Internship report on machine learning analysis of multivariate medical data

Alexandre SELVESTREL (Class of 2023)

Université Paris-Saclay, CNRS, Centrale-Supélec,

Laboratoire des signaux et systèmes

Supervisors: Arthur Tenenhaus, Laurent Lebrusquet

Academic Supervisor for the Data and Information Sciences mention:

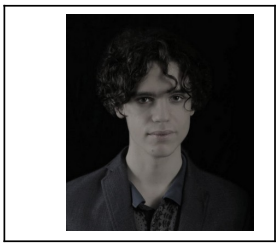
Antonio-José Silveti-Falls

Academic Supervisor for the Major Project Management filière:

Peggy Vicomte

Internship carried out from 2024/07/01 to 2024/12/09

Defense on November 29, 2024



CentraleSupélec Paris

Fiche-résumé du rapport de stage de fin d'études

FILIERE : Filière Management de grands projets MENTION : Sciences des données et de l'information
PROMOTION : P2023

NOM et Prénom de l'Élève: SELVESTREL Alexandre

NOM et adresse de la Société: Laboratoire des signaux et systèmes, Bâtiment IBM, Parc Paris-Saclay Université Rue Alfred Kastler, 91400 Orsay

Sujet du stage: Analyse par machine learning de données médicales multivariées

Résumé précis du rapport de stage:
(10 lignes dactylographiées maximum)

Ce stage en machine learning a un objectif double. Le premier est de proposer une classification des tumeurs du foie à partir de leur image IRM. Et le second est d'étudier des modèles de machine learning tensoriels, qui semblent particulièrement adaptés à ces données, et d'en proposer une amélioration (qui est également testée sur données simulées). Les résultats sur données simulées montrent une amélioration de l'interpolation du coefficient de régression par le nouveau modèle. Pour les tumeurs du foie, c'est finalement un modèle plus simple entraîné sur des features créées à l'oeuil nu par les radiologues qui donne les meilleurs résultats. Ces résultats sont prometteurs et laissent envisager un déploiment futur dans les hôpitaux d'une classification automatique des tumeurs du foie.

Date: 19/11/2024

Synthèse (version française)

Présentation générale. L'objectif de mon stage était de réaliser une classification automatique (via du machine learning) de tumeurs du foie, en se basant sur des IRMs et sur quelques données cliniques (âge, sexe du patient ...). Cette classification a donné un cadre dans lequel tester et améliorer des modèles tensoriels récents [1, 2] ainsi que vérifier si ceux-ci donnent de meilleures performances que les autres modèles. Ce stage a été effectué au laboratoire des signaux et systèmes (L2S) en partenariat avec l'assistance publique des hôpitaux de Paris (AP-HP). Sur le versan médical, nous avons pu bénéficier de l'aide de Sébastien Mulé, professeur à la faculté de santé, Université Paris-Est Créteil (UPEC) et Radiologie, chef du département imagerie de l'hôpital Henri Mondor.

Enjeux. Ce stage s'inscrit dans la cadre de la collaboration entre le L2S et l'AP-HP. Du point de vue du L2S, il s'agit de mettre à l'épreuve des méthodes de machine learning particulières, basées sur des tenseurs et qui semblent spécifiquement adaptées aux données étudiées. Par ailleurs, en me formant au machine learning appliqué au domaine médical, le laboratoire s'assure qu'en poursuivant en doctorat, je disposerai des compétences nécessaires pour être immédiatement opérationnel.

Pour l'AP-HP, l'enjeu est de faire progresser la recherche sur le cancer du foie. En effet, la détermination de la nature de la tumeur du foie d'un patient est un problème complexe auquel il n'existe pas de solution complètement satisfaisante à l'heure actuelle. Or, les médecins disposant des IRMs des patients malades, il serait dommage de ne pas les utiliser pour tenter de proposer un outil de diagnostic automatique. De plus, cet outil pourraît être utile aux médecins pour déterminer de nouveaux indicateurs qui caractérisent la classe d'une tumeur -en observant ce qui est considéré comme important par le modèle de classification automatique.

Solutions et résultats:. Nous avons commencé par implémenter des modèles de machine learning basiques (régression logistique lasso et random forest) sur les données de cancer du foie. Cela nous a permis d'établir une valeur de référence pour la performance de la classification (AUC = 0.68). Nous avons ensuite cherché à améliorer ce score en développant une régression logistique tensorielle. Par ailleurs, nous avons proposé une variante de ce modèle tensoriel en intégrant la structure en blocs des données (section 2.2). Ce nouveau modèle nous a permis de retrouver le score de départ, mais pas de l'améliorer.

Afin de s'assurer de la pertinence de notre modèle, nous avons alors cherché à tester son efficacité sur des données simulées. Sur ces données, notre modèle tensoriel a montré des performances bien meilleures que les modèles non tensoriels, permettant même de mieux retrouver le coefficient de régression que d'autres modèles tensoriels déjà existant. Cela nous a permis de conclure que ce modèle était pertinent dans certains cas et que l'impossibilité d'améliorer les performances sur les données médicales était probablement due à la mauvaise qualité de ces données. Après plus d'un mois de travail pour améliorer la qualité des données, je me suis rendu à l'Hôpital Henri Mondor afin de parler de mes résultats avec Sébastien Mulé sur son lieu de travail (et non au L2S comme les fois précédentes). Cette visite a permis de découvrir l'existence d'un autre jeu de données, ommises jusqu'à présent, beaucoup plus simples (seulement une quinzaine de variables par individu), donnant des résultats bien meilleures que les données précédentes quand on les traite par machine learning (AUC = 0.95). Nous avons été surpris par l'arrivée au dernier moment de ces données qui, bien que de bonne qualité, ne sont pas adaptées aux modèles tensoriels. Nous ne les mentionnons donc pas dans la partie "article" du rapport mais nous les présentons juste après.

Synthesis (english version)

Overview. The aim of my internship was to carry out an automatic classification (via machine learning) of liver tumors, based on MRI scans and some clinical data (patient age, sex, etc.). This classification provided a framework within which to test and improve recent tensor models [1, 2] as well as verify whether these perform better than other models. The internship was carried out at the Laboratoire des signaux et systèmes (L2S), in partnership with the Assistance Publique des Hôpitaux de Paris (AP-HP). On the medical side, we benefited from the help of Sébastien Mulé, teacher at the Faculty of Health, Université Paris-Est Créteil (UPEC) and Radiology, Head of the Imaging Department at Henri Mondor Hospital.

Stakes. This internship is part of the collaboration between L2S and AP-HP. From the point of view of the L2S, the aim is to put to the test particular machine learning methods, based on tensors, which seem specifically adapted to the data under study. What's more, by training me in machine learning applied to the medical field, the laboratory has ensured that if I go on to do a PhD, I will have the necessary skills to be immediately operational.

For AP-HP, the challenge is to advance research into liver cancer. Determining the nature of a patient's liver tumor is a complex problem for which there is currently no completely satisfactory solution. Since doctors have access to MRI scans of patients with liver tumours, it would be a shame not to use them to try and offer an automatic diagnostic tool. In addition, this tool could be useful for radiologists to determine new indicators that characterize the class of a tumor - by observing what is considered important by the automatic classification model

Solutions and results:. We started by implementing basic machine learning models (lasso logistic regression and random forest) on liver cancer data. This allowed us to establish a reference value for classification performance (AUC = 0.68). We then sought to improve this score by developing a tensor logistic regression. Furthermore, we proposed a variant of this tensor model by integrating the block structure of the data (section 2.2). This new model allowed us to recover the initial score, but not to improve it. In order to ensure the relevance of our model, we then sought to test its effectiveness on simulated data. On this data, our tensor model showed much better performance than non-tensor models, even allowing us to find the regression coefficient better than other existing tensor models. This allowed us to conclude that this model was relevant in certain cases and that the impossibility of improving performance on medical data was probably due to the poor quality of this data. After more than a month of work to improve the quality of the data, I went to Henri Mondor Hospital to discuss my results with Sébastien Mulé at his workplace (and not in the L2S as in previous times). This visit allowed me to discover the existence of another dataset, omitted until now, much simpler (only about fifteen variables per individual), giving much better results than the previous data when processed by machine learning (AUC = 0.95). We were surprised by the last-minute arrival of these data, which, although of good quality, are not suitable for tensor models. We therefore do not mention them in the "article" part of the report but present them just after.

Contents

1 Introduction **7**

2 Methodology **9**
- 2.1 Tensorial data and notations . . . 9
- 2.2 Machine learning models . . . 10
  - 2.2.1 Multiway logistic regression . . . 10
  - 2.2.2 Multiway and Multibloc Logistic Regression (MMLR) . . . 12
- 2.3 Simulated data generation . . . 15
  - 2.3.1 Regression parameter structure . . . 15
  - 2.3.2 Generation of explanatory variables . . . 15

3 Real dataset **20**
- 3.1 Presentation of real data . . . 20
- 3.2 feature extraction in 3D . . . 20
- 3.3 Feature extraction in 2D . . . 21
- 3.4 Extraction of healthy liver parts . . . 22

4 results **24**
- 4.1 Simulated data . . . 24
- 4.2 real data . . . 27

5 Conclusion **27**

6 Latest results: not mentioned in the article **29**

7 Retrospective and Perspectives on the Internship **30**
- 7.1 Possible extensions of the work . . . 30
- 7.2 Assessment and Reflection . . . 30
- 7.3 Competency framework . . . 32
  - 7.3.1 Complexity . . . 32
  - 7.3.2 Engineering Profession . . . 33
  - 7.3.3 Innovate and Undertake . . . 33
  - 7.3.4 Value Creation . . . 34
  - 7.3.5 Intercultural . . . 34
  - 7.3.6 Digital . . . 35
  - 7.3.7 Persuade . . . 36
  - 7.3.8 Project Team . . . 36
  - 7.3.9 Ethics and Sustainability . . . 37

8 Acknowledgments **38**

Appendix A Hyperparameters for simulated data 40
Appendix A.1 Data generation 40
Appendix A.2 Cross validation of models 40

Appendix B Parameters used for feature extraction with pyradiomics 41

Appendix C Importance of features 42

Multiway multiblock logistic regression to classify liver tumors from MRI images

Alexandre SELVESTREL

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-sur-Yvette, 91190, France

Abstract

In this paper, we introduce a novel tensor logistic regression model designed for cases where the data is distributed across multiple tensors (referred to as blocs), each potentially having a different number of modes. We call this model Multiway Multibloc Logistic Regression (MMLR) and we also provide its fitting procedure. This model extends existing tensor logistic regressions [1, 2] to accommodate more complex data structures. This model is tested on simulated data and on real data (liver cancer MRI) in order to compare it to other models, tensor (multiway logistic regression [2]) or non-tensor (lasso and group lasso [3]). The results of the MMLR are better than those of non-tensor models on simulated data and comparable to the state of the art on real data. In particular on simulated data, the regression coefficient is found more finely by the MMLR than by all the other models tested.

Keywords: multiway data, multiblock data, MRI, tensor

1. Introduction

In many scientific fields, datasets are organized into multidimensional arrays that go beyond simple vectors or matrices. These arrays, called tensors, provide a natural way to represent data with multiple dimensions, also called in that case "modes". A tensor generalizes the concept of a matrix to higher dimensions: a vector is a tensor of order 1, a matrix is a tensor of order 2, and tensors of order 3 or more can encode relationships across three or more modes.

To formalize this, let us consider a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ of order 3, where I, J, and K represent the dimensions along the three modes. Here, I might correspond to the number of individuals in a study, J to the number of observed variables, and K to the number of modalities or conditions under which these variables are measured. Each entry x_{ijk} in the tensor encodes the value of the j-th variable for the i-th individual under the k-th modality (see Fig. 1). This structure allows the tensor to capture multi-dimensional relationships in a compact and meaningful way. For instance, while a matrix can describe how features vary across individuals, a tensor can extend this description to include variations across multiple modalities or acquisition phases.

Tensors are particularly useful in applications where the relationships across modes are essential for understanding the data. For example, in medical imaging, radiomic features extracted from scans can be organized into tensors, where the modes might correspond to patients, extracted features, and different imaging phases. Preserving this structure enables models to exploit the inherent organization of the data, rather than flattening it into a less informative matrix.

Logistic regression, a classical tool for binary classification, has been extended to incorporate tensor data, allowing the regression parameters to reflect the multi-dimensional structure. These tensor logistic regression models have shown promise in leveraging the relationships across modes to improve both predictive accuracy and interpretability [1, 2]. However, existing methods often assume that the data can be represented as a single tensor, which may not hold for more heterogeneous datasets. In this work, we propose the MMLR, an extension designed to handle datasets that naturally decompose into multiple tensors, each potentially with a different number of modes.

