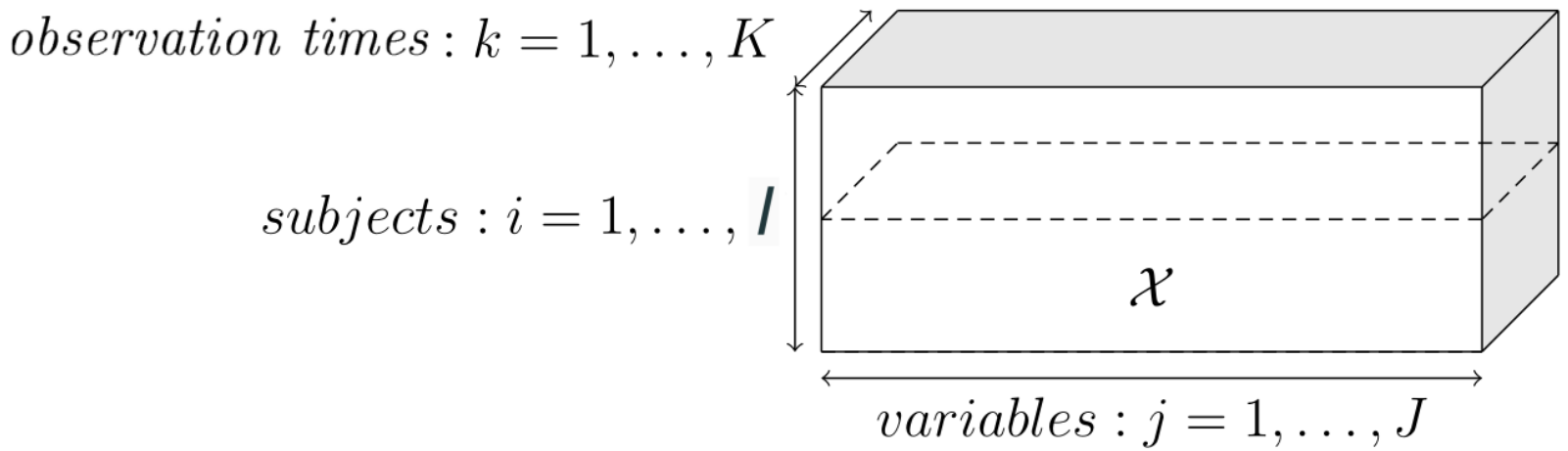


Figure 4: Type of data: tensorial

Multibloc data

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

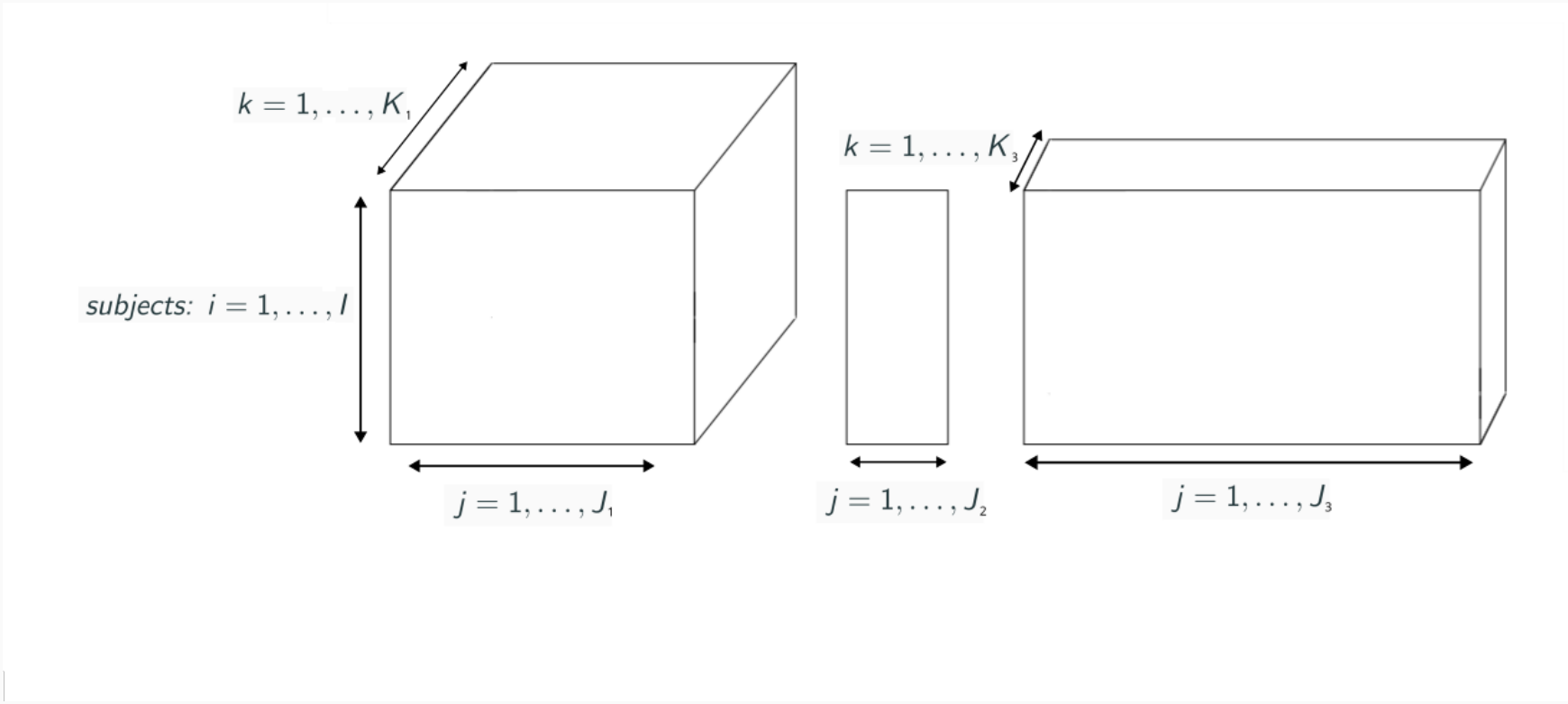


Figure 5: Type of data: multibloc

Table of contents

Machine learning models

Logistic regression

Classical machine learning (works with 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 $\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = y_i|x_i)$

