Tensor multiblock logistic regression

SELVESTREL Alexandre

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

Supervisors : Arthur Tenenhaus, Laurent Lebrusquet

Medical partner : Henri Mondor hospital, radiologist: Sébastien Mulé

January 15, 2025

Liver tumors classification

6th most widespread cancer and 4th mortality cause by cancer

Classification:

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

Non invasive method: MRI images with injection of contrast agent

Some MRI images

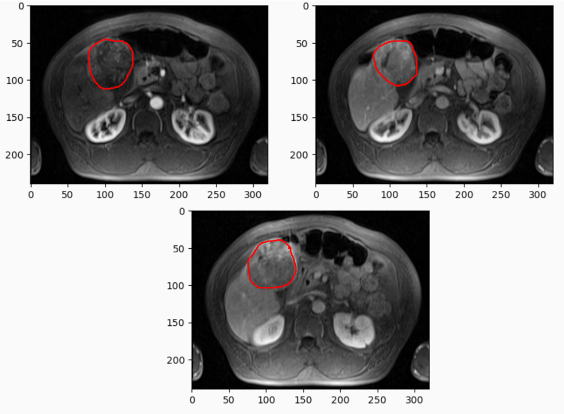


Figure 1: Example of MRI images of a HCC liver tumor (arterial, portal, late) from From Henri Mondor hospital: the 3 images look quite similar

Some MRI images

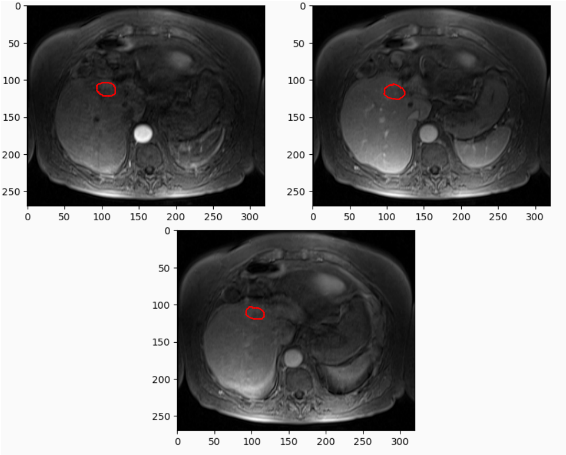


Figure 2: Example of MRI images of a CCK liver tumor (arterial, portal, late) from From Henri Mondor hospital

Available data

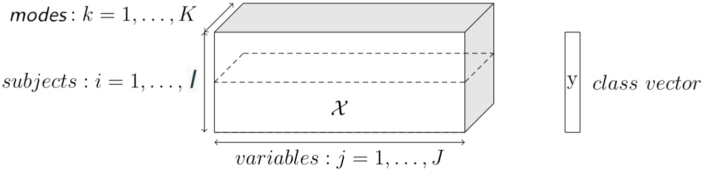
- MRI images in 3D of liver tumors (arterial, portal, late)
- gender (63 men, 27 women)
- age at disease (average: 63 years old)

Same variables extracted from each MRI image at 3 times → **tensor data**

Features grouped by blocks: grey levels intensity, shape, texture, univariate (age and gender)→ **multiblock data**

Tensor data

A given subject *i* is represented by a horizontal slice in the features tensor



$modes : k = 1, \dots, K$
$subjects : i = 1, \dots, I$
$\mathcal{X}$
y
$class\ vector$
$variables : j = 1, \dots, J$

Figure 3: Type of data: tensorial

Warning: Often strong correlation between features from the same variable across different modalities → adapt the model to this structure.

Multiblock data

Each block is a tensor with its own structure.

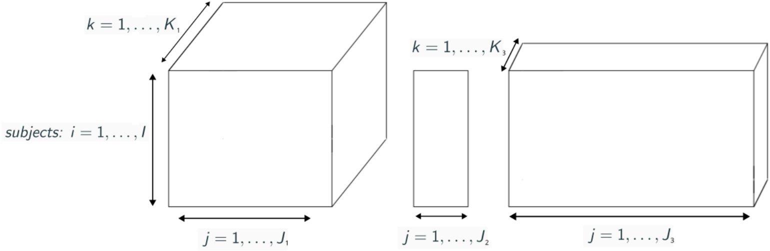


Figure 4: Type of data: multiblock

Table of contents

1. Logistic model for tensor multiblock data
2. Maximizing penalized likelihood
3. Tests on simulated data and application to liver tumor classification