To evaluate the proposed approach, we test it on both simulated data and a real-world application in liver cancer classification. Liver tumors, which are among the most prevalent and deadly forms of cancer, are primarily classified into two types: hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCK). Effective classification is crucial for determining appropriate treatment, but existing diagnostic methods face significant limitations. Microscopy, while reliable, is invasive and provides only a partial view of the tumor. Non-invasive radiographic imaging, such as MRI with contrast injection, captures a global view but is challenging to interpret due to overlapping characteristics of HCC and CCK.

In this study, we represent liver tumors as third-order tensors constructed from radiomic features extracted from MRI scans. For each patient, four MRI images are acquired at different time points, corresponding to distinct phases of contrast enhancement: arterial, portal, venous, and late. These temporal phases capture specific characteristics of the tumor's vascularization, providing complementary information. The features extracted from each image are organized into a tensor with dimensions $I \times J \times K$, where I is the number of patients, J is the number of features measured from each MRI image, and $K = 4$ corresponds to the four acquisition phases. This tensorial representation enables the application of our MMLR, which we compare against classical logistic regression, lasso-penalized logistic regression, and group lasso methods [3].

This paper is organized as follows. In section 2.2, we present tensor-based logistic regression models, starting with existing approaches and introducing the proposed MMLR. Then, section 2.3 and 3 describe the data used in this study, respectively simulated datasets and real-world MRI data from liver cancer patients. Finally, in section 4, we detail the results of our experiments, comparing the performances of the MMLR to state of the art methods on both types of data.

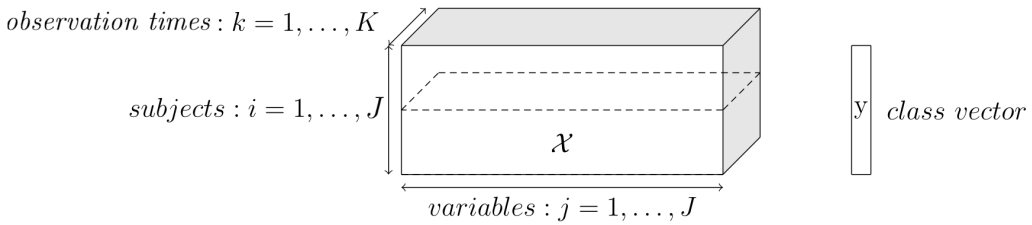


Fig. 1: Example of order 3 tensor from [2]

2. Methodology

2.1. Tensorial data and notations

We designate as tensor valued data any data where the explanatory variables are structured along several dimensions. To avoid confusion with the notion of dimension of a vector space we call these dimensions modes in the following. Then, instead of having a matrix of explanatory variables $\mathbf{X} = (x_{ij})_{i\in[\![1,I]\!],j\in[\![1,J]\!]}$ (where i is the individual and j is the quantity of interest), we get a tensor of explanatory variables $\underline{\mathbf{X}} = (x_{ijk_1k_2\ \ldots\ k_M})_{i\in[\![1,I]\!],j\in[\![1,J]\!],k_1\in[\![1,K_1]\!]\ \ldots\ k_m\in[\![1,K_M]\!]}$ (where i is the individual, j is the quantity of interest and where for $m \in [\![1, M]\!]$, k_m is the k_m-th modality of the m-th mode of the data). We use the notations introduced in [4], especially concerning matricization (see section 2.4 of [4]). For a self-contained document, we report some of them below in addition to the notations specific to this paper.

- The concatenation of two matrices $\mathbf{A}$ and $\mathbf{B}$ by juxtaposing their columns side by side is denoted $[\mathbf{A}\ \mathbf{B}]$.
- To avoid overuse of the symbol ^{T}, we also define a notation to designate the juxtaposition of two matrices one below the other. Thus, the matrix defined by block with $\mathbf{A}$ above $\mathbf{B}$ is denoted $[\mathbf{A}; \mathbf{B}]$. It can also be written $[\mathbf{A}^T\ \mathbf{B}^T]^T$ but this multiplies the ^{T} symbols, which impairs legibility.
- Since vectors are column matrices, using the same notation, we write the concatenation of two vectors $\mathbf{u}$ and $\mathbf{v}$ as follows: $[\mathbf{u}; \mathbf{v}]$.
- The vector (column) whose elements are $(u_i)_{i\in[\![1,I]\!]}$ is denoted $(u_1, u_2, \ \cdots\ u_I)$.
- If $\mathbf{X}$ is a matrix of explanatory variables, $\mathbf{x}_i$ is the vector (column) composed of the i-th row of $\mathbf{X}$.
- The vector of length I filled with 1 is denoted by $\mathbb{1}_I$.
- We denote $\mathrm{Diag}(\mathbf{u})$ the diagonal matrix whose diagonal is the vector $\mathbf{u}$.
- We denote $\mathbf{X}_{(1)}$ the transformation of a tensor $\underline{\mathbf{X}}$ into a matrix (by unfolding it), using the first mode to form the rows. This operation is called matricization. The order of the columns is lexicographic in the index of the modes. For a third order tensor, it gives $\mathbf{X}_{(1)} = [\mathbf{X}_{::1}\ \ \ldots\ \ \mathbf{X}_{::K}]$
- The Kronecker product of two matrices is denoted "$\otimes$"

2.2. Machine learning models

In this section, we introduce the tensor models used in this article: the multiway logistic regression and the MMLR. For clarity and to avoid overcomplicating the notation, we focus on the case where $\underline{\mathbf{X}}$ is a third-order tensor, noting that the extension to higher-order tensors is straightforward.

2.2.1. Multiway logistic regression

We present the lasso penalized rank-R multiway logistic regression. This model is described in [2]. The fundamental idea of the model is to decompose the parameter $\boldsymbol{\beta}_{\text{tens}} \in \mathbb{R}^{JK}$ associated with the tensor explanatory variables of the logistic regression as:

$$\boldsymbol{\beta}_{\text{tens}} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \tag{1}$$

with for all $r \in [\![1, R]\!]$, $\boldsymbol{\beta}_r^J \in \mathbb{R}^J$ and $\boldsymbol{\beta}_r^K \in \mathbb{R}^K$. To take account of the M tabular variables (non tensorial), we associate them with a coefficient $\boldsymbol{\beta}_{\text{tab}} \in \mathbb{R}^M$. In this way, the parameter $\boldsymbol{\beta}$ of the logistic regression is written: $[\boldsymbol{\beta}_{\text{tens}}; \boldsymbol{\beta}_{\text{tab}}]$.
As usual with logistic regressions, we consider that each realization of the explained variable y_i ($i \in [\![1, I]\!]$) follows an independent Bernoulli law conditionally on $\mathbf{x_i}$. For logistic regression, this probability is parametrzied by $\boldsymbol{\beta}$ and defined as

$$\mathbb{P}(y_i = 1 \,|\, \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta} - \beta_0)} \tag{2}$$

where $\beta_0 \in \mathbb{R}$ is the intercept.

We set $\boldsymbol{\beta}^J = \left[\boldsymbol{\beta}_1^J; \ \cdots \ ; \boldsymbol{\beta}_R^J\right]$ and $\boldsymbol{\beta}^K = \left[\boldsymbol{\beta}_1^K; \ \cdots \ ; \boldsymbol{\beta}_R^K\right]$. In order to promote sparsity, we add L1 type penalties. It yields to consider the following optimization problem:

$$\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}} = \underset{\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}}}{\operatorname{argmin}} \left[-\sum_{i=1}^{I} \log(\mathbb{P}(y_i = 1 \,|\mathbf{x}_i)) + \lambda \left(\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 \right) \right] \tag{3}$$

Optimization is performed by alternating directions between $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}\right]$ and $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$. We note that optimizing the loss function in each of these directions boils down to perform a simple logistic regression with a lasso penalty. Indeed, if we denote C the loss function of classical logistic regression penalized by lasso (for any $K_0 \in \mathbb{N}^*$):

$$C : \begin{cases} \mathbb{R} \times \mathbb{R}^{K_0} \times \mathbb{R}^{I \times K_0} \times \mathbb{R}^I \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ \qquad (\beta_0, \boldsymbol{\beta}, \mathbf{X}, \mathbf{y}, \lambda) & \longmapsto & -\sum_{i=1}^{I} [y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))] + \lambda \|\boldsymbol{\beta}\|_1 \end{cases} \tag{4}$$

optimizing the overall loss function with respect to $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{uni}}\right]$ is equivalent to solve

$$\underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^{JR+M}}{\operatorname{argmin}} C(\beta_0, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda) \tag{5}$$

Where $\mathbf{Q}^J$ and $\mathbf{Z}^J$ are defined as follows:

$$\mathbf{Z}^J = [\mathbf{Z}_1^J \ \cdots \ \mathbf{Z}_R^J \ \mathbf{X}_{\text{tab}}] \tag{6}$$

$$\text{where } \forall r \in [\![1, R]\!], \quad \mathbf{Z}_\mathbf{r}^\mathbf{J} = \sum_{k=1}^{K} (\beta_r^K)_k \, \mathbf{X}_{::k} \quad (\mathbf{Z}_r^J \in \mathbb{R}^{N\times J}) \tag{7}$$

$$\text{and} \quad \mathbf{Q}^J = \text{Diag}([\|\boldsymbol{\beta}_1^K\|_1^{-1}\mathbb{1}_J; \ \cdots \ ; \|\boldsymbol{\beta}_R^K\|_1^{-1}\mathbb{1}_J; \mathbb{1}_M]) \tag{8}$$

Girka et al. [2] demonstrate this result by noting that for $i \in [\![1, I]\!]$,

$$\mathbf{x}_{(1)i}^T \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\right) = \sum_{r=1}^{R} \left[(\mathbf{x}_{(1)i}^T \left(\boldsymbol{\beta}_r^K \otimes \mathbf{I}_J\right)\right] \boldsymbol{\beta}_r^J \tag{9}$$

$$= \sum_{r=1}^{R} (\mathbf{z}_r^J)_i^T \boldsymbol{\beta}_r^J \tag{10}$$

and that

$$\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 = \|\mathbf{R}_{\text{tens}}^J \boldsymbol{\beta}^J\|_1 \tag{11}$$

$$\text{with} \quad \mathbf{R}_{\text{tens}}^J = \text{Diag}([\|\boldsymbol{\beta}_1^K\|_1\mathbb{1}_J; \ \cdots \ ; \|\boldsymbol{\beta}_R^K\|_1\mathbb{1}_J]) \tag{12}$$

Thus,

$$(\mathbf{x}_{\text{tot}})_i^T \boldsymbol{\beta} = (\mathbf{z}_i^J)^T [\boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}] \tag{13}$$

$$\text{and} \quad \sum_{i=1}^{I} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 = \|(\mathbf{Q}^J)^{-1}\boldsymbol{\beta}\|_1 \tag{14}$$

This justifies the previous results

For optimization with respect to $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$, the method follows the same steps. The only difference concerns the definition of $\mathbf{Z}^K$. It is:

$$\mathbf{Z}^K = [\mathbf{Z}_1^K \ \cdots \ \mathbf{Z}_R^K \ \mathbf{X}_{\text{tab}}] \tag{15}$$

$$\text{with} \forall r \in [\![1, R]\!] \quad \mathbf{Z}_r^K = \sum_{j=1}^{J} (\beta_r^J)_j \mathbf{X}_{:j:} \tag{16}$$

This is justified by:

$$\mathbf{x}_{(1)_i}^T \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \right) = \sum_{r=1}^{R} \left[\left(\mathbf{x}_{(1)_i}^T \left(I_K \otimes \boldsymbol{\beta}_r^J \right) \right] \boldsymbol{\beta}_r^K \right. \tag{17}$$

$$= \sum_{r=1}^{R} (\mathbf{z}_r^K)_i^T \boldsymbol{\beta}_r^K \tag{18}$$

Note: The rank-1 version of this model was firstly developed in [1].