To much features (vs n) $\rightarrow$ 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 β

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

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

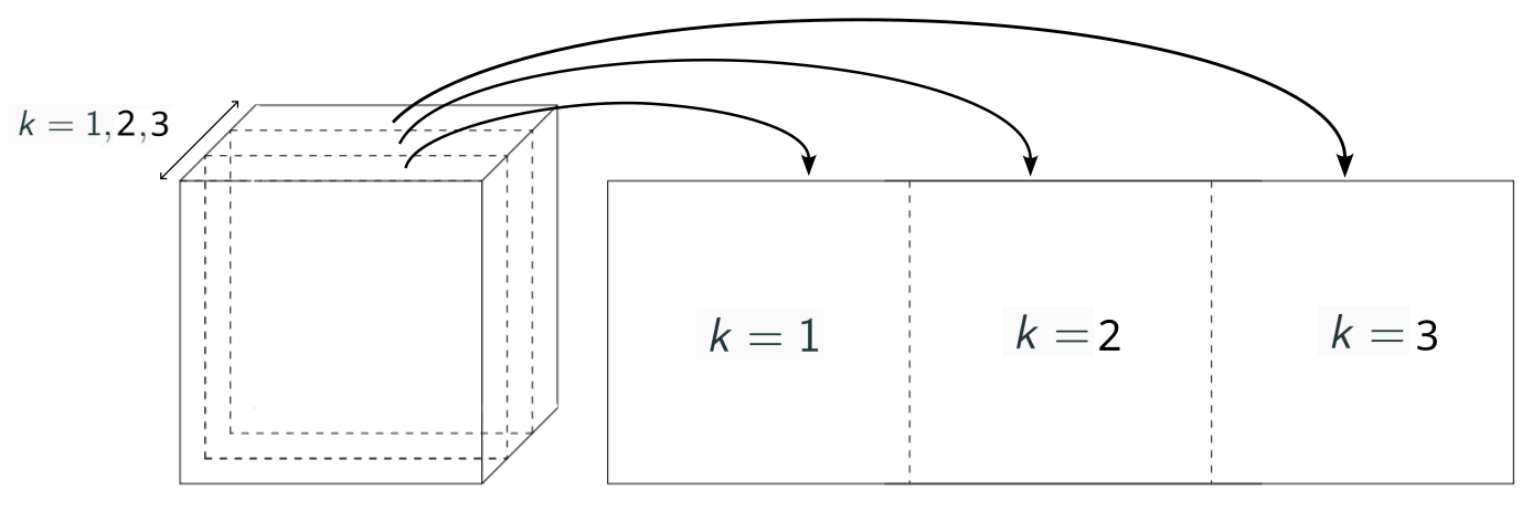


Figure 6: Unfolding of a tensor

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. But only using clustering + PCA $\rightarrow$ poor 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 $\mathcal{B}$) [2].

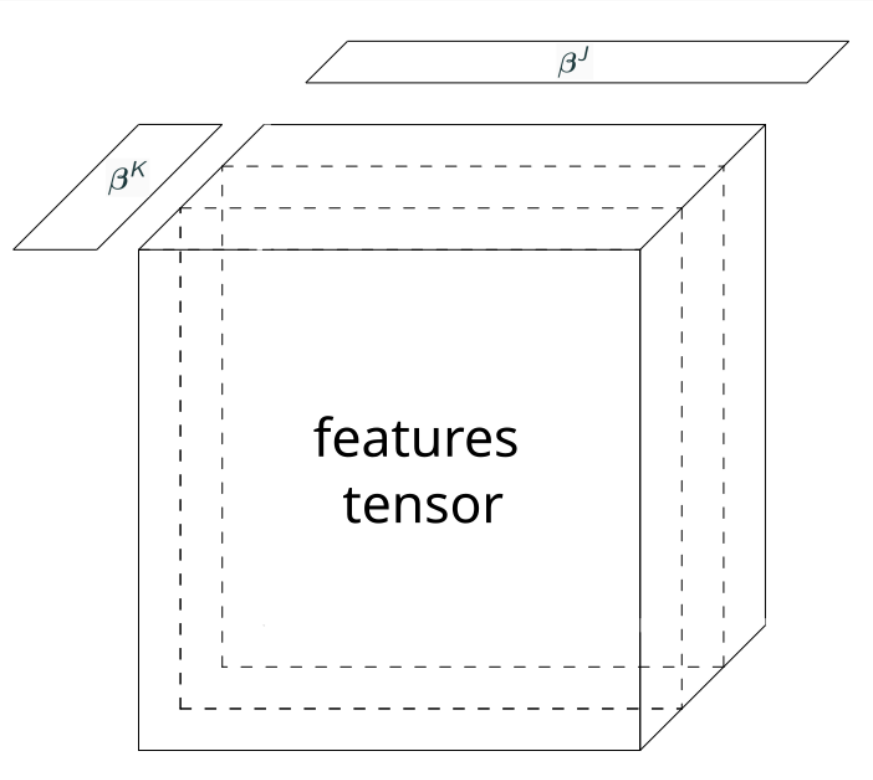


Figure 7: Tensor structure of $\mathcal{B}$

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 K modalities (e.g. times)