Logistic model for tensor multiblock data

Logistic regression (recall)

Generalized linear model (GLM) for classification:

$\mathbf{x}$: feature vector (explanatory variable)

Y : binary response (explained variable)

Likelihood for logistic regression

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

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

Naive approach for tensor data: unfolding

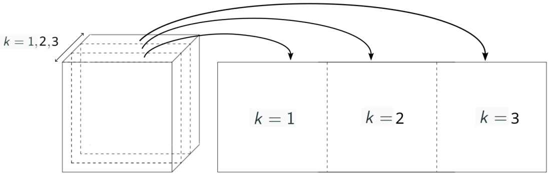


Figure 5: Unfolding a tensor

Naive unfolding

$$\mathbf{B} = (\beta_{j,k})_{j\in[\![1,J]\!],k\in[\![1,K]\!]} \rightarrow JK \text{ parameters to determine}$$

$$\mathbf{x}^T\boldsymbol{\beta} \rightsquigarrow \sum_{j,k} x_{j,k}\beta_{j,k} = \langle\,\mathbf{X}\,|\,\mathbf{B}\,\rangle$$

Limitation: No consideration of the tensor structure in the likelihood

Lasso penalization (recall)

To many features (vs I) → penalization to control variance of prediction (overfitting).

Search for easily interpretable model → choice of lasso:

Lasso

$$\text{penalization} = \lambda \|\beta\|_1 \qquad (\lambda > 0)$$

Function to maximize :

$$\text{penalized likelihood} = \log(\mathcal{L}(\beta)) - \lambda \|\beta\|_1$$

Limitation: No consideration of the tensor structure in the penalization.

Tensor regression model

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

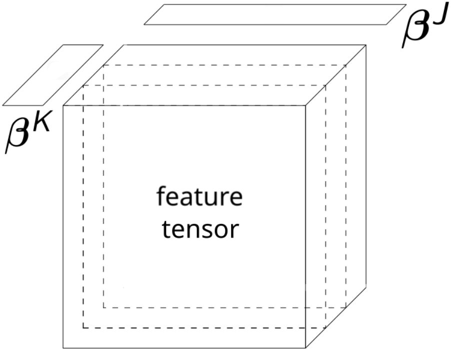


Figure 6: Tensor structure of **B**

Tensor regression model

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

Proposed rank 1 model

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

$$\mathbf{B} = \boldsymbol{\beta}^K \circ \boldsymbol{\beta}^J \qquad (\beta_{j,k} = \beta_j^J \beta_k^K)$$

β_j : impact of variable j

β_k : impact of modality k

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

Limits of rank 1

$\mathbf{B} = \beta^K \circ \beta^J$ implies a complete separation between columns and rows:

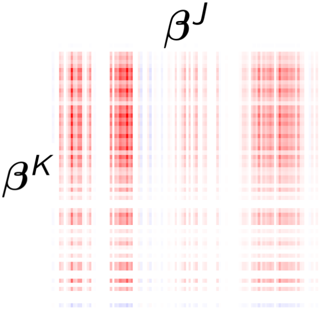


Figure 7: Heatmap of a rank 1 matrix: each pixel represents a number (from 0 in white to 1 in red)

This can be too simplistic.

Extension to rank R [1]

Rank R lasso penalized tensor model

$$\text{Summing rank 1 together : } \mathbf{B} = \sum_{r=1}^{R} \beta_r^J \circ \beta_r^K$$

$$\text{lasso like penalization} = \lambda \sum_{r=1}^{R} \|\beta_r^J \circ \beta_r^K\|_1 = \lambda \sum_{r=1}^{R} \|\beta_r^J\|_1 \|\beta_r^K\|_1$$

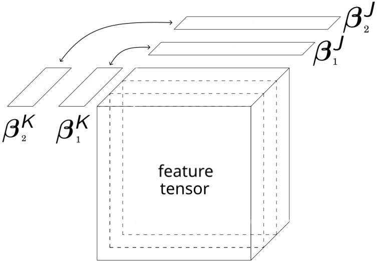


Figure 8: Tensor structure of β for rank 2

Blocks of variables

Aim: GLM framework for multiblock tensor data.