2.2.2. Multiway and Multibloc Logistic Regression (MMLR)

This section presents lasso penalized MMLR. This model draws heavily on the multiway logistic regression we have just presented, while also taking into account a block structure of tensor data. More precisely, each of these blocks $\underline{\mathbf{X}}^l$ is a tensor and has its own independent regression coefficient $\boldsymbol{\beta}_l$. We also allow each block to have its own rank R_l. As tabular quantities are not measured according to several modalities, they are not placed in any particular block. They will be included in the model in the same way as in the multiway case. Mathematically, we define the model as follows:

Let $L \in \mathbb{N}^*$ denote the number of blocks. For any $l \in [\![1, L]\!]$, let J_l be the number of tensorial variables in block l. The matrix of explanatory variables is defined by $\mathbf{X}' = [\mathbf{X}_{(1)}^1 \cdots \mathbf{X}_{(1)}^L]$ and $\boldsymbol{\beta}$ is the regression coefficient with respect to this matrix. The structure of $\boldsymbol{\beta}$ is given by blocks. It is:

$$\boldsymbol{\beta} = \left[\sum_{r=1}^{R_1} \boldsymbol{\beta}_{(1,r)}^K \otimes \boldsymbol{\beta}_{(1,r)}^J ; \ \cdots \ ; \sum_{r=1}^{R_L} \boldsymbol{\beta}_{(L,r)}^K \otimes \boldsymbol{\beta}_{(L,r)}^J \ ; \boldsymbol{\beta}_{\text{tab}} \right] \tag{19}$$

With for all $l \in [\![1, L]\!]$, we have $r \in [\![1, R_l]\!]$, $\boldsymbol{\beta}_{(l,r)}^J \in \mathbb{R}^{J_l}$ and $\boldsymbol{\beta}_{(l,r)}^K \in \mathbb{R}^K$.

We call $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ the vectors

$$\boldsymbol{\beta}^J = \left[\boldsymbol{\beta}_{(1,1)}^J ; \ \cdots \ ; \boldsymbol{\beta}_{(1,R_1)}^J ; \ \cdots \ \cdots \ ; \boldsymbol{\beta}_{(L,1)}^J \ \cdots \ ; \boldsymbol{\beta}_{(L,R_L)}^J \right] \tag{20}$$

$$\boldsymbol{\beta}^K = \left[\boldsymbol{\beta}_{(1,1)}^K ; \ \cdots \ ; \boldsymbol{\beta}_{(1,R_1)}^K ; \ \cdots \ \cdots \ ; \boldsymbol{\beta}_{(L,1)}^K \ \cdots \ ; \boldsymbol{\beta}_{(L,R_L)}^K \right] \tag{21}$$

In a similar way to what is done in the multiway model, we adapt the lasso penalty, so that the new optimization problem becomes:

$$\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}} = \underset{\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}}}{\text{argmin}} \left(- \sum_{i=1}^{I} \log(\mathbb{P}(y_i = 1 \,|\mathbf{x}_i)) + \sum_{l=1}^{L} \sum_{r=1}^{R_l} \|\boldsymbol{\beta}_{(l,r)}^K \otimes \boldsymbol{\beta}_{(l,r)}^J\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 \right) \tag{22}$$

Once again, this problem is solved by alternating optimization directions $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}\right]$ and $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$. Each of these two problems can be reduced to a lasso-penalized classical logistic regression

Indeed, optimizing according to $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}\right]$ is equivalent to searching

$$\underset{(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}}\, C(\beta_0, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda) \tag{23}$$

Where $\mathbf{Q}^J$ and $\mathbf{Z}^J$ are defined as follows:

$$\mathbf{Z}^J = [\mathbf{Z}^J_{(1,1)} \ \cdots \ \mathbf{Z}^J_{(1,R_1)} \ \cdots \ \cdots \ \mathbf{Z}^J_{(L,1)} \ \cdots \ \mathbf{Z}^J_{(L,R_L)} \ \mathbf{X}_{\text{tab}}] \tag{24}$$

$$\text{where } \forall r \in [\![1, R_l]\!], \quad \mathbf{Z}^J_{(l,r)} = \sum_{k=1}^{K} \left(\beta^K_{(l,r)}\right)_k \mathbf{X}^l_{::k} \quad \left(\mathbf{Z}^J_{(l,r)} \in \mathbb{R}^{I\times J_l}\right) \tag{25}$$

$$\mathbf{Q}^J = \text{Diag}([\|\boldsymbol{\beta}^K_{(1,1)}\|_1^{-1}\mathbb{1}_{d_1}; \ \cdots; \|\boldsymbol{\beta}^K_{(1,R_1)}\|_1^{-1}\mathbb{1}_{d_1}; \ \cdots \cdots; \|\boldsymbol{\beta}^K_{(L,1)}\|_1^{-1}\mathbb{1}_{J_L}; \ \cdots; \|\boldsymbol{\beta}^K_{(L,R_L)}\|_1^{-1}\mathbb{1}_{J_L}; \mathbb{1}_M]) \tag{26}$$

The demonstration of this result is similar to that of the multiway case. Indeed, we note that

$$(\mathbf{x}')_i^T \left[\sum_{r=1}^{R_1} \boldsymbol{\beta}^K_{(1,r)} \otimes \boldsymbol{\beta}^J_{(1,r)}; \ \cdots; \sum_{r=1}^{R_L} \boldsymbol{\beta}^K_{(L,r)} \otimes \boldsymbol{\beta}^J_{(L,r)}\right] = \sum_{l=1}^{L}\sum_{r=1}^{R_l} \left(\mathbf{x}^l\right)_i^T \left(\boldsymbol{\beta}^K_{(l,r)} \otimes \boldsymbol{\beta}^J_{(l,r)}\right) \tag{27}$$

$$= \sum_{l=1}^{L}\sum_{r=1}^{R_l} \left[\left(\mathbf{x}^l\right)_i^T \left(\boldsymbol{\beta}^K_{(l,r)} \otimes I_{J_l}\right)\right] \boldsymbol{\beta}^J_{(l,r)} \tag{28}$$

$$= \sum_{l=1}^{L}\sum_{r=1}^{R_l} \left(\mathbf{z}^J_{(l,r)}\right)_i^T \boldsymbol{\beta}^J_{(l,r)} \tag{29}$$

And that

$$\sum_{l=1}^{L}\sum_{r=1}^{R_l} \|\boldsymbol{\beta}^K_{(l,r)} \otimes \boldsymbol{\beta}^J_{(l,r)}\|_1 = \|\mathbf{R}^J_{\text{tens}}\boldsymbol{\beta}^J\|_1 \tag{30}$$

$$\text{with} \quad \mathbf{R}^J_{\text{tens}} = \text{Diag}([\|\boldsymbol{\beta}^K_{(1,1)}\|_1\mathbb{1}_{d_1}; \ \cdots; \|\boldsymbol{\beta}^K_{(1,R_1)}\|_1\mathbb{1}_{d_1}; \ \cdots \cdots; \|\boldsymbol{\beta}^K_{(L,1)}\|_1\mathbb{1}_{J_L}; \ \cdots; \|\boldsymbol{\beta}^K_{(L,R_L)}\|_1\mathbb{1}_{J_L}; \mathbb{1}_M]) \tag{31}$$

We deduce that

$$[\mathbf{x}'_i; \mathbf{x}_{\text{tab}_i}]\boldsymbol{\beta} = (\mathbf{z}^J_i)^T[\boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}] \tag{32}$$

$$\text{and} \quad \sum_{l=1}^{L}\sum_{r=1}^{R_l} \|\boldsymbol{\beta}^K_{(l,r)} \otimes \boldsymbol{\beta}^J_{(l,r)}\|_1 + \|\boldsymbol{\beta}_{\text{uni}}\|_1 = \|(\mathbf{Q}^J)^{-1}\boldsymbol{\beta}\|_1 \tag{33}$$

Wich justifies the previous results.

For optimization with respect to $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$, the method is similar. The only difference

concerns the form of $\mathbf{Z}^K$. It is written as:

$$\mathbf{Z}^K = [\mathbf{Z}^K_{(1,1)} \ \cdots \ \mathbf{Z}^K_{(1,R_1)} \ \cdots \ \cdots \ \mathbf{Z}^K_{(L,1)} \ \cdots \ \mathbf{Z}^K_{(L,R_L)} \ \mathbf{X}_{\text{tab}}] \tag{34}$$

$$\text{where } \forall r \in [\![1, R_l]\!], \quad \mathbf{Z}^K_{(l,r)} = \sum_{j=1}^{J_l} \mathbf{X}^l_{:j:} \left(\beta^J_{(l,r)}\right)_j \quad \left(\mathbf{Z}^K_{(l,r)} \in \mathbb{R}^{I \times K}\right) \tag{35}$$

The justification of that last result is similar to the one used in the multiway case.

The algorithm associated with the lasso penalized MMLR is summarized below

Fitting algorithm for MMLR:

Inputs

- $\epsilon > 0$, $\lambda > 0$, $R \in \mathbb{N}^*$
- $\boldsymbol{\beta}^{K(0)} \in \mathbb{R}^{LRK}$

Treatment

- $q \leftarrow 0$

Repeat

- Construct $\mathbf{Z}^J$ according to eqs. (24) and (25)
- Construct $\mathbf{Q}^J$ according to eq. (26)
- $(\beta_0^{J(q)}, \boldsymbol{\beta}^{J(q)}) \longleftarrow \underset{(\beta_0,\boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{RJ+M}}{\operatorname{argmin}} (C(\beta_0, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda))$
- Construct $\mathbf{Z}^K$ according to eqs. (34) and (35)
- Construct $\mathbf{Q}^K$ by adapting eq. (26)
- $(\beta_0^{K(q)}, \boldsymbol{\beta}^{K(q)}) \longleftarrow \underset{(\beta_0,\boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{LRK+M}}{\operatorname{argmin}} (C(\beta_0, (\mathbf{Q}^K)^{-1}\boldsymbol{\beta}, \mathbf{Z}^K\mathbf{Q}^K, \mathbf{y}, \lambda))$
- $q \leftarrow q + 1$

until $|C^K - C^J| < \epsilon |C^J|$ (where C^K and C^J are respectively $C(\beta_0^{J(q)}, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}^{J(q)}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda))$ and $C(\beta_0^{K(q)}, (\mathbf{Q}^K)^{-1}\boldsymbol{\beta}^{K(q)}, \mathbf{Z}^K\mathbf{Q}^K, \mathbf{y}, \lambda))$

Return $(\beta_0^{K(q)}, \boldsymbol{\beta}^{K(q)}, \boldsymbol{\beta}^{J(q)})$

Notes:

- The worst-case complexity of the tensor algorithms presented here is $O(J+K+R)$, compared with $O(JK)$ for non-tensor algorithms. So, if J and K are large and supposing $R \ll \min(J, K)$ then tensor algorithms are more efficient, as shown in [2].
- With the MMLR, we can deal with the case where each block is a tensor of different order. All we need to do is optimize several times according to the same $\boldsymbol{\beta}$ mode in blocks with fewer modes than the others.

- We decided to optimize the loss function completely in one direction before turning to the other one instead of alternating one step in each direction because the first procedure was more stable and could be implemented efficiently using the glmnet package in R [5].

2.3. Simulated data generation

To test the MMLR, we perform tests on simulated data. In this section, we explain how we generate this data.

2.3.1. Regression parameter structure

In order, to mimic the structure of real multiblock tensor data, we have organized our simulated data into several tensors (i.e. blocks). This enables us to compare the performance of the MMLR with other logistic models.

The multiway and multiblock aspect of the MMLR is reflected in its regression parameter $\boldsymbol{\beta}$. This is why we have chosen to generate our data in such a way that the optimal regression parameter $\boldsymbol{\beta}^*$, which minimizes the classification error, has a multiblock multiway structure. To make the reconstruction of the regression parameter as visual as possible, we reused the pictograms presented in [6]. A modified version of these pictograms is presented in Fig. 2 and Fig. 3. Thus, $\boldsymbol{\beta}^*$ is in fact composed exclusively of 0 and 1. The 1 are arranged to form simple geometric patterns when the beta vector is split into several lines (Fig: 2). The result is $\boldsymbol{\beta}^*$ in the form of a second-order tensor, each column of which is associated with a different explanatory variable and each row with a different observation modality. As pictograms are simple, the rank of the tensor is expected to be low in relation to the number of variables and modalities.

To add a multiblock aspect to $\boldsymbol{\beta}^*$, instead of choosing just one pictogram, we consider several pictograms: each pictogram defining one block of $\boldsymbol{\beta}^*$ (Fig: 3). Thus, each pictogram, seen as a 2nd-order tensor, is of low rank, but the concatenation of several pictograms produces a tensor of higher rank. It is this concatenation which, after being unfolded into a single line, constitutes $\boldsymbol{\beta}^*$. This renders the single multiway logistic regression model less relevant (which will need to have a high rank to correctly reconstruct $\boldsymbol{\beta}^*$), without putting the MMLR at a disadvantage (which separates $\boldsymbol{\beta}^*$ into several tensors of lower rank: one per pictogram).