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

β_j : impact of variable j

β_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 $\boldsymbol{\beta}$ looks like:

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

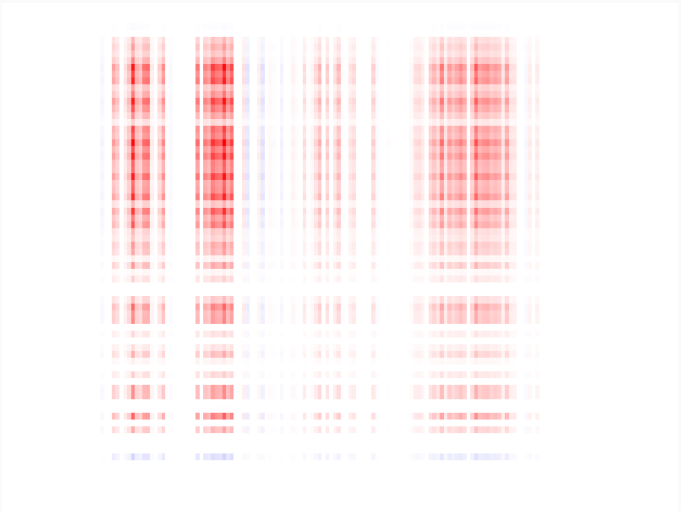


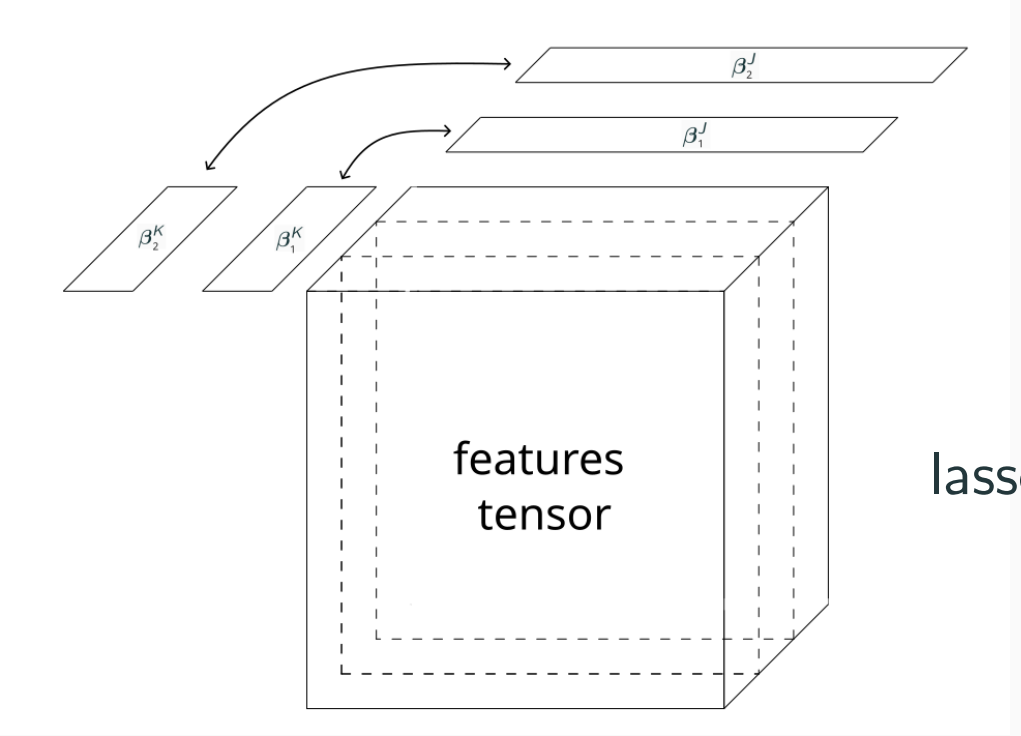
Figure 9: Example of rank 1 matrix (all values allowed)

This can be too simplistic

Rank R multiway logistic regression [1]

$$\text{Summing rank 1 together : } \beta_{j,k} = \sum_{r=1}^{R} \beta^J_{j,r}\beta^K_{k,r}$$

$$\beta_{j,k}x_{j,k} = \left(\sum_{r=1}^{R} \beta^J_{j,r}\beta^K_{k,r}\right) x_{j,k} = \sum_{r=1}^{R} \beta^J_{j,r}\beta^K_{k,r}\, x_{j,k}$$



$$\text{lasso} \rightsquigarrow \sum_{r=1}^{R} \left(\|\boldsymbol{\beta}^J_{(1,r)}\|_1 \|\boldsymbol{\beta}^K_{(1,r)}\|_1 \right)$$