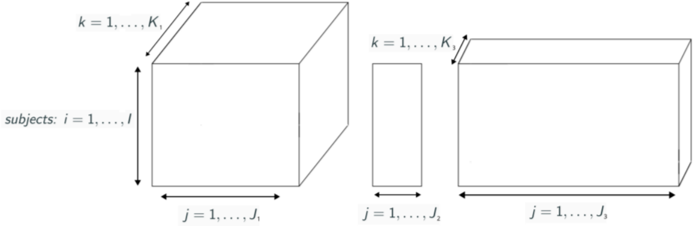


Figure 9: One tensor per type of variable in multiblock data

β^J and β^K have different meanings and sizes for each block of data.

Logistic model for tensor multiblock data

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

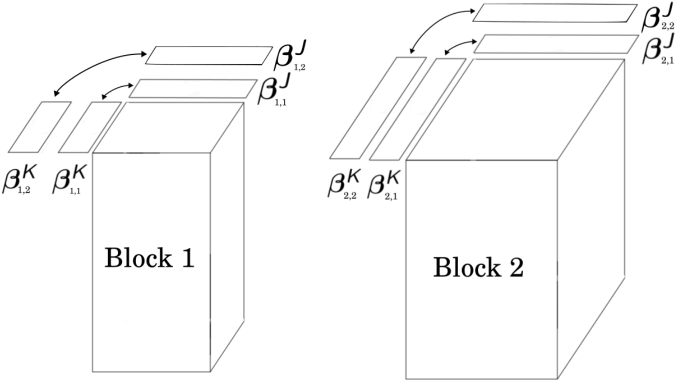


Figure 10: Tensor multiblock model for rank 2

Tensor multiblock logistic regression

Scalar product for L blocks

$$\mathbf{x}^T\beta \rightsquigarrow \sum_{\ell} \langle \mathbf{X}^\ell \,|\, \mathbf{B}_\ell \rangle = \sum_{\ell=1}^{L} \sum_{j,k} x^\ell_{j,k} (\beta_\ell)_{j,k}$$

Regression coefficient for block ℓ

$\mathbf{B}_\ell$ can have any rank R_ℓ

$$\mathbf{B}_\ell = \sum_{r=1}^{R_\ell} \boldsymbol{\beta}^J_{\ell,r} \circ \boldsymbol{\beta}^K_{\ell,r}$$

Penalization

$$\text{Lasso like penalty} = \lambda \sum_{\ell,r} \|\boldsymbol{\beta}^K_{\ell,r} \circ \boldsymbol{\beta}^J_{\ell,r}\|_1 = \lambda \sum_{\ell,r} \|\boldsymbol{\beta}^K_{\ell,r}\|_1 \|\boldsymbol{\beta}^J_{\ell,r}\|_1$$

Maximizing penalized likelihood

likelihood term

Scalar product of x and β

$$\mathbf{x}^T\beta \rightsquigarrow \sum_{\ell,r,j,k} x^{\ell}_{j,k}(\beta^{J}_{\ell,r})_j(\beta^{K}_{\ell,r})_k \tag{1}$$

$$= \sum_{\ell,r,j}\left(\sum_{k} x^{\ell}_{j,k}(\beta^{K}_{\ell,r})_k\right)(\beta^{J}_{\ell,r})_j \tag{2}$$

Partial optimization problem

Optimizing the likelihood along mode $J \Leftrightarrow$ solving a logistic regression on weighted aggregated data $\sum_{k} x^{\ell}_{j,k}(\beta^{K}_{\ell,r})_k$

Algorihm

Similar result for mode K. Possibility to optimize the likelihood by alternating between modes (can be easily adapted for lasso penalization)

Tests on simulated data and application to liver tumor classification

Data generation: example in 2D

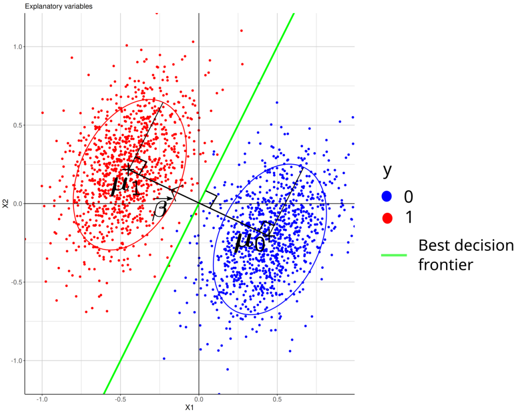


Figure 11: Example of explanatory variables for $\beta = (-2, 1)$

Possibility to choose $(\sigma_\beta, \sigma_{\text{noise}})$ to define the problem difficulty.