2.3.2. Generation of explanatory variables

The method generally used to simulate explanatory variables in regression models is to use a simple probability distribution (often the standardized normal distribution), identical for all individuals. The explained variable is then obtained by applying the regression model with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ to the explanatory variables. This is, for example, what is proposed in [6]. However, this method is unsuitable for binary classification, as it fails to control both the number of individuals in each class and the optimal error rate.

To overcome this difficulty, we decided to generate the explanatory variables differently.

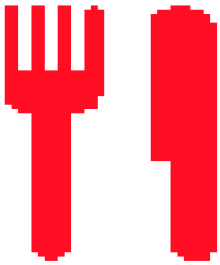


Fig. 2: Example of the pictogram used to generate $\boldsymbol{\beta}$.

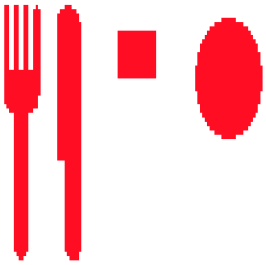


Fig. 3: Example of pictogram concatenation used to generate $\boldsymbol{\beta}$.

For each individual class, we chose to generate the explanatory variables according to a multivariate normal distribution. The two classes have the same covariance matrix, but different means. These means and covariance matrices are chosen to ensure that $\boldsymbol{\beta}^*$ is indeed the normal vector to the best class-separation hyperplane. To prove this, we will demonstrate that the method used ensures that this hyperplane is the Bayes classifier minimizing the classification error for the simulated data.

Proposition 1. *Noting respectively $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ the mean vectors of the J explanatory variables of the two classes and $\boldsymbol{\Sigma}$ the covariance matrix of these same variables, if we impose*

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \text{ is colinear with } \boldsymbol{\beta}^*. \tag{36}$$

$$\boldsymbol{\Sigma} = \mathbf{P}\mathbf{D}\mathbf{P}^T \quad \text{with } \mathbf{P} \in \mathcal{O}(n). \tag{37}$$

$$\text{The first column of } \mathbf{P} \text{ is colinear to } \boldsymbol{\beta}^*. \tag{38}$$

then the Bayes estimator (which, by definition, is the best estimator for minimizing the probability of classification error) has for decision frontier a hyperplane with normal vector $\boldsymbol{\beta}^*$.

Proof.
In a binary classification, the Bayes estimator is:

$$g^* : \begin{cases} \mathbb{R}^n \longrightarrow \{0,1\} \\ \mathbf{x} \longmapsto \begin{cases} 1 & \text{if } E[Y|X=\mathbf{x}] \geq 0.5 \\ 0 & \text{else} \end{cases} \end{cases} \tag{39}$$

Given that X and Y admit densities with respect to the lebesgue measure and the counting measure respectively, we have:

$$E(Y|X=\mathbf{x}) = \frac{1}{f_X(\mathbf{x})} \int y f_{(X,Y)}(\mathbf{x}, y) dy \tag{40}$$

Since Y admits a density with respect to the counting measure, this integral can be rewritten:

$$E(Y|X = \mathbf{x}) = \frac{1}{f_X(\mathbf{x})} \sum_{y \in \{0,1\}} y f_{(X,Y)}(\mathbf{x}, y) \tag{41}$$

And therefore

$$E(Y|X = \mathbf{x}) = \frac{f_{(X,Y)}(\mathbf{x}, y = 1)}{f_X(\mathbf{x})} \tag{42}$$

Which means

$$E(Y|X = \mathbf{x}) = \frac{f_{(X|Y)}(\mathbf{x}|y = 1)P(Y = 1)}{f_{X|Y}(\mathbf{x}|y = 1)P(Y = 1) + f_{X|Y}(\mathbf{x}|y = 0)P(Y = 0)} \tag{43}$$

Now, by hypothesis, we know that for $y \in \{0, 1\}$, $f_{X|Y}(\,.\,|y)$ is the density of $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Also, $P(Y = 1)$ and $P(Y = 0)$ correspond exactly to the proportion of individuals generated in each class and are therefore known. For the sake of simplicity, let's note: $P(Y = 1) = p_1$ and $P(Y = 0) = p_0$. Consequently

$$E(Y|X = \mathbf{x}) \geq \frac{1}{2} \tag{44}$$

$$\iff \frac{p_1 \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_1)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{2}\right)}{p_1 \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_1)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{2}\right) + p_0 \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_0)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}{2}\right)} \geq \frac{1}{2} \tag{45}$$

$$\iff \frac{1}{1 + \frac{p_0}{p_1} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_0)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}{2} + \frac{(\mathbf{x}-\boldsymbol{\mu}_1)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{2}\right)} \geq \frac{1}{2} \tag{46}$$

$$\iff \frac{(\mathbf{x} - \boldsymbol{\mu}_0)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)}{2} - \frac{(\mathbf{x} - \boldsymbol{\mu}_1)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)}{2} \geq \log\left(\frac{p_0}{p_1}\right) \tag{47}$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive symmetric, we can associate it with the positive semidefinite bilinear

form it induces, which we denote $\langle .,. \rangle_{\Sigma^{-1}}$. Thus:

$$E(Y|X = \mathbf{x}) \geq \frac{1}{2} \tag{48}$$

$$\iff \langle \mathbf{x} - \boldsymbol{\mu}_0, \mathbf{x} - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} + \langle -\mathbf{x} + \boldsymbol{\mu}_1, \mathbf{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} \geq 2 \log\left(\frac{p_0}{p_1}\right) \tag{49}$$

$$\iff \langle \mathbf{x} - \boldsymbol{\mu}_0, \mathbf{x} - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} + \langle -\mathbf{x} + \boldsymbol{\mu}_1, \mathbf{x} - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} + \langle -\mathbf{x} + \boldsymbol{\mu}_1, \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \rangle_{\Sigma^{-1}} \geq 2 \log\left(\frac{p_0}{p_1}\right) \tag{50}$$

$$\iff \langle \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \mathbf{x} - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} - \langle \boldsymbol{\mu}_1 - \mathbf{x}, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} \geq 2 \log\left(\frac{p_0}{p_1}\right) \tag{51}$$

$$\iff \langle 2\mathbf{x} - \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} \geq 2 \log\left(\frac{p_0}{p_1}\right) \tag{52}$$

$$\iff \mathbf{x}^T \mathbf{P} \mathbf{D}^{-1} \mathbf{P}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \geq \log\left(\frac{p_0}{p_1}\right) + \frac{1}{2} \langle \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} \tag{53}$$

$$\tag{54}$$

By hypothesis, the first column of $\mathbf{P}$ is collinear with $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. We denote $\mathbf{v}$ this column and λ the real such that $\mathbf{v} = \lambda(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Since $\mathbf{P}$ is orthogonal, all its other columns are orthogonal to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. We therefore have, noting d_1 the first real of the diagonal of $\mathbf{D}$:

$$\mathbf{x}^T \mathbf{P} \mathbf{D}^{-1} \mathbf{P}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = (\mathbf{x}^T \mathbf{v} \;\; 0 \;\; 0 \;\; \cdots \;\; 0) \mathbf{D}^{-1} \begin{pmatrix} \mathbf{v}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{55}$$

$$= \lambda^2 \mathbf{x}^T (\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) \, d_1^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{56}$$

And therefore

$$E(Y|X = \mathbf{x}) \geq \frac{1}{2} \tag{57}$$

$$\iff \mathbf{x}^T (\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) \geq \frac{d_1}{\lambda^2 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2} \log\left(\frac{p_0}{p_1}\right) + \frac{d_1}{2\lambda^2 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|^2} \langle \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \rangle_{\Sigma^{-1}} \tag{58}$$

$$\tag{59}$$

By hypothesis, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is colinear to β^*. Since the term on the right is independent of $\mathbf{x}$, the decision frontier of the Bayes classifier is indeed a hyperplane with normal vector $\boldsymbol{\beta}^*$.

The hyperparameters used to generate the simulated variables are presented in appendix

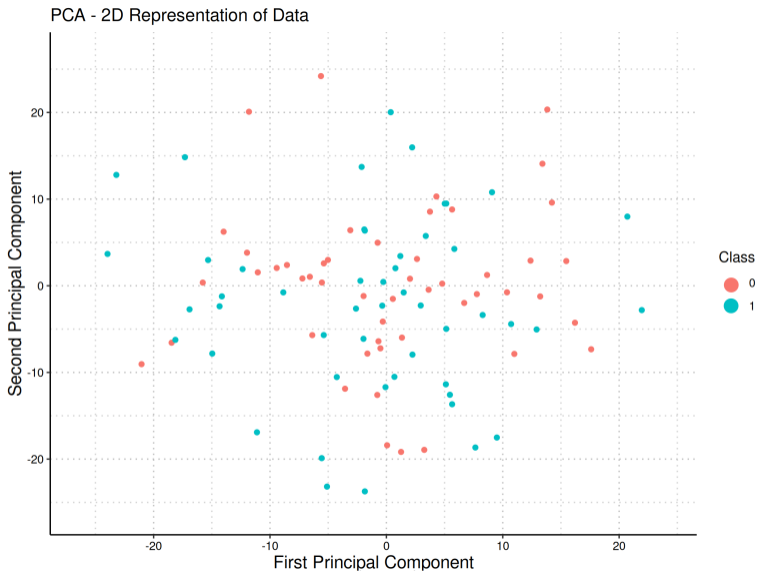


Fig. 4: Plane projection of the first two principal components of the explanatory variables simulated for 100 individuals when $\boldsymbol{\beta}^*$ is given by the concatenation of pictograms in Fig: 3

Appendix A. These are chosen experimentally to enable our models to reconstruct $\boldsymbol{\beta}^*$ without a simple 2-means algorithm being able to separate them (see appendix Appendix A). The projection onto the plane of the first two principal components of the explanatory variables is shown in figure 4. It shows that the classes are difficult to separate with the naked eye.

Table 1: Number of patients with usable MRI at the times indicated in the column for each tumor class. The total number of patients with each tumor class is entered in the total column.

class	Arterial	Portal	Venous	Late	All times	all times except venous	total
HCC	84	81	83	78	72	74	86
CCK	18	18	14	18	12	16	19
Mixed	35	36	32	34	29	31	37

3. Real dataset

3.1. Presentation of real data

The actual data on which we are working comes from a cohort of 142 patients with liver tumors. 86 of them have HCC tumors, 19 have CCK tumors and 37 have mixed tumors. Mixed tumors display both CCK and HCC characteristics. However, they are still poorly understood, and some doctors even prefer to categorize them as HCC or CCK depending on which aspect predominates in the tumor. Therefore, they are not studied in this article.

Each patient underwent four MRI radiographs of the liver, one at each time point after contrast injection. These were arterial, portal, venous and late. However, not all MRIs are usable. The patient may move during the MRI, rendering it unusable. A summary table (Table 1) is provided in order to specify the number of usable MRIs by temporality. Clinical data are also available: age at tumor detection, gender and patient alpha fetoprotein (AFP) levels. However, since the AFP levels of 22% patients are missing from the data, we decided to exclude this clinical variable.

On each of the MRI, the tumor area is displayed and saved as a mask superimposed on the MRI. The MRIs and masks are in .nii format. Although taken at four different times, the four MRIs are very similar. In particular, the MRIs at venous and late time are extremely similar and often redundant in the eyes of radiologists. We will take this opportunity to eliminate the venous time MRIs, as this is the time for which there are the most missing MRIs. We propose two possible extractions for features. A 3D extraction, where features are extracted from the entire tumor volume, and a 2D extraction, where features are extracted from each tumor section.

3.2. feature extraction in 3D

We use the pyradiomics package [7] to extract an array of 3D features for each tumor. Only the original (unfiltered) image is used to extract these features. We extract all the first-order parameters (relative to gray levels), 3D shape parameters (volume, surface, etc.), and texture parameters (based on co-occurrence matrix, gradient matrix, etc.) proposed by the package (except those considered deprecated or duplicative: for example, we eliminate glcm joint average as it is redundant with glcm sum average). The result is 106 features for each radio. Shape parameters are averaged over all extracted temporalities, as we consider that the shape of a tumor has no reason to change between different MRIs.

The exact parameters used for pyradiomics extraction are given in appendix Appendix B. To ensure that the extraction is consistent from one tumor to the next, all tumors have been

resampled to the same scale. On each (x, y, z) axis, the spacing used is half the median spacing on that axis (calculated over all available MRIs). The idea behind this spacing is to avoid losing too much information by increasing the voxel size of higher-resolution MRIs without having to completely interpolate lower-resolution MRIs. Image interpolations are performed using cubic splines, while mask interpolations are based on the closest interpolation method (to guarantee mask connectivity).