Figure 10: Tensor structure of $\boldsymbol{\beta}$

Blocs of variables

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

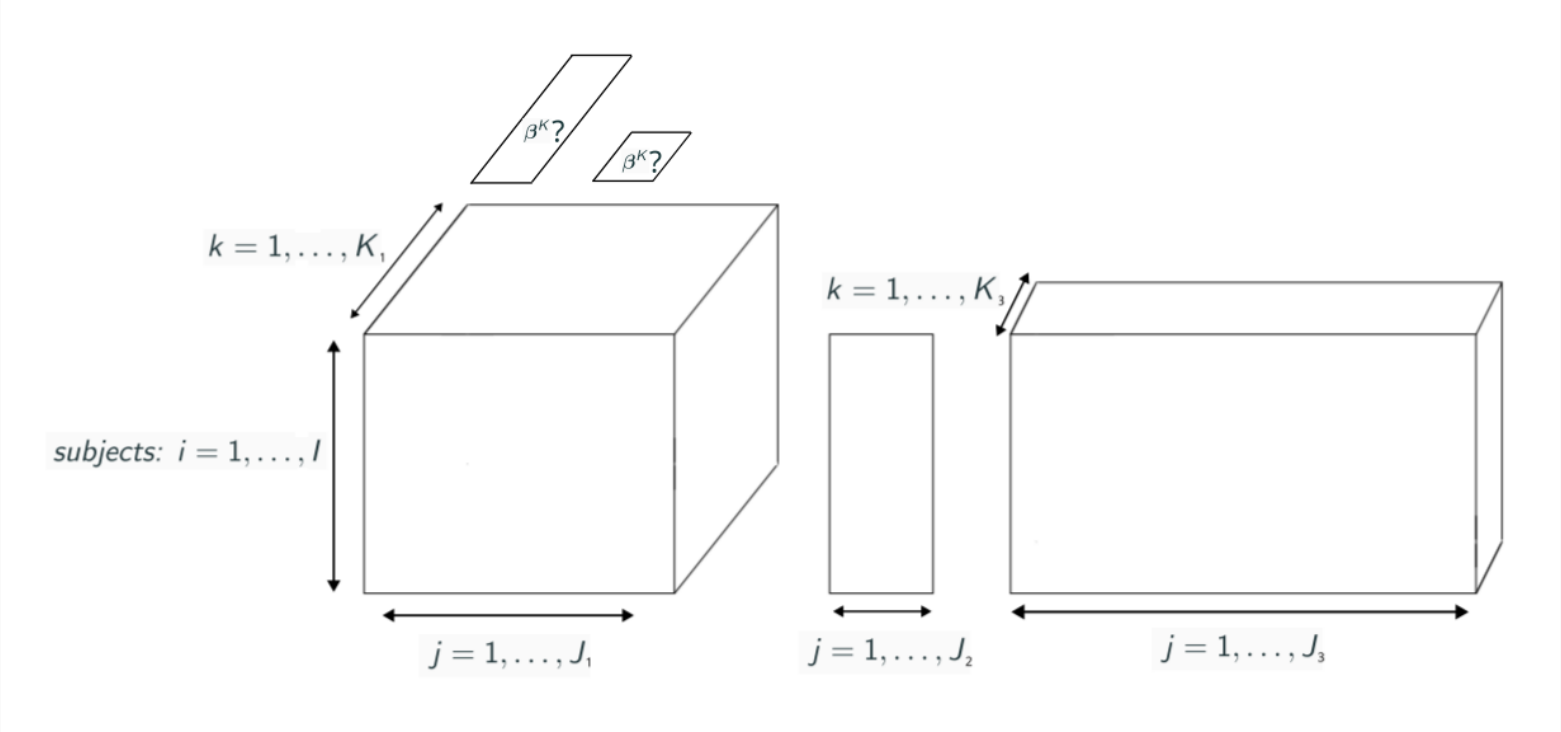


Figure 11: Problem if blocs have different sizes

Multiway model with blocs of variables

If $K_1 = K_2$, blocks can be glued, but the structure is lost

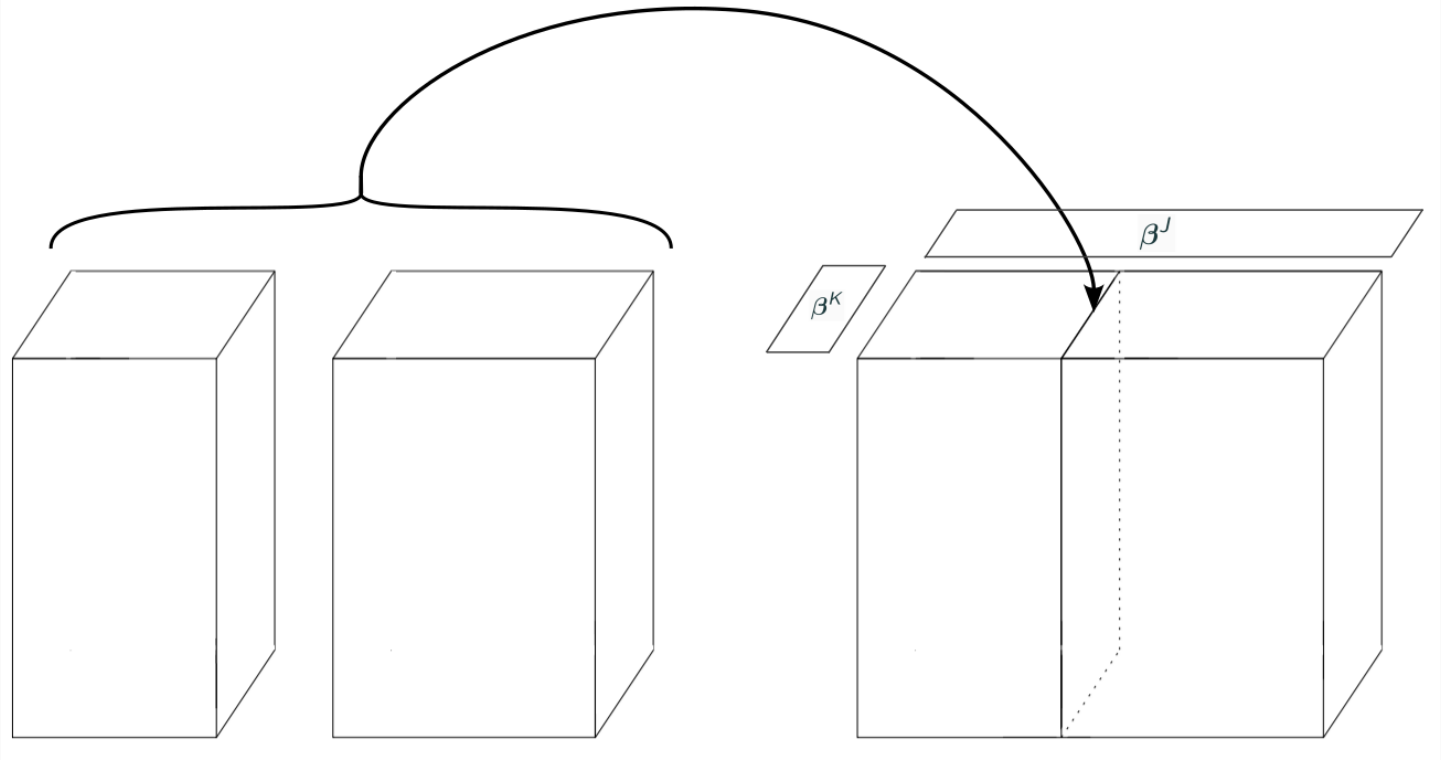


Figure 12: Multiway model with blocs of variables