AUC on simulated data

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

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

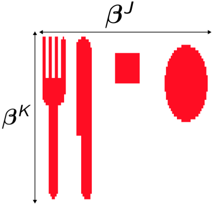


Figure 12: Pictogram for non multiblock models

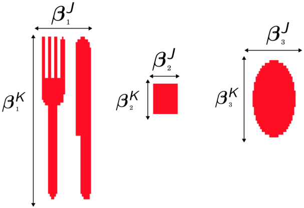
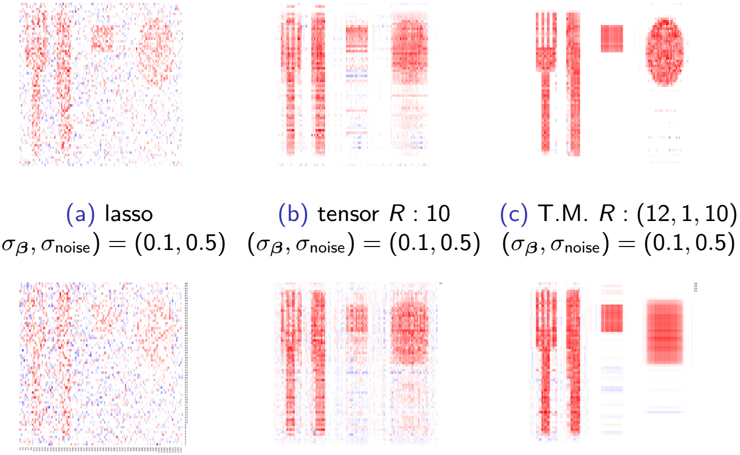


Figure 13: Pictogram for tensor multiblock model

Reconstructed β



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

(b) tensor $R: 10$
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.5)$

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

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

(e) tensor $R: 10$
$(\sigma_\beta, \sigma_{\text{noise}}) = (0.1, 0.8)$

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

Results on liver tumor data

lasso	group lasso (block)	group lasso (time)	group lasso (var)	tensor	tensor blocks
0.74 ± 0.04	0.78 ± 0.03	0.76 ± 0.03	0.73 ± 0.03	0.77 ± 0.03	0.77 ± 0.03

Cross validated AUC on 3D real data

Performances of tensor models similar to those of the best model (here group lasso, with grouping of features by block), but better explainability (less parameters to determine).

Conclusion

State of the art performances.

Better than state of the art in simulated data when few overlapping between classes.

Scales better than regular logistic regression for high order tensors.

Lowers the complexity of the regression model an therefore reduces overfitting.

Good interpretability (sparse + displays importance of each block, mode and variable in $\boldsymbol{\beta}$).

Further work

Testing on other real datasets wether the performances on the simulated dataset can be replicated.

Testing other penalizations (group lasso, elastic net).

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

Improving the optimization, by using coordinate descent (as done in glmnet [3] in R).

Bibliography

- Fabien Girka, Pierrick Chevaillier, Arnaud Gloaguen, Giulia Gennari, Ghislaine Dehaene-Lambertz, Laurent Le Brusquet, and Arthur Tenenhaus.
  Rank-R Multiway Logistic Regression.
  In *52èmes Journées de Statistique*, Nice, France, 2021.
  les 52èmes journées de Statistique 2020 sont reportées ! Elles auront lieu du 7 au 11 Juin 2021.

- Laurent Le Brusquet, Gisela Lechuga, and Arthur Tenenhaus.
  Régression Logistique Multivoie.
  In *JdS 2014*, page 6 pages, Rennes, France, June 2014.

- Rob Tibshirani, Trevor Hastie, and Jerome Friedman.
  Regularized paths for generalized linear models via coordinate descent.
  Journal of Statistical Software, 33, 02 2010.