The advantage of 3D extraction is that each MRI image is summarized in a relatively small number of features (compared with 2D extraction). What's more, since the parameters are calculated on the tumor in its entirety, they do not omit any part of it. This idea is confirmed by the absence of improvement in the performance of the logistic lasso regression when features from the 2D extraction are added to the 3D parameters: the 3D features seem to stand on their own. The weakness of the 3D extraction is that it requires a complete segmentation by the radiologist of every tumor in the training database, which is very time-consuming.

3.3. Feature extraction in 2D

The first step in this extraction process is to determine the slices we wish to extract from the tumor. We choose the axial plane for the slices, as this is the one used by radiologists when analyzing a tumor. As for the extraction parameters, they are again given in Appendix Appendix B. However, we cannot simply extract slices at regular intervals along the vertical axis, for two reasons:

Firstly, tumor size varies from patient to patient. Thus, a certain spacing between slices will lead to the extraction of 3 slices of tumors in some patients and 10 slices in others. However, the machine learning models we use need to compare the same features in all patients. Secondly, slices with a very small piece of tumor are not very significant for our analysis. However, extracting at regular intervals will lead to the extraction of such slices at the beginning and end of certain elongated tumors (along vertical axis). We'd therefore like to give more importance to slices where the area of the tumor region is larger (without completely ignoring slices with less tumor on them).

We therefore propose an extraction where we first specify the number of slices n_{slices} to be extracted from each tumor. We begin by interpolating the cumulative distribution of tumor volume by depth (along the vertical axis) for each tumor (see Fig. 5). This curve is then inverted to obtain the depth distribution as a function of the cumulative tumor volume covered. A slice is then extracted at each of the following depths:

$$(i - 0.5)\frac{\text{area}_{\max}}{n_{\text{slices}}} \qquad \text{for i } \in [\![1, n_{\text{slices}}]\!] \tag{60}$$

We have experimented other extraction methods, in particular trying to extract precisely the same depths for each MRI of the same tumor, while taking into account the fact that the patient may have moved slightly between two MRIs. However, as the results were of lesser quality, we will not develop these approaches here.

The feature extraction used for each slice is almost the same as the one used for the 3D

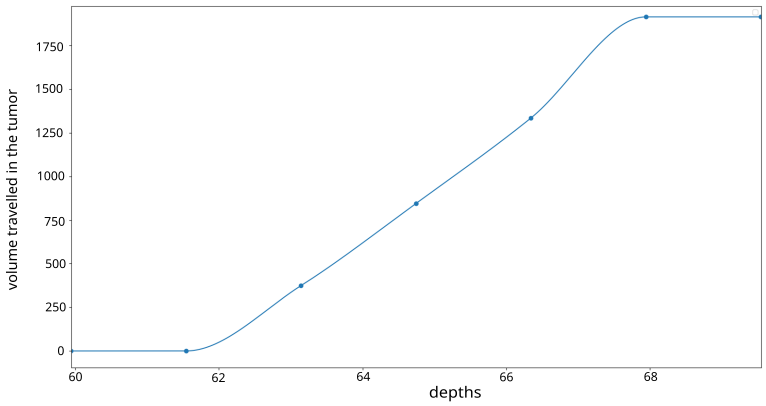


Fig. 5: graph of the cumulative volume distribution (in mm^3) of the third CCK patient's tumor according to depth (in mm) for a given tumor. The points correspond to the slices recorded in the sitk image (with its initial spacing). The curve is obtained by interpolating these points using Hermite cubic splines.

tumor (in the previous section), except for shape parameters. In fact, 2D shape parameters (instead of 3D shape parameters) are now extracted by pyradiomics. 2D shape parameters are always averaged over all MRIs of the same tumor (variations in tumor shape between MRIs result solely from changes in the way radiologists cut masks, and therefore do not provide information on the tumor itself).

The advantage of this type of extraction is that the radiologist may only need to segment a limited number of slices that are "representative" of the tumour, instead of the whole tumor. Indeed, the work carried out to find the right slices to extract could be replaced by an estimate made by the radiologist's naked eye. Although it would be necessary to check that the results are not affected by this change, this method seems simpler to generalize to large datasets (the segmentation of a couple of slices being less time consuming for a radiologist than the segmentation of the whole tumor). The disadvantage of this extraction method is that it loses the information contained in the slices that were not selected, as well as that concerning the overall shape of the tumor (which cannot be reduced to the shape observed on a few slices). This results in a slightly poorer performance of all models on these data (cf section 4).

3.4. Extraction of healthy liver parts

We wanted to add the features obtained by performing the extraction on portions of healthy liver. Radiologists generally compare the luminosity of the tumor area with the rest of the liver, so it seemed appropriate to do the same with our model.

To do this, a small strip of tissue was extracted around the tumor area. To ensure that no area outside the liver or crossed by a blood vessel was included, we decided to extract only areas of low local variance and whose luminosity was greater than that of the black background. By adding a 3D connectivity criterion, we can extract a 3D area of healthy liver large enough to perform a 3D extraction of firstorder and texture features (the shape

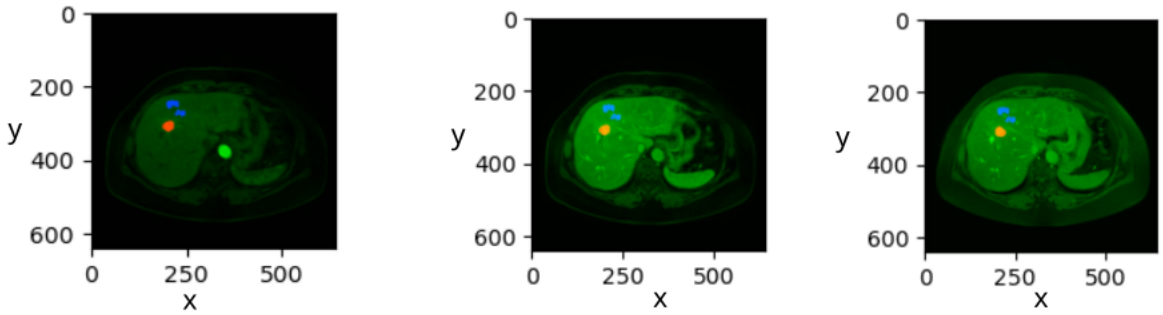


Fig. 6: MRIs of a CCK tumor slice with the tumor area in red and the peripheral area of extracted healthy liver in blue. From left to right, arterial, portal and late MRIs. Axes are graduated in mm.

of the extracted area being of no interest).

To ensure that the same area of healthy tissue was extracted from each MRI of the same tumor, we decided to crop the healthy tissue on the late MRI only (this was when our extraction method was most visually successful). We then applied the same trimming to the other MRIs, shifting the extracted area slightly to take account of the patient's movements. These movements were estimated by comparing the tumor areas on each slice and trying to increase as much as possible the intercorrelation of the area curves between each MRI and the late MRI. This procedure is performed along all three spatial axes. Finally, the extracted zone can be visualized, as in Fig. 6 to check that the extraction is proceeding correctly.

However, we did not perceive any improvement in the performance of our models by adding these features. We therefore decided not to include them in the rest of our study.

Table 2: Area under curve for each model on simulated data. For the group lasso model, it is possible to group variables by block, mode or variable. The type of grouping used is indicated in brackets

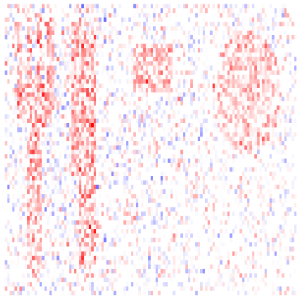
number of individuals	lasso	goup lasso (by block)	goup lasso (by mode)	group lasso (by variable)	multiway	multiway multibloc
3000	0.83	0.86	0.94	0.94	0.99	0.99
500	0.59	0.65	0.64	0.63	0.92	0.63

4. results

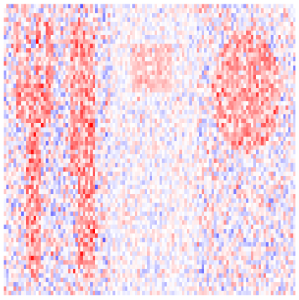
In this section, we compare the performance of the tensor models presented in 2.2 with non-tensorial approaches, namely lasso and group lasso logistic regressions. These alternative methods operate directly on the matricized tensor $\mathbf{X}_{(1)}$. To approximate the structure-awareness of tensor models, the group lasso is applied with feature grouping based on modes or variable names. This comparison allows to evaluate the potential performance benefits of explicitly preserving the tensor structure in the model.

4.1. Simulated data

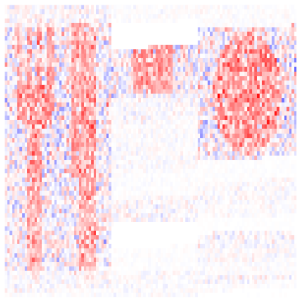
We performed tests on simulated data generated with the parameters presented in Appendix A. The pictogram to retrieve is the one in figure 3. We evaluated the performance in two settings: one with a lot of individuals to classify (3000 individuals in the training data set) and another with fewer individuals to classify (500 individuals in the training dataset). The testing dataset is always composed of 1000 individuals. We consider the area under curve to determine the performance of each model as it is a robust and fine grained indicator of the efficience of each model to separate the two classes. We show the results in Table 2. For each model, the hyperparameters used during cross-validation are provided in Appendix A. In each case, we present the pictogram reconstructed here:



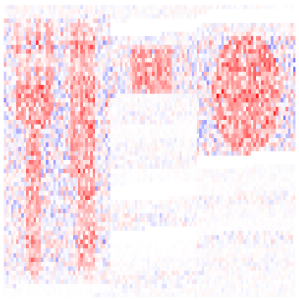
(a) lasso



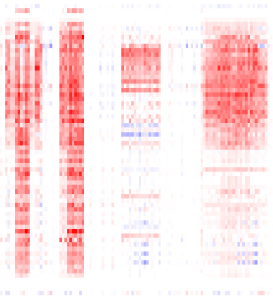
(b) group lasso (block)



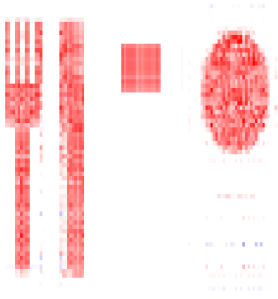
(c) group lasso (variable)



(d) group lasso (mode)

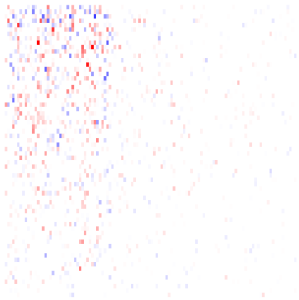


(e) multiway

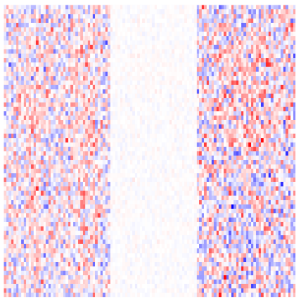


(f) multiway multibloc

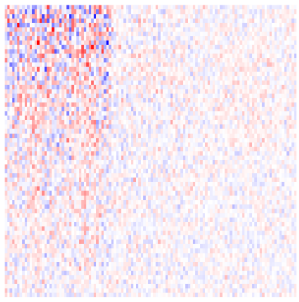
Fig. 7: Pictograms reconstructed by the different models for 3000 individuals in the training dataset. The name of the model used is indicated in the legend of each figure.



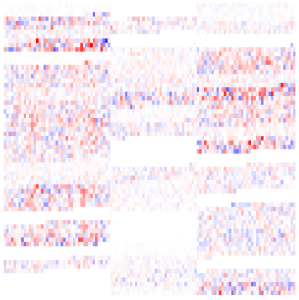
(a) lasso



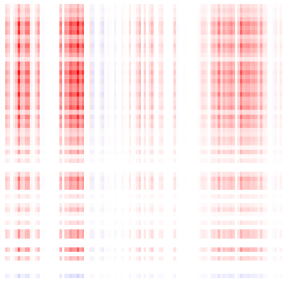
(b) group lasso (block)



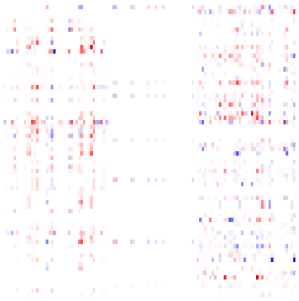
(c) group lasso (variable)



(d) group lasso (mode)



(e) multiway



(f) multiway multibloc

Fig. 8: Pictograms reconstructed by the different models for 500 individuals in the training dataset. The name of the model used is indicated in the legend of each figure.