Multiway Multiblock Logistic Regression

Solution: giving each block its own β^J and β^K

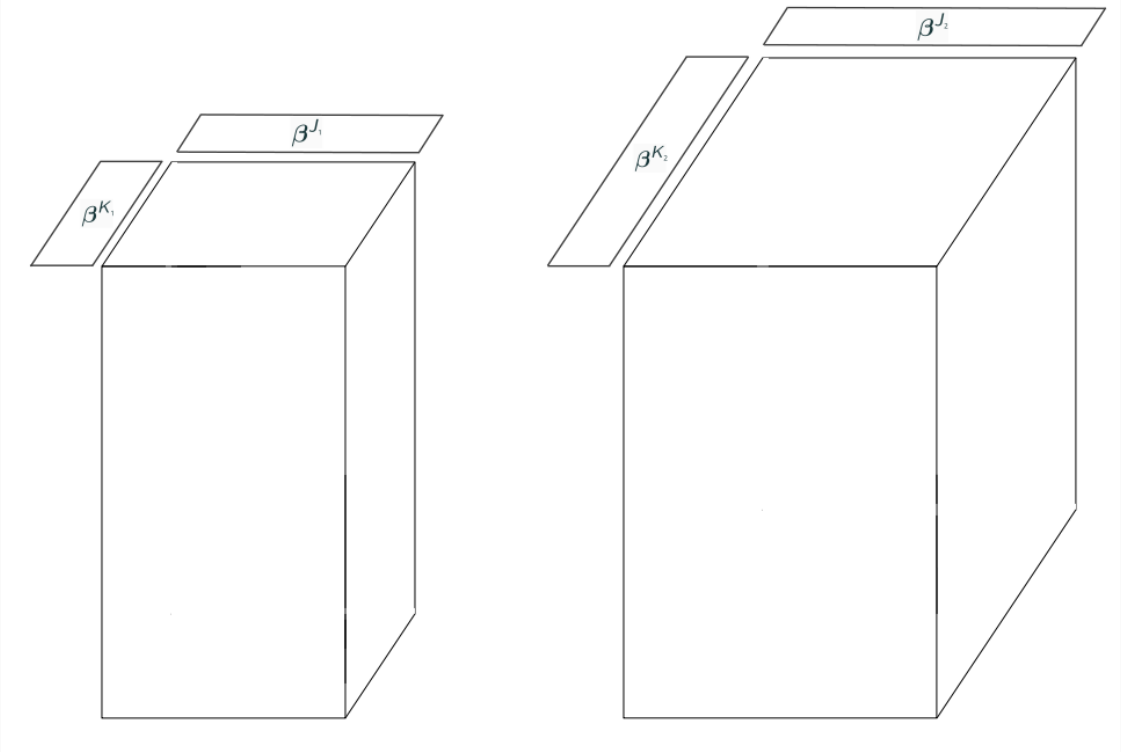


Figure 13: Multiway multiblock model for rank 1

Multiway 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 $\boldsymbol{\beta}^{l}$ 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 coefficients β (several blocks)
- Quality of the classification (AUC)
- Quality of the reconstruction of β (pictograms [4])

