Rank-R Multiway Logistic Regression

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

Rank-R Multiway Logistic Regression

Fabien Girka [1], Pierrick Chevaillier [1], Arnaud Gloaguen [2,3], Giulia Gennari [4], Ghislaine Dehaene-Lambertz [4], Laurent Le Brusquet [2] & Arthur Tenenhaus[2]

[1] *CentraleSupélec, 91190, Gif-sur-Yvette, France,*
[2] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.*
[3] *UNATI, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France,*
[4] *INSERM, UMR992, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France*
fabien.girka@supelec.fr, pierrick.chevaillier@supelec.fr, arnaud.gloaguen@centralesupelec.fr

Résumé. Beaucoup de données ont une structure intrinsèque tensorielle lorsque, par exemple, plusieurs modalités de la même variable ont été observées sur chaque individu. Les approches multivoie deviennent alors un choix naturel pour analyser ces données. Les versions standards de ces approches consistent à imposer au vecteur de poids d'être une décomposition de PARAFAC de rang 1. Pour certaines applications, les données peuvent cependant s'avérer trop complexes pour que cette décomposition soit valide. Ce papier présente une version de la régression logistique multivoie associée à une décomposition de rang R, l'objectif étant de proposer une méthode de classification applicable aux situations où la contrainte de rang 1 serait trop restrictive. Un algorithme de directions alternées est proposé pour la régression logistique multivoie de rang R. Les performances de cette méthodes sont évaluées sur des données d'électroencéphalogrammes (EEG).

Mots-clés. Analyse multivoie, régression logistique, EEG

Abstract. Data often has an inherent tensor structure (e.g. data where the same set of variables is collected at different occasions). To deal with such data, multiway models become a natural choice. Standard multiway models impose weight vectors to be rank-1 PARAFAC decomposition. However in some applications, this constraint appears to be too restrictive. This paper presents a more general version of multiway logistic regression (MLR) associated to a rank-R decomposition. The objective of such an approach is to propose a classification model that can cope with situations where rank-1 constraint may be too restrictive. An alternating direction algorithm is proposed for rank-R MLR and its performances are evaluated on electroencephalogram (EEG) data.

Keywords. Multiway analysis, logistic regression, EEG

1 Introduction

Multiway data appears in many research fields as neuroscience, chemometrics or social networks to name a few. It occurs when the same set of variables is collected through

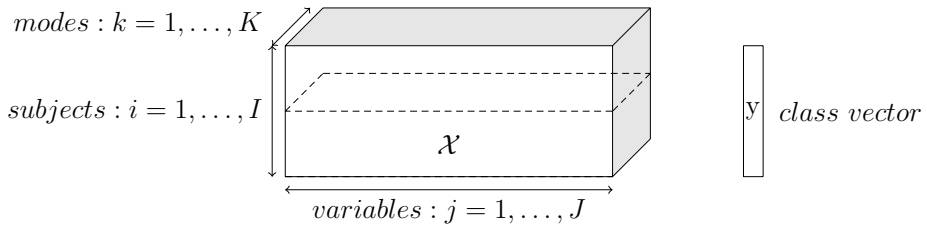


Figure 1: Three-way data, each sample is represented by K vectors of length J

different modes. For example, this is the case for spatio-temporal data (several images collected at different time steps) or when a set of measures is acquired through multiple sensors. See [1] for an overview of methods for the analysis of multiway data.

Let $\mathcal{X} = \{\mathrm{X}_{ijk}\}_{1\leq i\leq I, 1\leq j\leq J, 1\leq k, \leq K}$ be a third order tensor of dimension $I \times J \times K$ where I is the number of subjects, J the number of variables and K the number of modes (see Figure 1). The roles of variables and modes are symmetrical and thus can be interchanged.

A first non-multiway option to deal with such tensor data is to unfold the tensor by concatenating its frontal slices $\mathrm{X}_{..k}$ next to each other, leading to $\mathrm{X} = \begin{bmatrix}\mathrm{X}_{..1} & \dots & \mathrm{X}_{..K}\end{bmatrix}$ which is a matrix of size $I \times JK$ on which classic statistical methods can be applied. However, such an approach leads to issues: (i) A problem of size JK could be computationally impractical for standard computers, (ii) the higher the dimension/order of the tensor, the more the number of variables obtained by flattening, the higher the risk of overfitting is, (iii) the results are not easy to interpret as the considered model does not permit a separate interpretation of the influence of the variables and modes.

A second approach largely used in the multiway literature when the model to construct uses the tensor variables through a linear form is to impose a Kronecker constraint on the weight vector: $\boldsymbol{\beta} = \boldsymbol{\beta}^J \otimes \boldsymbol{\beta}^K$. Taking into account the multiway structure of the data with a Kronecker constraint reduces the degree of freedom from JK to $J + K$, which may limit overfitting effects and computation time. Moreover, the study of the contributions of the variables and modes is made easier by analysing each vector $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ separately.