We can see that, in the presence of a large number of individuals (3,000), the MMLR is the most successful at reconstructing pictograms. In terms of performance, measured by the area under curve, its performance in this setting is similar to that of the multiway model. The other models (group lasso and classical logistic regression with lasso) perform much less well than the two tensor models. In the presence of a small number of individuals (500), the multiway model performed best. In fact, it captures the pictogram structure in a very small number of parameters (the rank chosen by cross validation is equal to 1 in this configuration). Even imposing a rank of 1 on the MMLR requires the calculation of more coefficients (around 2 times more in the case of our pictograms), which can lead to greater over-interpretation.

These simulations clearly demonstrate the usefulness of tensor models for classifying high-dimensional data. They show that the structure of the $\boldsymbol{\beta}$ coefficient is more easily found by the MMLR (when $\boldsymbol{\beta}$ has a block structure, as is the case for our pictograms). However, results also indicate that the multiway model performs better than the MMLR in classification tasks when working with small datasets.

Table 3: Average area under curve obtained with each model on real data for 50 different trainings (with different partition between training set and testing set). For the group lasso model, it is possible to group variables by block, mode or variable. The type of grouping used is indicated in brackets. The confidence intervals provided are normal-based at a 95% confidence level.

Type of data	lasso	goup lasso (by block)	goup lasso (by time)	group lasso (by variable)	multiway	multiway multibloc
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	goup lasso (by block)	goup lasso (by slice)	group lasso (by time)	group lasso (by variable)	multiway	multiway multibloc
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

4.2. real data

We carried out tests on simulated data both using the 3D extraction (as described in 3.2) and the 2D extraction (as described in 3.3). Results are presented for each model in table 3. All results are averaged over 50 different training sessions. An analysis of the importance of each feature of the data studied is also proposed in Appendix C

The performance obtained on medical data are not good enough for our models to be used in real conditions. In particular, in the course of our tests, we found that no model achieved an accuracy of better than 50% in the detection of HCC tumors. Overall, 3D data give better results than 2D data. This may be explained by the fact that 2D data do not take into account the entire tumor. Experiments combining 2D and 3D parameters have been carried out, but have never exceeded the performance of 3D data alone.

With both extraction procedures, the MMLR performs well compared to other models, but never better than all the studied non-tensorial models. This indicates that the structure of the optimal $\boldsymbol{\beta}$ parameter is likely not tensorial. Indeed, a simple grouping of features (enabled by the group lasso), is sufficient to obtain similar results. In terms of computation time, the MMLR is the most time-consuming, as it requires cross-validation on the rank to be used. In particular, as the number of times and slices in the data is close to 1, there is no gain in computation time compared with non-tensor models.

5. Conclusion

In this article, we proposed the MMLR model, a novel multiway multiblock logistic regression method, along with an efficient algorithm for its training. The results obtained on simulated data demonstrate the effectiveness of tensor-based methods, particularly when the regression coefficient $\boldsymbol{\beta}^*$ exhibits a tensor structure. The MMLR model is especially well-suited for efficiently capturing this structure when data are organized into blocks. Moreover, the

classic multiway tensor model, which represents the state of the art, achieves the best overall classification performance on simulated data, and our MMLR model comes close while offering added flexibility for multiblock scenarios.

On the liver cancer dataset, tensor methods do not show a clear advantage over non-tensorial approaches. This observation suggests that, while the features are well represented in tensor form, the regression coefficient $\boldsymbol{\beta}^*$ might not adhere to this structure. Additionally, the limited size of the training dataset - only 16 CCK tumors with complete MRI data - likely hinders the potential of the models to achieve strong generalization.

Future work could aim to improve model performance by extending the feature set. For example, incorporating features defined by radiologists based on clinical expertise, in addition to pyradiomics descriptors, would likely enhance the models' ability to capture relevant patterns in the data. Another promising direction involves extending the MMLR model by incorporating a group lasso penalty to exploit potential group structures within blocks.

Finally, to facilitate reproducibility and further exploration, the code used for simulations and experiments in this study is publicly available at `https://github.com/AlexandreSelvestre/Stage-3A.git` .

6. Latest results: not mentioned in the article

As indicated in the summary of this internship report, two weeks ago I went to visit Sébastien Mulé at Henri Mondor Hospital to talk to him about the pre-processing of liver cancer data. It was then that he remembered the existence of another database concerning the liver cancers of the patients studied in the article. In fact, for each individual, the radiologist had also indicated the presence or absence on the MRI images of 13 markers (presence of necrosis, presence of luminal enhancement in the late phase, etc.) which are used by radiologists to determine the class (HCC or CCK) of the tumor. This translates into 13 binary variables in the database, to which we add the patient's gender. Unlike the features extracted by pyradiomics, the features in this database are based exclusively on medical criteria. Furthermore, each marker indicated by the radiologist takes into account all 4 MRI images, whereas features extracted by pyradiomics only took into account one image at a time and were therefore extracted image by image. This explains why, unlike the features extracted by pyradiomics, those in the new database are not tensorial.

In order to compare these data with those studied in the article, a lasso logistic regression was performed on these data. Averaging the area under curve over 50 trainings, we find:

$$\text{AUC} = 0.96 \pm 0.02 \qquad \text{balanced accuracy} = 0.85 \pm 0.05$$

where confidence intervals are of the normal type and calculated at the 95% threshold. We can see that the two most important features in the classification (where feature importances are calculated as in Appendix C) are, in descending order, late luminal enhancement and non-peripheral washout. The first of these two features is considered the most important by radiologists (when they try to determine with naked eye the class of a tumor), which is consistent with our model. These results are much better than those obtained with data extracted by pyradiomics. The quality of these results is particularly high given the low number of patients in the training database. Moreover, lasso logistic regression may not be the best model on these data, and we can hope to improve these results still further by using models more suited to binary data, such as random forest. These results are left out of the article section for the following reasons:

- They have no connection with tensors, whereas tensors are at the heart of the article. In order to practice writing reports in article format, it was therefore decided that I should write the section on actual data as if the latest data had not been communicated to me.

- As these results were obtained late in the course, we did not have much time to develop them.

I am well aware, however, that the latest results completely exceed those obtained in the article, and that, with a view to publication, it would be necessary to study another set of real data, more suited to tensor models.

7. Retrospective and Perspectives on the Internship

7.1. Possible extensions of the work

The work carried out on tensor models during this internship shows that these models can be genuinely useful when the regression parameter has a tensor structure, as in the simulated data. In particular, it demonstrates the effectiveness of the MMLR to find the regression coefficient $\boldsymbol{\beta}$. However, this model could be further generalized. We could change the penalty used to, for example, mix the L1 and L2 penalties, or mix the group lasso with tensor models. We could also try to adapt variable selection procedures such as those proposed in [8], to improve computation speed in very high-dimensional tensors.

With regard to the liver cancer data, it would be interesting to train several tabular machine learning models on the latest data (random forest, group lasso, boosting etc...) in order to propose the most accurate model possible and compare it with the methods already used by radiologists. Indeed, while the performance of the algorithms on the data presented in the article section of this report left no doubt as to their inferiority to that of radiologists, on the new data, this becomes more uncertain. Finally, it would also be good to add mixed tumors to the analysis, to see whether the models are also capable of distinguishing them. In practice, they make up around a fifth of tumours, and even if they are less well understood than HCC or CCK tumours, it remains interesting to be able to distinguish them.

7.2. Assessment and Reflection

This internship enabled me to acquire the technical skills needed to pursue a thesis under the right conditions in the same laboratory. Indeed, I was able to familiarize myself with complex classical machine learning models, with the R language, with tensor data and with the laboratory's calculator (and its formalism). As the data I will be working on in my thesis (the results of blood samples analysis taken from lung cancer patients) are of a similar nature to those studied in this internship, I will be able to reuse most of these tools. What's more, I have been able to see how the laboratory is organized and I am used to working with my supervisors. In particular, I think I have found the right balance between taking the initiative and asking my supervisors for help.

I was also able to see how the distance between the various players could complicate the process of obtaining results. Indeed, I did not wait for the appointment at the hospital a few weeks before the end of the internship to meet Sébastien Mulé. On the contrary, he came to the laboratory regularly (around once every six weeks) to discuss research progress. But at no time did we think to ask him if data other than that which he had provided were available. And it was only at the hospital that he thought to mention the existence of such data. I think there are several reasons for this situation:

- On the one hand, our distance from the hospital meant that our work was necessarily compartmentalized from that of the radiologists. This compartmentalization was accentuated by the fact that we had only one medical intermediary: Sébastien Mulé. As a result, the slightest omission on his part could have major consequences for our work (since we could not cross-reference our information with that of other radiologists).

- Furthermore, as my research placement was not part of a larger project, there was no dedicated project manager. My supervisors were the closest to this role, but they did not have the freedom to go out to the hospital regularly to talk to the doctors, as a project manager might have done. This had consequences for the flow of information.

For my thesis, I will be keeping a close eye on these points, to make sure I do not miss out on any key information. So, with my supervisors, I will make sure I regularly meet as many stakeholders as possible (doctors, other PhD students working on the same data but from a different angle, their supervisors etc...), preferably at their place of work, to make sure I do not make any mistakes or forget anything. The question of the project manager should be less of an issue as the thesis I am about to pursue is part of a project with the Gustave Roussy foundation. As such, there is already a project manager in charge of coordinating the various research projects. My supervisors have already given me the opportunity to learn more about the project by inviting me to a progress meeting with all the stakeholders on Thursday September 26, 2024.

Apart from the complexity of managing the risks associated with communicating information between the laboratory and the hospital, the management of other risks proved satisfactory. In fact, I regularly backed up my work on a GitHub server to ensure that I would not risk losing everything in the event of a problem with my personal computer. I also had an external hard disk on which I saved my work (albeit at less regular intervals), to avoid any major data loss in the event of a GitHub handling error. In addition, any demonstration written down on paper had to be quickly written up in LaTeX (a requirement of my supervisors), to avoid, for example any loss of time in the event of the research notebook being lost. As far as the risks associated with managing deadlines were concerned, we were also quite successful in planning the use of the calculator well in advance (more than three months before the end of the course). This turned out to be essential, because between the cyber-attack suffered by the University of Paris-Saclay, the calculator taking longer to get used to than expected (requiring lengthy discussions with the technical department) and maintenance lasting two weeks instead of one, I might never have had time to carry out my simulations if we had started later.

The work carried out as part of this internship is useful for both the laboratory and the hospital. Trainings with simulated data on the MMLR are interesting, both for assessing the model's performance and for reflecting on data generation. For the hospital, the results obtained on the new real data set are promising, and encourage further research in this direction. Finally, although the work carried out on pyradiomics did not produce satisfactory results, it did enable us to exclude this extraction method from the scope of our studies. Indeed, the features obtained by pyradiomics had already been the subject of several research activities/projects with students at Centrale-Supélec or Henri Mondor hospital (master's thesis at Henri Mondor hospital, integration teaching at Centrale-Supélec etc.) and it was therefore important to discover their limits. This internship report condenses all the important information obtained during the internship, and will serve as a reference for the laboratory in the years to come, should any questions arise about the work that has been produced. In order to be as precise and concise as possible (and therefore as useful

as possible for the laboratory), the research content has not been simplified in the article section.

On a personal level, I am convinced of the merits of the research I have carried out. Machine learning is indeed the subject of much debate, and I would not have wanted to get involved in a project whose consequences could have been harmful to society. For example, the research I carried out was primarily aimed at improving a precise diagnosis, and not at replacing the radiologist (whose role goes far beyond this task). What's more, classic machine learning models such as those studied do not have the opacity of deep learning models, and remain highly specific. Nor do they have the invasive and potentially uncontrollable character of more "general" intelligence models such as language models. Finally, from the patient's point of view, it is important to note that the data used in this course are anonymized and that patients have given their consent for their RMI images to be used for research purposes.

7.3. Competency framework

In accordance with Centrale-Supélec's requirements for internship reports, this section presents a self-assessment of the engineer's nine skills, based on the experiences I had during this internship..

