Régression Logistique Multivoie

Laurent Le Brusquet, Gisela Lechuga, Arthur Tenenhaus

Regression Logistique Multivoie

Laurent Le Brusquet[1], Gisela Lechuga[1] & Arthur Tenenhaus[1]

[1] *Supélec Sciences des Systèmes - EA4454 (E3S), Gif-sur-Yvette, France,*
laurent.lebrusquet@supelec.fr, gisela.lechuga@supelec.fr, arthur.tenenhaus@supelec.fr

Résumé. Le papier étend la régression logistique aux données multivoie, c'est-à-dire aux données pour lesquelles, pour chaque individu, plusieurs modalités de la même variable ont été observées. Les données ont ainsi naturellement une structure tensorielle. Pour la régression logistique proposée, les coefficients de la fonction de discrimination sont contraints par une structure tensorielle identique à celle des données. L'intérêt de cette contrainte est double. D'une part elle permet une étude séparée de l'influence des variables et de l'influence des modalités, conduisant ainsi à faciliter l'interprétation du classifieur obtenu. D'autre part, elle permet de restreindre le nombre de coefficients à estimer, et ainsi de limiter à la fois la complexité calculatoire et le phénomène de sur-apprentissage. La méthode proposée est illustrée sur données simulées.

Mots-clés. Données multivoie, régression logistique logistique.

Abstract. In this paper, we propose a formulation of logistic regression for multiway (i.e. data where the same set of variables is collected at different occasions). More specifically, multiway logistic regression (MLR) constraints the coefficients of the logistic model to a tensorial structure that fits the natural structure of the data. Expected improvements of MLR compared with Logistic Regression are (i) better interpretability of the resulting model that allows studying separately the effects of the variables and the effects of modalities, and (ii) limit the number of coefficients to be estimated that decreases the computational burden and allows a better control of the overfitting issue. An aternating directions algorithm is proposed for MLR and the performances are evaluated on simulated data.

Keywords. Multiway analysis, logistic regression

1 Introduction

Ce papier présente une extension de la régression logistique binaire au cas des données multivoie. L'aspect "multivoie" concerne les données explicatives qui ne sont pas représentées par une matrice, comme dans le cas de la régression logistique classique, mais par un tenseur. Le papier présente le cas des tenseurs d'ordre 3 bien que l'approche présentée puisse s'appliquer aux tenseurs d'ordre supérieurs.

Le cas des tenseurs d'ordre 3 se rencontre lorsqu'un ensemble de variables est collecté à différentes occasions. Citons l'exemple des données spatio-temporelles (plusieurs images

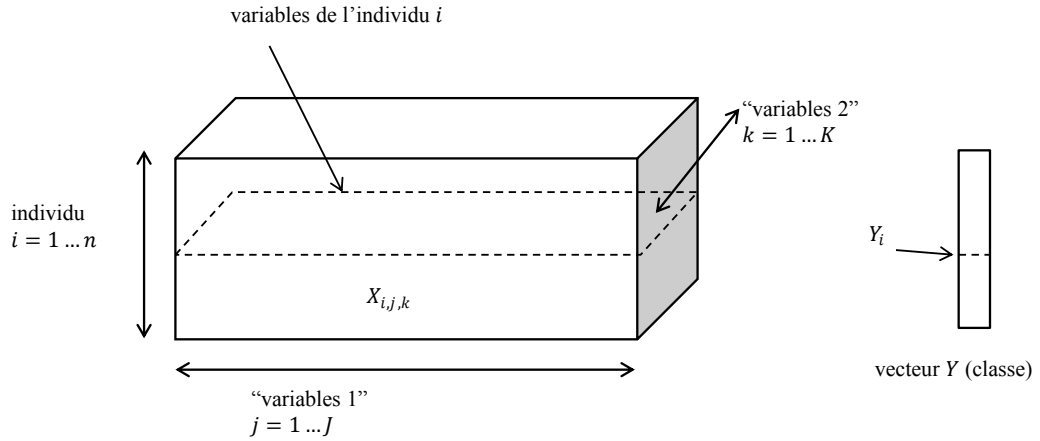

Figure 1: Données organisées selon un tenseur d'ordre 3 : chaque individu est représenté par K vecteurs de taille d.

collectées à différents instants), de mesures spectrales obtenues selon différentes modalités (par exemple à différentes profondeurs) ou bien de mesures vectorielles (par exemple à différentes instants) acquises avec différents capteurs. La litérature propose de nombreuses méthodes d'analyse de données multivoie prenant en compte la structure tensorielle des données (se référer par exemple à l'article de Bro (2000)).