Annex

Lasso penalty in fitting algorithm

Rewriting the penalty

$$\text{penalty} \propto \sum_{\ell,r} \left(\|\boldsymbol{\beta}^{K}_{\ell,r}\|_1 \|\boldsymbol{\beta}^{J}_{\ell,r}\|_1 \right) \tag{3}$$

$$= \sum_{\ell,r} \left(\left\| \|\boldsymbol{\beta}^{K}_{\ell,r}\|_1 \boldsymbol{\beta}^{J}_{\ell,r} \right\|_1 \right) = \sum_{\ell,r} \left(\left\| \|\boldsymbol{\beta}^{J}_{\ell,r}\|_1 \boldsymbol{\beta}^{K}_{\ell,r} \right\|_1 \right) \tag{4}$$

Lasso penalty in fitting algorithm

Rewriting the penalty

$$\text{penalty} \propto \sum_{\ell,r} \left(\|\boldsymbol{\beta}^K_{\ell,r}\|_1 \|\boldsymbol{\beta}^J_{\ell,r}\|_1 \right) \tag{3}$$

$$= \sum_{\ell,r} \left(\left\| \|\boldsymbol{\beta}^K_{\ell,r}\|_1 \boldsymbol{\beta}^J_{\ell,r} \right\|_1 \right) = \sum_{\ell,r} \left(\left\| \|\boldsymbol{\beta}^J_{\ell,r}\|_1 \boldsymbol{\beta}^K_{\ell,r} \right\|_1 \right) \tag{4}$$

Strategy

dilate $\boldsymbol{\beta}^J_{\ell,r}$ by $\|\boldsymbol{\beta}^K_{\ell,r}\|_1$ and $x^\ell_{j,k}$ by $\|\boldsymbol{\beta}^K_{\ell,r}\|_1^{-1}$, so

$$\mathbf{x}^T \boldsymbol{\beta} \rightsquigarrow \sum_{j,\ell,r} \left(\sum_k x^\ell_{j,k} (\beta^K_{\ell,r})_k \right) (\beta^J_{\ell,r})_j$$

does not change but

$$\|\boldsymbol{\beta}^J\|_1 \rightsquigarrow \sum_{\ell,r} \left(\|\boldsymbol{\beta}^K_{\ell,r}\|_1 \|\boldsymbol{\beta}^J_{\ell,r}\|_1 \right)$$

Lasso penalty in fitting algorithm

New optimization problem

After the dilations presented in the previous slide, we get:

$$\tilde{x}_{\ell,r,j} = \sum_{k} x^{\ell}_{j,k} \|\boldsymbol{\beta}^{K}_{\ell,r}\|_1^{-1} \tag{5}$$

$$\tilde{\beta}^{J}_{l,r,j} = (\beta^{J}_{\ell,r})_j \|\boldsymbol{\beta}^{K}_{\ell,r}\|_1 \tag{6}$$

So that

$$\mathbf{x}^{T}\boldsymbol{\beta} \rightsquigarrow \langle \underline{\tilde{\mathbf{X}}} \mid \underline{\tilde{\mathbf{B}}}^{J} \rangle \tag{7}$$

$$\text{penalty} = \lambda \|\underline{\tilde{\mathbf{B}}}\|_1 \tag{8}$$

Thus, it is possible to do standard logistic lasso regression with $\underline{\tilde{\mathbf{X}}}$ (unfolded) as features and λ as penalty to find the coefficients $\underline{\tilde{\mathbf{B}}}^{J}$, which can be easily related to those of the vectors $(\boldsymbol{\beta}^{J}_{\ell,r})$.

Everything works symetrically for mode K.

Stopping criterion

Penalized likelihood

$$C = \log(\mathcal{L}(\beta)) - \text{penalty}$$

Before the t-th optimization cycle, its value is C^t and after this cycle it becomes C^{t+1}.

Stopping criterion

$$|C^{t+1} - C^t| < \epsilon |C^t|$$

(typically $\epsilon = 10^{-4}$)

Data generation

Theorem for data generation

For a given $\boldsymbol{\beta}$ to be reconstructed (pictograms).

If the $(\mathbf{x}_i)_{i\in[\![1,I]\!]}$ are generated with 2 multivariate normal laws of means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ and common covariance matrix $\boldsymbol{\Sigma}$ such that:

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

Then $\boldsymbol{\beta}$ is the normal vector to the best separating hyperplane between the two classes (which is in this case the Bayes classifier.)

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