7.3.1. Complexity

Statement: *Analyze a system in its entirety, identify its scientific, economic, human and other dimensions, model with the appropriate scale and assumptions, solve the problem, and design all or part of a complex system.*

Strengths: I have worked with complex problems during this research internship and I have been able to adapt to studying them. In particular, I know that there is never just one solution to these problems. In the case of machine learning, sometimes it is the model that is at fault, other times the raw data, other times the pre-processing of this data, sometimes even the hardware used etc. So I have learned to be patient and not to close off possibilities from the outset. In particular, I think I have also grasped the human aspect of these problems, and the importance of good cooperation between all the parties involved in finding workable solutions.

Areas for improvement: I have not often been confronted with problems with a strong economic component. Even if I integrate budget constraints into my projects (this will be the case, for example, for the lung cancer monitoring tool we will be developing during my thesis), I have never had to make decisions that would have major economic consequences. I still have a lot to learn in this respect. On the other hand, as I am at the very beginning of my professional career, I have to remain humble about my knowledge of complex problems: I believe that experience is also a key element in solving them.

7.3.2. Engineering Profession

Statement: *In-depth knowledge of a particular discipline or sector should enable graduates to delve deeply into a field, to think in depth, to understand the difficulties and subtleties of a subject. This "expert" approach enables graduates to acquire complex ways of thinking that can then be transposed to other sectors, giving them the distance they need to approach future learning effectively. 2 axes are targeted: the business sector, and the cross-functional business line.*

Strengths: During this internship and my third year at Centrale-Supélec, I was able to deepen my knowledge of machine learning. In doing so, I was also able to make progress in mathematics, modeling and computer science, and I got into the habit of reading scientific articles on my own on subjects related to my work. I have also been able to discover the health sector and health research, which was previously unknown to me. All this gives me a more global vision, both of my field of expertise (I can deal with more machine learning problems, I know more about the "tricks", the recurring difficulties etc.) but also of the field of engineering sciences in general (by being at the interface of several professions).

Areas for improvement: My versatility, linked to my engineering training, certainly enables me to adapt easily to a wide range of professional environments, but it has prevented me from delving as deeply into a specific field as students who have completed a full master's degree in it. For example, my specialization in machine learning only began in my final year at Centrale-Supélec, whereas other university students had studied the field for several years. I am particularly lacking in knowledge related to the mastery of low-level programming languages, web tools and concepts, containment, compilation and advanced use of Linux (although I am bound to make progress on these subjects as I continue in machine learning).

7.3.3. Innovate and Undertake

Statement: *Entrepreneurial competence is understood in a broad sense (self-entrepreneur, capacity for action, intrapreneur, business creator, etc.): observing and allowing oneself to criticize the world as it is, questioning one's initial hypotheses, proposing alternatives integrating risks and uncertainty, concretely implementing innovative ideas, industrialization to deliver tangible results.*

Strengths: My research work (during this internship and during my research course at Centrale-Supélec) has accustomed me to proposing innovative solutions. In particular, I know that I should not hesitate to question the initial hypotheses of a given problem and to propose alternatives (as was the case for the processing of real data during this internship). I also know that before trying to innovate, I have to be sure of what already exists and try above all to reduce the difficulty of the problem posed (especially when there is no need for an optimal solution). Finally, working with a hospital has taught me the importance of always keeping the customer's final objective in mind (to avoid any drift).

Areas for improvement: I have never been in an entrepreneurial or managerial context, so I have not often had to make decisions that impact an entire business. Nor have I had to manage teams of workers. So the human aspect of innovation contexts is still rather theoretical for me.

7.3.4. Value Creation

Statement: *Starting from a company's definition of increased productivity and profitability, Value Creation is extended here to the contribution of progress and improvement in areas such as the efficiency of a process, the performance or reliability of a technical solution, or the environmental or societal impact; in such a way that the beneficiary (or beneficiaries), whether individual or collective (companies, organizations, society), can see and measure the effects.*

Strengths: I have already carried out group projects at Centrale-Supélec in response to a client's request (in the Major Project Management program, for example), and this internship was an opportunity to discover a scientific context where the major issue is added value for the client (the hospital). I know that it is important to clearly define the customer's expectations by discussing them with him beforehand. In particular, I have noticed that the customer is sometimes unaware of the extent of the possibilities available to him, and so it is important to be proactive during the exchange (this was the case here, for example, when we decided to study portions of the healthy zone in the tumor: Sébastien Mulé was not sure that we had the means to properly cut out these healthy zones ourselves). Defining evaluation metrics was also an important step at the start of the research internship, so I think I am pretty comfortable with this aspect.

Areas for improvement: I have never had to manage a major project, so I have not had to deal with the way in which these projects can gradually drift out of line with the customer's expectations. Furthermore, I have never had to deal with a situation where meeting the customer's expectations was linked to high financial stakes. I think I still need to gain experience in these areas.

7.3.5. Intercultural

Statement: *Evolve in multicultural and multilingual environments that expose you to ways of thinking, functioning and communicating that are different from your own, and step back from your own cultural filters to adapt, cognitively, behaviorally and emotionally, to intercultural environments.*

Strengths: This internship did not have any international component. In order to develop an answer on my current skills on this subject, I must call upon my international experience during the second semester of my gap year. It was a humanitarian internship in Greece, during which I had to adapt to the language barrier (I do not speak Greek) as well as to

ways of thinking that were sometimes different from my own (although having remained in the European Union, apart from a generally more assertive conservatism than in France, I did not perceive huge differences). During this internship, I was able to work with children of different nationalities and see how a humanitarian project was organized on several continents. Therefore I know what it means to work in a multicultural environment.

I am fluent in English (level C1) and German (level B2) and I have already had the opportunity to work with people of different nationalities in a scientific context at Centrale-Supélec.

Areas for improvement: I have never had to work in a country where the culture was completely different from my own (on another continent). What's more, my command of English, while good, is far from perfect, and I still have to rely heavily on automatic translators when I write.

7.3.6. Digital

Statement: *Digital technologies are developing at breakneck speed, and their adoption by individuals and businesses is transforming the economy and society, introducing more sharing, cooperation and empowerment. This movement is accelerating and leading companies to transform themselves, or disappear. This transformation is not only technological, but also organizational and cultural. Students are at ease in this digital world where they innovate and "disrupt". They understand the techniques and sciences that underpin the digital revolution.*

Strengths: Being specialized in machine learning, I am used to working with numerical methods, both for minimizing objective functions (as was the case during this internship) and for approximating densities. I also know that it is important to check their effectiveness in simple cases, by evaluating their margins of error and comparing them with existing methods wherever possible (as was the case when comparing the different regression models in this internship). Although it was not the focus of this internship, I know the basics of algorithm proofs and have taken specific courses on them in the context of optimization algorithms.

I am also aware of the constraints linked to the performance of machines and calculators, and I have learned to make the best use of available capacities (by parallelizing calculations, for example) or by using common strategies (recall tables, approximations, etc.) to save calculation time. I am aware of the importance of not using more resources than necessary for calculations (this is expressly requested when using the laboratory's calculator, for both environmental and financial reasons). Finally, during the many group projects I carried out at Centrale-Supélec, I got into the habit of using collaborative work tools such as git (a tool also used during this internship) and of using the correct version for each tool (very important during this internship for pyradiomics to work: it does not function with the latest versions of Python).

Areas for improvement I never did software development or web development, and I have

never taken a course on cybersecurity. I have already touched on these notions from afar during generic first and second year courses, but that is about it. However, I will probably end up having to learn more about these subjects if I want to continue in computer science. I have more knowledge of databases, but I have never needed to use or build one in SQL during my projects. Finally, I have never had to work with massive data, so I have not had to use big data tools like Hadoop or Spark (although I did take a course on them in my second year). Again, I think I will probably have to learn more about these subjects if I want to continue in machine learning.

7.3.7. Persuade

Statement: *Convincing means presenting a point of view - or a proposal - in such a way as to enable an interlocutor to recognize its accuracy, relevance or force, and thus to appropriate it.*

Strengths: Writing this internship report (article section) gave me the opportunity to test myself against the requirements of a written scientific presentation. In particular, the methodology section (2), which was written before the rest of the report, underwent several revisions to make it more precise, which gave me a good idea of the level of clarity expected in a publication. During this internship, I also often had to present my latest advances to my supervisors orally, sometimes without any written support, and I think I have made a lot of progress in this area. I am now better at justifying my choices and anticipating objections. Finally, this internship has forced me to adapt to my interlocutor, for example by simplifying mathematical explanations when talking to a doctor.

Areas for improvement: I need to take more into account readability in my reports. Indeed, I have often tended to write equations that are too heavy with too few comments in the Methodology section (2), which can put off an uninitiated reader (this is what required the most corrections in the article part of this report).

7.3.8. Project Team

Enoncé *Build, mobilize and train a collective to work as a team, demonstrate different forms of leadership, enrich the team with external resources and expertise, and work in project mode.*

Strengths: Since no other intern or PhD student was working on my research topic, the leadership and task distribution aspect was less present than in other internships. However, I quickly understood that it would be necessary for me to be proactive, by regularly making decisions myself and testing my proposals before submitting them to my supervisors. In addition, I also realized that I had to question my supervisors' proposals and not hesitate to ask them for details on the real expectations of each task. This was the case, for example, in the pre-processing part of the real data, where I was able to discover several points where the extraction of the real data was suboptimal (some parameters to be extracted were omitted, there were different scales for each RMI image of the same patient, etc.)

Areas for improvement: The communication with the radiologist could have been improved, which would have avoided us changing real data a few weeks before the end of the internship. In addition, even if my internship was not considered as a project, in the sense that the final goal was only to have the best possible results and not to develop a finished tool, an organization closer to project management could probably have helped. In particular, giving one of the two supervisors the full responsibility of project manager would have probably improved communication. However, I think that from a logistical point of view (the supervisors are already researchers themselves and also supervise doctoral students, the hospital is far from the laboratory, etc.), this was not possible.

7.3.9. Ethics and Sustainability

Statement *Analyze and anticipate the possible consequences of one's actions, of the decisions of organizations and economic models of the structures to which one contributes; arbitrate an ethical dilemma; act in an inclusive manner when faced with questions of diversity; respect scientific ethics.*

Ethical questions cannot be reduced to binary reasoning (strengths/areas for improvement) so this separation is abandoned here.

The main ethical dimension of this internship concerns the impact of the research carried out. Indeed, machine learning is a powerful tool whose use can be negative on our society. For example, we can think of the use of language models by foreign powers to write waves of negative comments under political articles and videos or the impact of the recommendation algorithms of Youtube, Facebook or Twitter, which tend to imprison users in filter bubbles (by only showing them what confirms their position) and amplify the sharing of hateful and misleading content compared to other content. In this internship, however, this type of use is not really possible. Indeed, most of the algorithms that are debated are based on deep learning (artificial neural networks), which is not the subject of this internship. Furthermore, the algorithms developed here are specific to a certain type of data and it therefore seems unlikely that their use will be diverted for malicious purposes. On the contrary, we can guess the immediate interest of the research carried out, namely the improvement of the diagnosis of liver cancer. In my opinion, therefore, the ethical impact of this internship is positive.

Furthermore, since the data is anonymized and was collected from patients who gave their consent for their RMI images to be used for research purposes, this data is collected ethically. In addition, as stated earlier in this report, there is no question of replacing the work of a human being with this type of model. And even by extending this type of work, it seems difficult to imagine arriving at models capable of replacing humans (these models learn a specific task and are not capable of generalizing to other tasks). Finally, regarding this report itself, I was careful to write it as accurately as possible, without modifying or exaggerating the results obtained, even when it did not go in the direction I would have initially hoped.

8. Acknowledgments

I warmly thank my two supervisors Arthur Tenenhaus and Laurent Lebrusquet for their kindness and availability throughout this internship. I also thank them for the motivation they helped give me in this research work, even when the results were not those expected. I thank them for accepting after this internship to welcome me in this laboratory where I feel so good since the beginning of this internship. Finally, I thank them for all the help they gave me during this internship, whether on the computer, mathematical or administrative level.

I thank Sébastien Mulé for the time he was willing to devote to us in addition to his working hours, despite his already very busy schedule, and for all the trips he agreed to make to the laboratory. I also thank him for the many explanations he gave us about liver cancer and for his offer to host me in his hospital for a day (which led to the discovery of the new database studied).

Finally, I would like to thank the technical support of the laboratory calculator, without which the simulations could not have been carried out.