Illustration in 2D

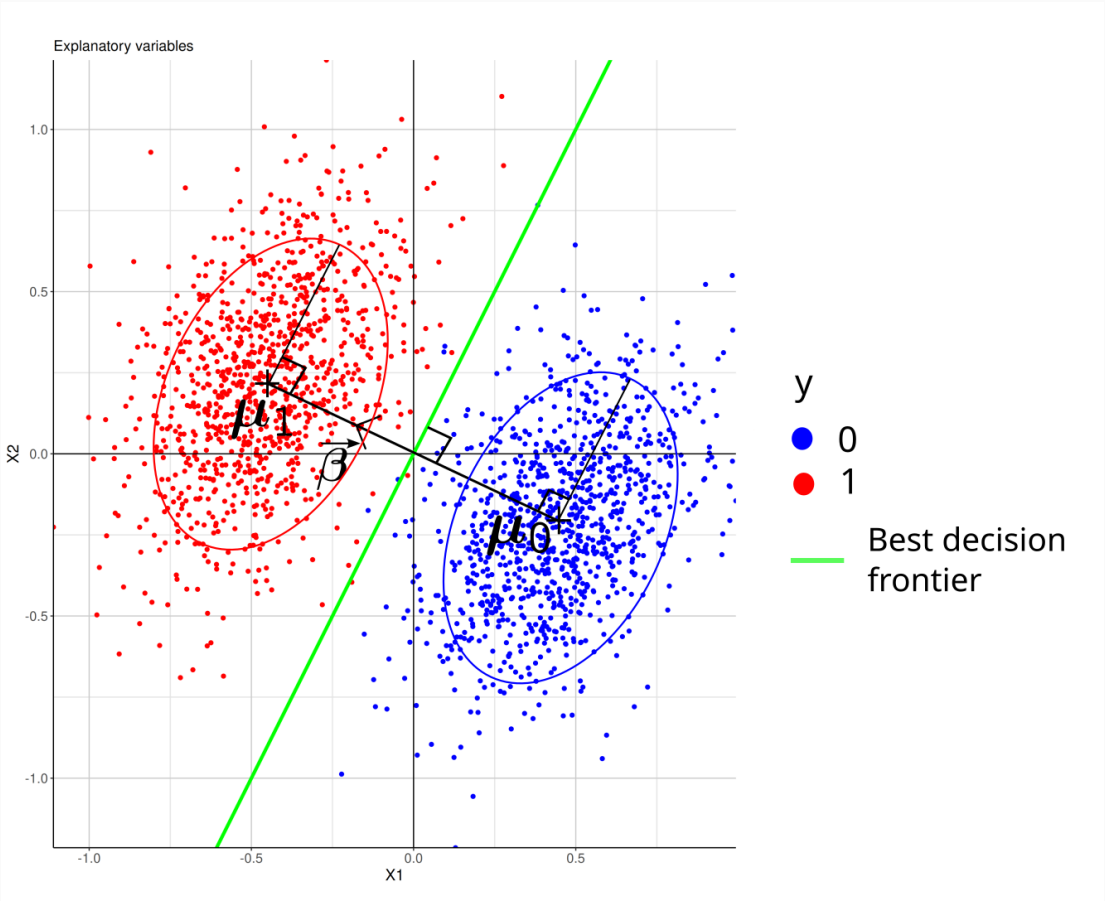


Figure 14: Example of explanatory variables for $\beta = (-2, 1)$

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 $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and common covariance matrix $\boldsymbol{\Sigma}$ such that:

- $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ colinear to $\boldsymbol{\beta}$
- One of the principal axis of $\boldsymbol{\Sigma}$ colinear to $\boldsymbol{\beta}$

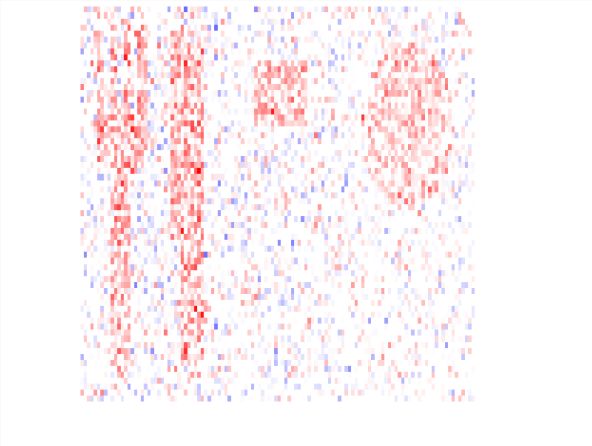
Separation of classes linked with eigenvalues of $\boldsymbol{\Sigma}$ (to be compared with $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|$)

AUC simulated data

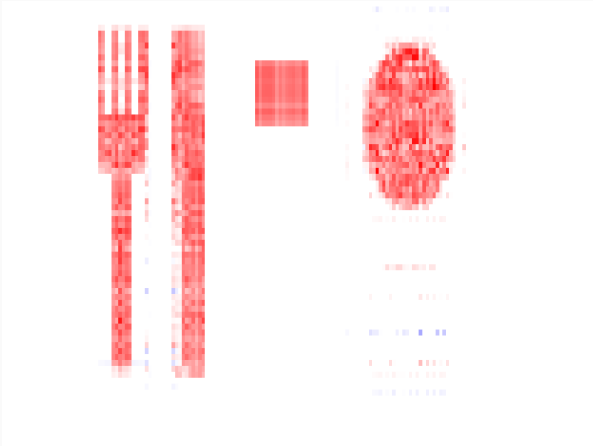
Table 1: AUC for each model on simulated data for 3000 individuals

$(\sigma_\beta, \sigma_{\text{noise}})$	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

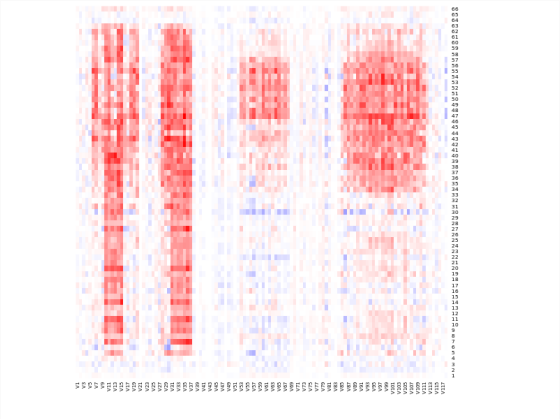
Reconstructed β



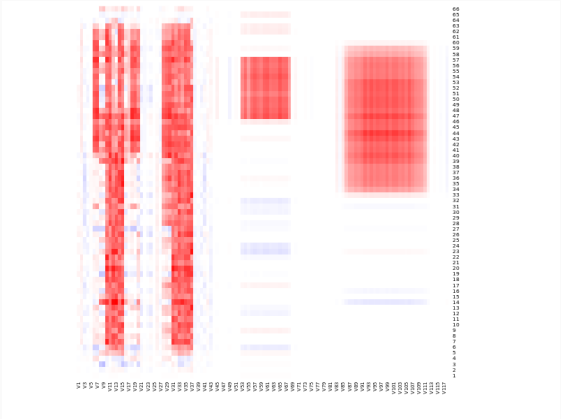
(a) lasso
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.5)$