This rank-1 constraint was used for example in Multiway Logistic Regression (MLR) [2]. However, such a constraint can be too restrictive when effect of variables and modes are not strictly parallel. Therefore a higher rank decomposition for $\boldsymbol{\beta}$ can be considered :

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

Similar rank R approaches has been studied in [3] for Generalized Linear Model and [4] for Support Vectors Machines. In this paper, we propose a rank-R Multiway Logistic Regression.

This paper is organized as follows: Section 2 presents a short reminder of (regularized) logistic regression. The rank-R logistic regression is described in Section 3. Section 4 presents an application on EEG data.

2 Regularized Logistic Regression

Logistic regression can be directly used on tensor data by unfolding the tensor in a matrix. Let x_i be the vector of the KJ observed variables of subject i (x_i is then the i^{th} line of the unfolded tensor X), y_i its class (either 0 or 1). Logistic regression relies on maximising the conditional log-likelihood $\sum_{i=1}^{n} \log \mathbb{P}(y_i \mid \mathrm{x}_i)$ under the assumption that the conditional probabilities log-ratio is linear:

$$\log\left(\frac{\mathbb{P}(y=1 \mid \mathrm{x}_i)}{1-\mathbb{P}(y=1 \mid \mathrm{x}_i)}\right) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i$$

with β_0 and $\boldsymbol{\beta}$ parameters of the model. From this model it comes the following expression of the regularized log-likelihood:

$$\mathcal{C}(\beta_0, \boldsymbol{\beta}, \mathrm{X}, \mathrm{y}, \lambda) = \sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i) - \log(1+\exp(\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i)) - \lambda g(\boldsymbol{\beta}) \tag{2}$$

where $\lambda > 0$ is a regularization parameter that can be tuned and $g(\beta)$ is a penalty term.

3 Multiway Logistic Regression

In this section, rank-R multiway logistic regression (R-MLR) is presented. R-MLR is defined as the following optimization problem:

$$\max_{\beta_0,\ \boldsymbol{\beta}^K,\ \boldsymbol{\beta}^J} \ \mathcal{C}(\beta_0, \boldsymbol{\beta}, \mathrm{X}, \mathrm{y}, \lambda) \quad \text{s.t.} \quad \boldsymbol{\beta} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \tag{3}$$

where $\boldsymbol{\beta}^J = \left[(\boldsymbol{\beta}_1^J)^{\mathsf{T}} \ \ldots \ (\boldsymbol{\beta}_R^J)^{\mathsf{T}}\right]^{\mathsf{T}}$ and $\boldsymbol{\beta}^K = \left[(\boldsymbol{\beta}_1^K)^{\mathsf{T}} \ \ldots \ (\boldsymbol{\beta}_R^K)^{\mathsf{T}}\right]^{\mathsf{T}}$. An alternating direction algorithm that monotonically converges is proposed to solve the optimization problem (3). First, we can note that $\boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i$ can be expressed as:

$$\left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right)^{\mathsf{T}} \mathrm{x}_i = \sum_{r=1}^{R} (\boldsymbol{\beta}_r^J)^{\mathsf{T}} \left((\boldsymbol{\beta}_r^K)^{\mathsf{T}} \otimes \mathrm{I}_J\right) \mathrm{x}_i \doteq \sum_{r=1}^{R} (\boldsymbol{\beta}_r^J)^{\mathsf{T}} \mathrm{z}_{r,i}^J \tag{4}$$

In addition, two types of regularisation are considered in this paper. First an ℓ_2 penalty with $g(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$ that can be expressed in terms of $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ as follows:

$$\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} = \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right)^\mathsf{T}\left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right) = \sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J\|_2^2\|\boldsymbol{\beta}_r^K\|_2^2 + 2\sum_{i=1}^{R}\sum_{j=i+1}^{R} (\boldsymbol{\beta}_i^J)^\mathsf{T}\boldsymbol{\beta}_j^J(\boldsymbol{\beta}_i^K)^\mathsf{T}\boldsymbol{\beta}_j^K$$
$$= (\boldsymbol{\beta}^J)^\mathsf{T}\mathrm{R}^J\boldsymbol{\beta}^J$$

where $\mathrm{R}^J = \left((\mathrm{B}^K)^\mathsf{T}\mathrm{B}^K\right) \otimes \mathrm{I}_J$, with $\mathrm{B}^K = \begin{bmatrix}\boldsymbol{\beta}_1^K & \dots & \boldsymbol{\beta}_R^K\end{bmatrix}$. As long as the columns of B^K are not colinear, R^J is symmetric positive definite and $\mathrm{Q}_{\ell_2}^J = (\mathrm{R}^J)^{-\frac{1}{2}} = \left((\mathrm{B}^K)^\mathsf{T}\mathrm{B}^K\right)^{-\frac{1}{2}} \otimes \mathrm{I}_J$ is well defined.