References

[1] L. Le Brusquet, G. Lechuga, A. Tenenhaus, Régression Logistique Multivoie, in: JdS 2014, Rennes, France, 2014, p. 6 pages.
URL https://centralesupelec.hal.science/hal-01056558

[2] F. Girka, P. Chevaillier, A. Gloaguen, G. Gennari, G. Dehaene-Lambertz, L. Le Brusquet, A. 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.
URL https://centralesupelec.hal.science/hal-03051752

[3] L. Meier, S. Van De Geer, P. Bühlmann, The Group Lasso for Logistic Regression, Journal of the Royal Statistical Society Series B: Statistical Methodology 70 (1) (2008) 53–71. arXiv:https://academic.oup.com/jrsssb/article-pdf/70/1/53/49796502/jrsssb_70_1_53.pdf, doi:10.1111/j.1467-9868.2007.00627.x.
URL https://doi.org/10.1111/j.1467-9868.2007.00627.x

[4] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM Review 51 (3) (2009) 455–500. arXiv:https://doi.org/10.1137/07070111X, doi:10.1137/07070111X.
URL https://doi.org/10.1137/07070111X

[5] R. Tibshirani, T. Hastie, J. Friedman, Regularized paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (02 2010). doi:10.1163/ej.9789004178922.i-328.7.

[6] H. Zhou, L. Li, H. Zhu, Tensor regression with applications in neuroimaging data analysis, Journal of the American Statistical Association 108 (2013) 540–552. doi:10.1080/01621459.2013.776499.

[7] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational Radiomics System to Decode the Radiographic Phenotype, Cancer Research 77 (21) (2017) e104–e107. arXiv:https://aacrjournals.org/cancerres/article-pdf/77/21/e104/2934659/e104.pdf, doi:10.1158/0008-5472.CAN-17-0339.
URL https://doi.org/10.1158/0008-5472.CAN-17-0339

[8] J. Fan, J. Lv, Sure independence screening for ultra-high dimensional feature space (2008). arXiv:math/0612857.
URL https://arxiv.org/abs/math/0612857

Appendix A. Hyperparameters for simulated data

Appendix A.1. Data generation

In our simulations, we unfold the β^l of each pictogram line by line into a vector (rather than a matrix) and concatenate these vectors to obtain $\boldsymbol{\beta} = [\boldsymbol{\beta}^1; \cdots \boldsymbol{\beta}^L]$. Let N be the size of $\boldsymbol{\beta}$. We then use the following parameters to generate the simulated data:

- $\boldsymbol{\mu}_0 = \mathbb{0}_N$
- $\boldsymbol{\mu}_1 = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$
- $\mathbf{P}$ is obtained by completing in orthonormal basis $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$
- $d_1 = 0.01$
- For $i \in [\![2, N]\!]$, $d_i = 0.25$ (where (d_i) are the diagonal elements of $\mathbf{D}$)

On 1000 individuals generated, with 500 in each class, the accuracy obtained by the 2-means algorithm is 0.48: in other words, it doesn't do better than chance (even slightly worse in our case). However, the section 4 shows far better performance from our models

Appendix A.2. Cross validation of models

For the lasso and group lasso models, we cross-validate on 20 values of λ distributed according to a logarithmic scale between 10^{-5} and 10^{-13}.

For the muliway and multiway mulibloc models, we cross-validate on 5 values of λ distributed on a logarithmic scale between 10^{-3} and 10^{-6}. We also cross-validate on the rank used. The rank of $\boldsymbol{\beta}^*$ is bounded above by the sum of the ranks of the individual pictograms (around 15, 1 and 10 respectively). So we won't exceed 27 for rank in our cross validation. In fact, we hope to approximate β with a matrix of rank lower than 26 to ensure a certain sparsity in the model. We therefore cross-validate on ranks 1, 10 and 20 in the multiway model (to propose a model parameterization with a minimal rank, an intermediate rank and a high rank).

In the MMLR, we can choose a rank for each pictogram. In order not to overload the cross-validation, we limit ourselves to cross-validating on the rank of the first pictogram. As this is the highest-ranking pictogram, there are more different "choices" possible for the MMLR, depending on the desired sparsity. We therefore propose the ranks 1, 6 and 12 for this pictogram in the cross validation. For the other pictograms, we impose 1 and 10 respectively. Here again, we also cross-validate on 5 values of λ, distributed on a logarithmic scale between 10^{-3} and 10^{-6}.

Appendix B. Parameters used for feature extraction with pyradiomics

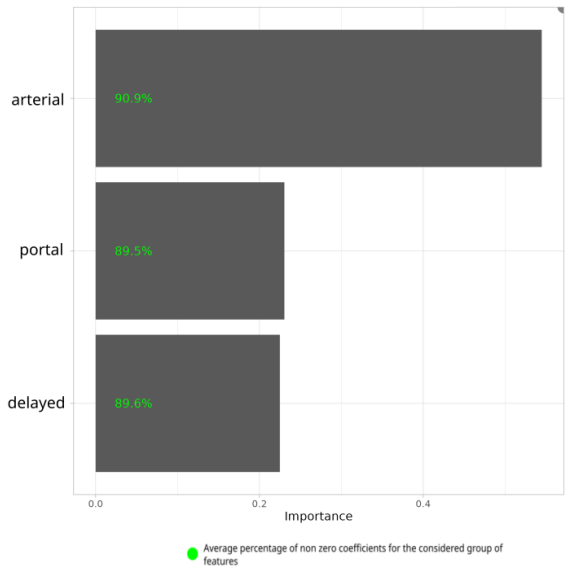
List of parameters used for feature extraction by pyradiomics:

- Bin width : 25
- Resampled Pixel Spacing : $[2, 2, 2]$ si l'extraction est en 3D, $[2, 2]$ si elle est en 2D
- interpolator : sitkBSpline
- force2D : True
- force2Ddimension : 2

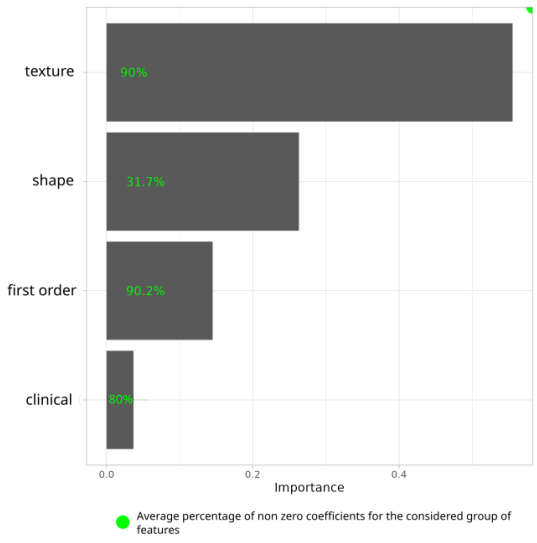
Appendix C. Importance of features

Importance graphs are given here for the best-performing models (in terms of AUC, see Table 3), namely the group lasso model with block grouping for 3D data and the classic lasso for 2D data. The importance of a feature is measured as the absolute value of the β coefficient in front of it. This value is then averaged over all simulations to find the feature's importance. Given the large number of features, we group them by block, mode and/or variable name. The importance of a group is given as the sum of the importances of its features. All group importances are finally renormalized so that their sum is 1 (we're interested in the relative importance of features in relation to each other).

On each stick of each bar chart, in addition to the importance, there is a percentage in green. This provides information on the number of times the coefficient in front of the features associated with the stick has been non-zero in the simulations. As all our models are penalized by the lasso, they tend to set the coefficients of the least important variables to zero. We can therefore calculate for each variable, over the 50 training sessions carried out, the percentage of times this coefficient was non-zero. The average of these percentages over all the features in a feature group (block, mode or variable) is shown in green on the graph. This average reflects the number of times the features in this group were deemed important by the model (i.e. their regression coefficient was non-zero).

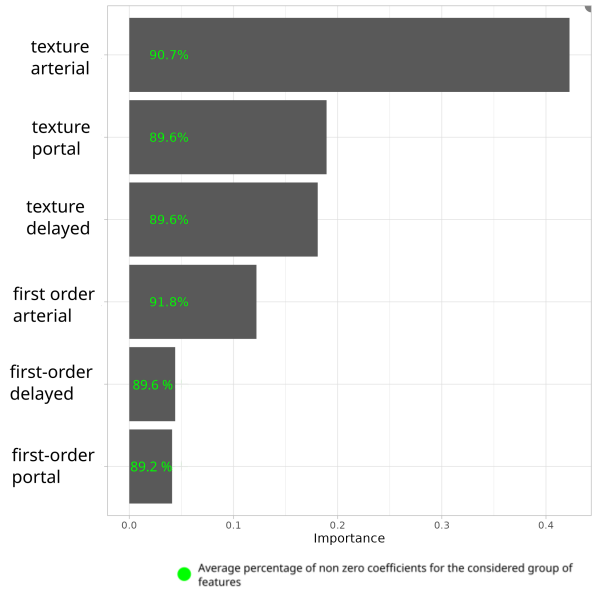


(a) Relative importance of time on 3D data (excluding clinical data)

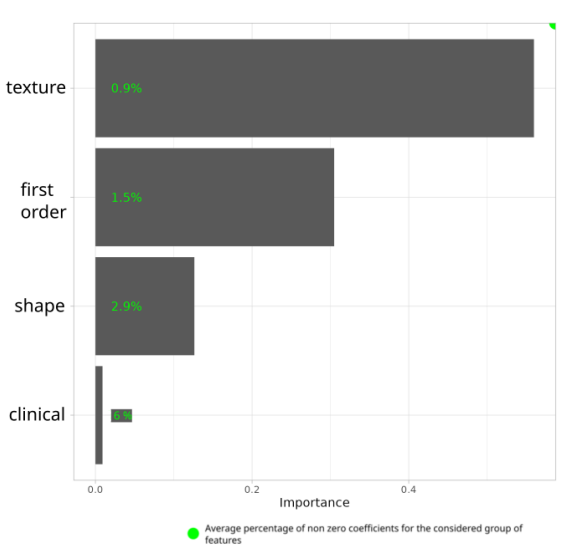


(b) relative importance of blocks on 3D data

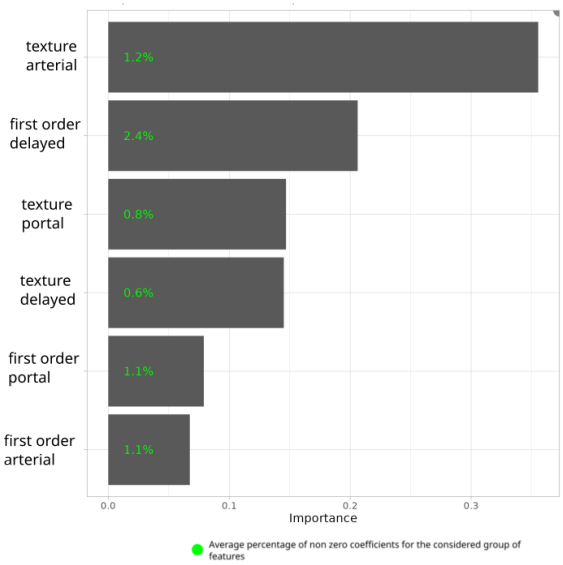
Fig. C.9: Relative importance of features for 3D data with the best-performing model for this data: group lasso with block grouping (part 1)



(c) relative importance of block times on 3D data (excluding clinical data)

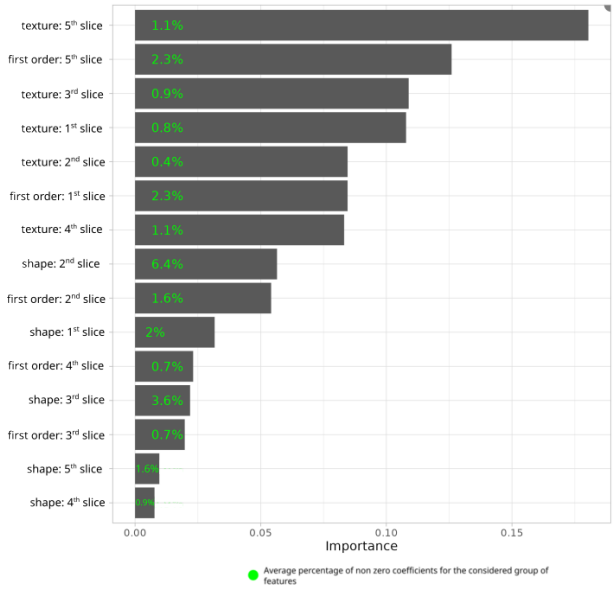


(a) Relative importance of blocks on 2D data



(b) relative importance of times per block on 2D data (excluding clinical data and shape featrues because they do not depend on time)

Fig. C.10: Relative importance of features for 2D data with the best-performing model on these data: lasso logistic regression



(c) relative importance of slices per block on 2D data (excluding clinical data)