(b) M.M., $R = (12, 1, 10)$
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.5)$



(c) multiway $R = 10$
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$



(d) M.M., $R = (6, 1, 1)$
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$

real data

Features extraction with pyradiomics [3]

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

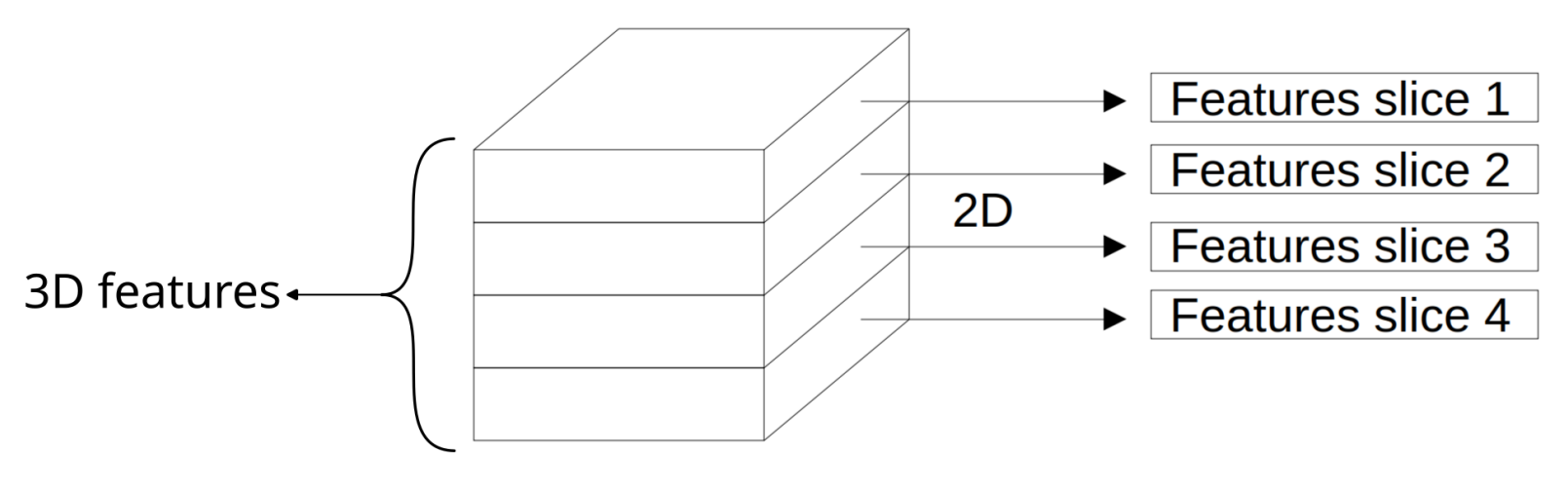


Figure 16: Features extraction with pyradiomics of an RMI image composed of 4 slices

Features extraction in 3D

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

But : Calculations of pyradiomics by pixels/ voxels.

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

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

Features 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)