Furthermore, by setting $g(\boldsymbol{\beta}) = \sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\|_1$, R-MLR with variable selection can also be defined. The interest of this structured sparsity-inducing norms is that, with a single parameter λ, the sparsity will spread (driven by the data) among ranks, variables and modes. This penalty term can be expressed as a function of $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$:

$$\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\|_1 = \|\mathrm{R}^J\boldsymbol{\beta}^J\|_1 \tag{5}$$

with $\mathrm{R}^J = \left(\|\boldsymbol{\beta}_r^K\|_1 \mathrm{I}_J\right)_{r\in\{1\dots R\}}$ a block diagonal matrix. As before, $\mathrm{Q}_{\ell_1}^J = (\mathrm{R}^J)^{-1} = \left(\|\boldsymbol{\beta}_r^K\|_1^{-1}\mathrm{I}_J\right)_{r\in\{1\dots R\}}$ is well defined.

As a consequence, the objective function of the optimization problem (3) can be expressed in terms $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ and therefore can be maximized with respect to β_0, $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ using an alternating direction algorithm. Indeed, optimising w.r.t. $(\beta_0, (\mathrm{Q}^J)^{-1}\boldsymbol{\beta}^J)$ can be seen as applying logistic regression to maximise criterion $\mathcal{C}^J = \mathcal{C}(\beta_0, (\mathrm{Q}^J)^{-1}\boldsymbol{\beta}^J, Q^J Z^J, \mathrm{y}, \lambda)$ with Z^J a matrix of size $I \times J$ and $\mathrm{z}_i^J = \begin{bmatrix}(\mathrm{z}_{1,i}^J)^\mathsf{T} & \dots & (\mathrm{z}_{R,i}^J)^\mathsf{T}\end{bmatrix}^\mathsf{T}$. As variables and modes play symmetric roles, the same can be done with $\boldsymbol{\beta}^K$. We can now derive the R-MLR algorithm presented in Algorithm 1.

4 Application on EEG data and discussion

The objective of this study was to identify whether the infant's brain encodes the phonetic features used by linguists to describe speech. 24 different consonant-vowel syllables were presented to 25 infants in a randomized order every 1000 ms during one-hour-long experimental sessions. Brain responses were recorded at 500 Hz with a high-density EEG net comprising 252 channels. After pre-processing, this EEG experiment yields 25 tensors of size 24 syllables $\times$ 500 time steps $\times$ 252 channels each. The consonants varied along the manner of articulation separating the 24 syllables in 2 classes that we want to predict.

Algorithm 1: Rank R Multiway Logistic Regression

Inputs: $\epsilon > 0,\ \lambda,\ R,\ \boldsymbol{\beta}^{K(0)}$, penalty

$q \longleftarrow 0$

repeat

$\mathrm{Z}_r^J = \sum_{k=1}^{K} \left(\boldsymbol{\beta}_r^{K(q)}\right)_k \mathrm{X}_{..k}$ for $r \in \{1, \ldots, R\}$

$\left(\mathrm{Z}^J, \boldsymbol{\beta}^{K(q)}\right) \longleftarrow \left(\left[(\mathrm{Z}_1^J)^\mathsf{T} \quad \ldots \quad (\mathrm{Z}_R^J)^\mathsf{T}\right]^\mathsf{T},\ \left[(\boldsymbol{\beta}_1^{K(q)})^\mathsf{T} \quad \ldots \quad (\boldsymbol{\beta}_R^{K(q)})^\mathsf{T}\right]^\mathsf{T}\right)$

if ℓ_1 penalty $Q^J \longleftarrow Q_{\ell_1}^J$ **else** $Q^J \longleftarrow Q_{\ell_2}^J$

$(\beta_0^{(q)}, (Q^J)^{-1}\boldsymbol{\beta}^{J(q)}) \longleftarrow \mathrm{argmax}_{\beta_0, \boldsymbol{\beta}}\, \mathcal{C}\left(\beta_0, (Q^J)^{-1}\boldsymbol{\beta}, Q^J\mathrm{Z}^J, \mathrm{y}, \lambda\right)$

$\mathrm{Z}_r^K = \sum_{j=1}^{J} \left(\boldsymbol{\beta}_r^{J(q)}\right)_j \mathrm{X}_{.j.}$ for $r \in \{1, \ldots, R\}$

$\left(\mathrm{Z}^K, \boldsymbol{\beta}^{J(q)}\right) \longleftarrow \left(\left[(\mathrm{Z}_1^K)^\mathsf{T} \quad \ldots \quad (\mathrm{Z}_R^K)^\mathsf{T}\right]^\mathsf{T},\ \left[(\boldsymbol{\beta}_1^{J(q)})^\mathsf{T} \quad \ldots \quad (\boldsymbol{\beta}_R^{J(q)})^\mathsf{T}\right]^\mathsf{T}\right)$

if ℓ_1 penalty $Q^K \longleftarrow Q_{\ell_1}^K$ **else** $Q^K \longleftarrow Q_{\ell_2}^K$

$(\beta_0^{(q)}, (Q^K)^{-1}\boldsymbol{\beta}^{K(q+1)}) \longleftarrow \mathrm{argmax}_{\beta_0, \boldsymbol{\beta}}\, \mathcal{C}\left(\beta_0, (Q^K)^{-1}\boldsymbol{\beta}, Q^K\mathrm{Z}^K, \mathrm{y}, \lambda\right)$

$q \longleftarrow q + 1$

until $|\mathcal{C}^K - \mathcal{C}^J| < \epsilon\ \mathcal{C}^J$;

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

Regularized logistic regression and R-MLR with R from 1 to 3 are evaluated and compared using Leave-One-Out cross validation: the model is trained on all infants except one and tested on the remaining one. The Area Under the ROC Curve (AUC), the regularisation parameter λ and the computational time used to run the 25 folds are reported in Table 1.

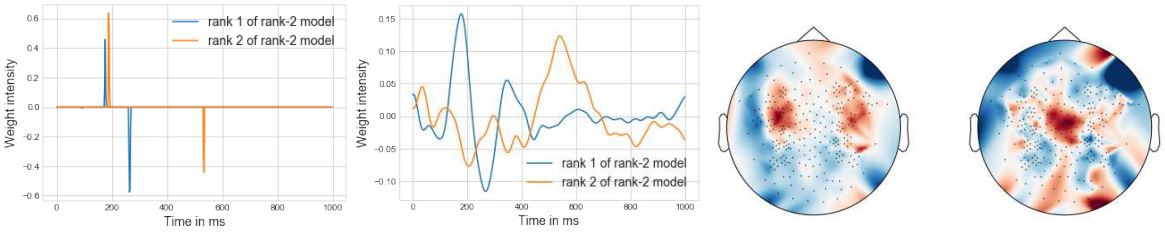
From Table 1, we show that, for complex data such as EEG data, higher rank MLR enable to outperform both rank-1 MLR and regularized logistic regression. For ℓ_1 penalty, finding the best regularization parameter for logistic regression is really challenging given the required computation time. MLR enables to cut down drastically this computation time and yields better results for rank 3.

Model	AUC	λ	Time (in s)
LR	0.822 ± 0.19	1000	892
1-MLR	0.815 ± 0.23	6000	282
2-MLR	0.852 ± 0.17	17500	340
3-MLR	0.857 ± 0.18	17500	498

Model	AUC	λ	Time (in s)
LR	0.830 ± 0.19	20	117000
1-MLR	0.816 ± 0.22	6.5	257
2-MLR	0.826 ± 0.18	10	409
3-MLR	0.853 ± 0.19	15	557

Table 1: Cross validation results by Leave-One-Out for ℓ_2 (left) and ℓ_1 (right) penalties

In addition, multiway models enable the graphical display of the weights and get insights into the importance of variables and modes separately (see Figure 2).



(a) ℓ_1 penalty (b) ℓ_2 penalty (c) ℓ_2 penalty

Figure 2: Weight visualisation for 2-MLR trained on all 25 infants: (a) and (b) show time step weights and (c) shows the topomap of electrode weights.

Topomaps associated with the ℓ_1 penalty (not displayed here due to lack of space) are consistent with the ℓ_2 penalty results. However, similarly to the time weights of the ℓ_1 penalty, they are very sparse with few (less than 5%) isolated selected channels. This variable selection leads spatial/time resolution that is lower than the usual phenomenon observed and described by neuroscientists. Hence, future works include to combine ℓ_1 and ℓ_2 penalties in order to try to catch smoother effects in time and space.

5 Conclusion

Rank-R Multiway Logistic Regression is presented in this paper and shows promising results in EEG application. While R-MLR is presented for third order tensors, it can be generalized to any higher order.

Bibliography

[1] Bro, R. (2000), Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications ICSLP Proceedings.
[2] Le Brusquet L., Lechuga G., Tenenhaus A. (2014), Régression Logistique Multivoie, 46ème Journée de Statistique.
[3] Zhou, Hua; Li, Lexin; Zhu, Hongtu (2013), Tensor Regression with Applications in Neuroimaging Data Analysis, Journal of the American Statistical Association, vol.108.
[4] T. Lyu, E. F. Lock, and L. E. Eberly (2017), Discriminating sample groups with multi-way data, Biostatistics (Oxford, England)