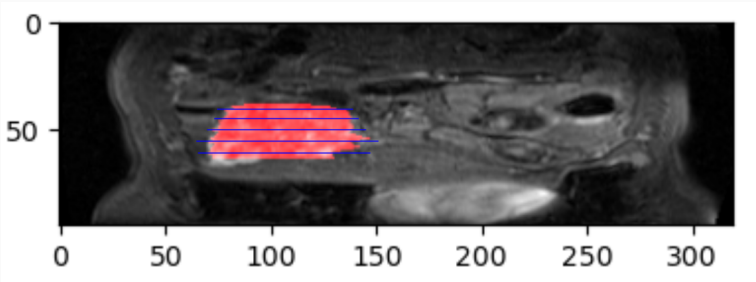


Figure 17: Extracting 5 slices in a big tumor

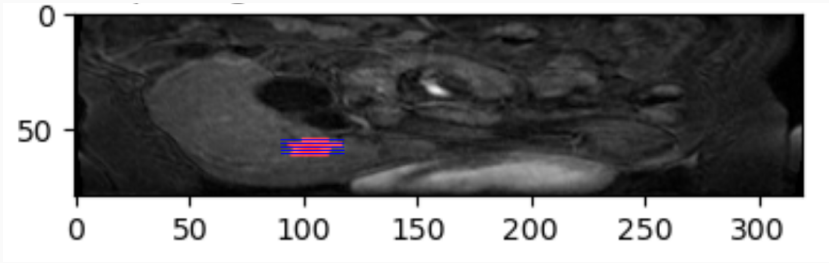


Figure 18: Extracting 5 slices in a small tumor

Features extraction in 2D

Every slice not equally informative

Features extraction in 2D

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

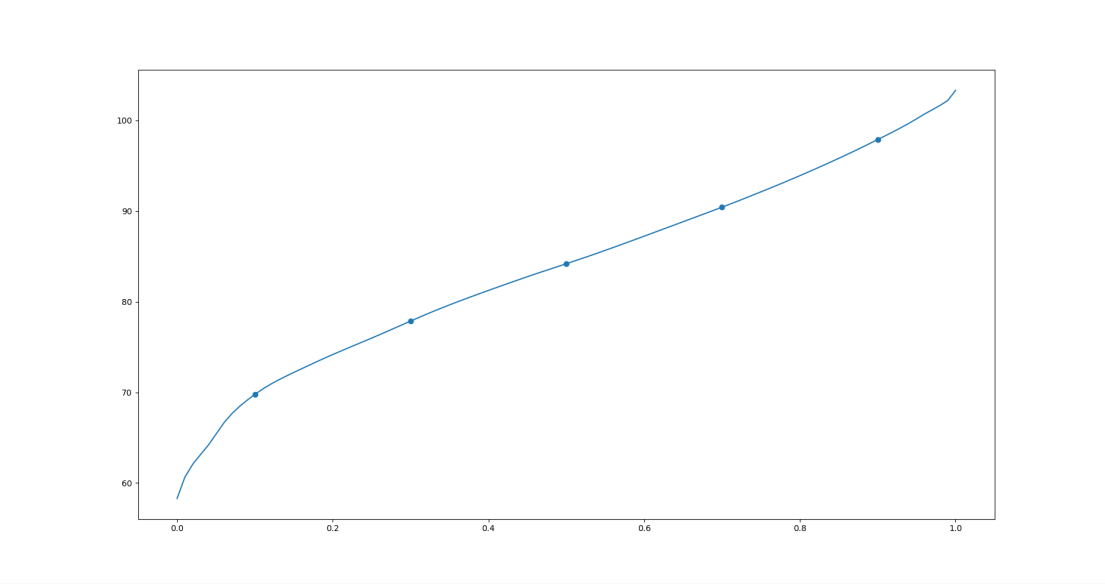


Figure 19: Curve of the depth travelled in the liver (in *mm*) as a function of the standardized cumulated volume of the tumor. The points represent the selected slices.

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

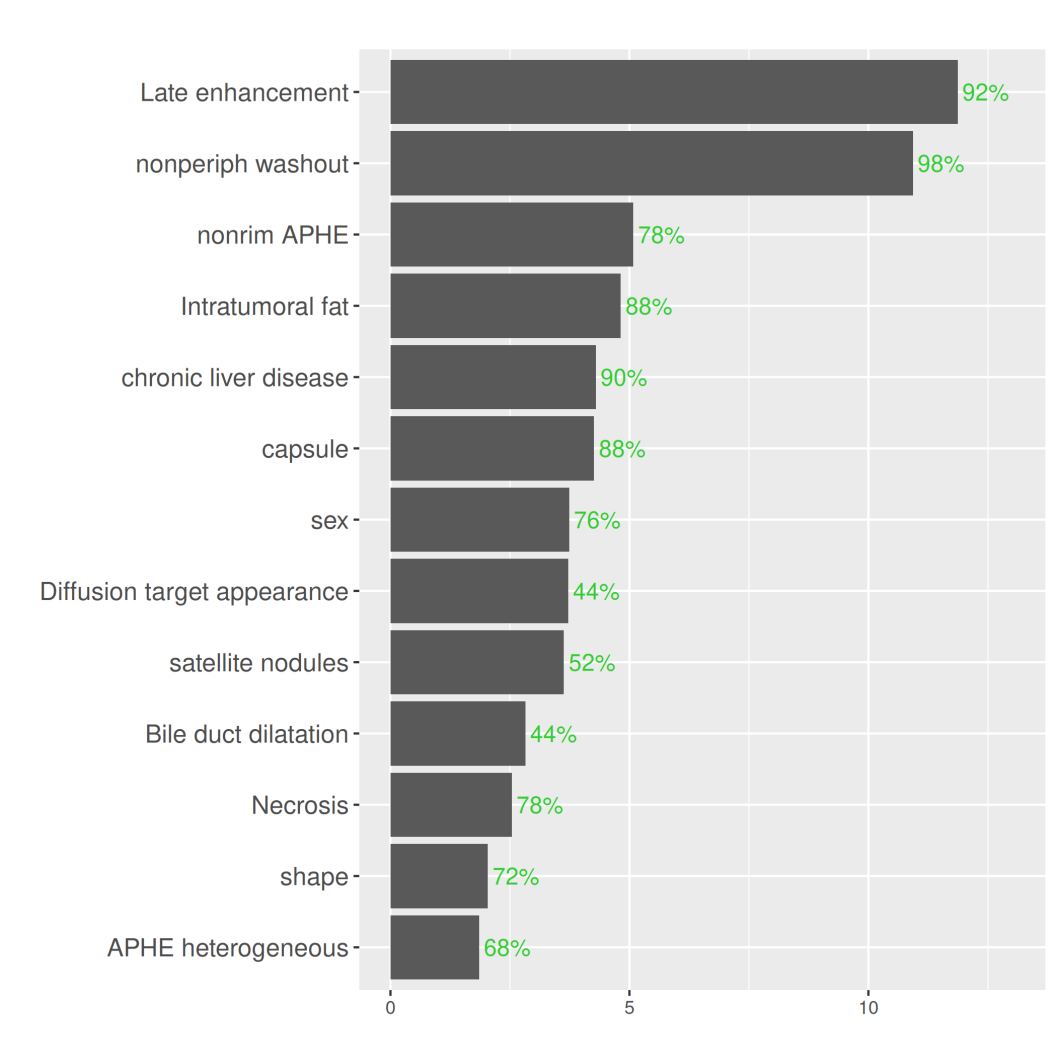


Figure 20: Features importance with lasso (in green percentage of runs with non null coefficient)

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 (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.