Soit donc $\{\mathbf{X}_{ijk}\}_{1\leq i\leq n, 1\leq j\leq J, 1\leq k\leq K}$ un tenseur d'ordre 3 de dimension $n \times J \times K$ où n désigne le nombre d'individus, J le nombre de variables et K le nombre de modalités (voir figure 1). Les rôles des variables et des modalités sont, pour l'approche proposée, symétriques et peuvent donc être échangés.

Une méthode simple pour traiter le cas des données multivoie consiste à "déplier" le tenseur par concaténation des matrices représentant les différentes modalités : on obtient ainsi une matrice de taille $n \times KJ$ sur laquelle on peut appliquer les méthodes statistiques classiques. Une telle approche n'est pas sans difficulté : outre le fait qu'elle conduit à un problème de taille $K \times J$, pouvant être trop volumineux pour les calculateurs standard, elle pose le problème de l'interprétabilité de l'opérateur obtenu. On peut aussi ajouter au nombre des difficultés le problème de sur-apprentissage, problème classique lorsque le nombre de degrés de liberté de l'opérateur à calculer devient trop important. Ce nombre étant étant en général égal au nombre de variables, on comprend aisément que les données tensorielles offre un cadre propice à la manifestation de ce problème, et cela d'autant plus que l'ordre du tenseur est élevé.

Lorsque l'opérateur à construire, que ce soit à des fins d'analyse, d'estimation ou de classification, fait intervenir les variables du tenseur par l'intermédiaire d'une forme linéaire $\sum_{j,k} \boldsymbol{\beta}_{j,k} X_{i,j,k}$, nous proposons d'imposer au vecteur de coefficients $\boldsymbol{\beta}$ cherché la même structure tensorielle que celle des données :

$$\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J \tag{1}$$

où $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ sont des vecteurs de longueurs K et J. En réduisant ainsi le nombre de degrés de liberté à $K + J$ (au lieu de $K \times J$), on espère limiter le sur-apprentissage, améliorer le temps de calcul, et, en analysant séparément les vecteurs $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$, interpréter séparément l'influence des variables et l'influence des modalités.

Le modèle (1) peut être utilisé pour toute méthode d'analyse statistique dès que les variables interviennent via une forme linéaire. C'est le cas notamment pour la régression multiple, l'analyse en composantes principales, l'analyse discriminante, la régression logistique. Un précédent papier (Le Brusquet et Tenenhaus (2013)) propose l'extension de l'analyse discriminante au cas multivoie. Dans ce papier, nous proposons d'adapter la régression logistique, dans sa version régularisée ou non.
La section 2 revient sur le critère optimisé en régression logistique. La section 3 intègre la contrainte (1) dans l'optimisation de ce critère.

2 Régression logistique standard

La régression logistique peut directement être appliquée aux données tensorielles en "dépliant" le tenseur pour former une matrice. La régression logistique standard est ici brièvement présentée, à la fois parce qu'elle servira de comparaison, mais aussi parce que l'extension proposée utilise le même critère. Pour un individu, notons $\mathbf{x}$ le vecteur contenant l'ensemble des $K \times J$ valeurs observées et y sa classe ($y = 1$ ou -1). La régression logistique repose sur la maximisation de la log-vraisemblance conditionnelle $\sum_{i=1\ldots n} \log \mathbb{P}(y_i/\mathbf{x}_i)$. Un modèle pour les probabilités conditionnelles est utilisé en faisant l'hypothèse que le log-ratio des probabilités conditionnelles est affine :

$$\log \frac{\mathbb{P}(y = 1/\mathbf{x})}{1 - \mathbb{P}(y = 1/\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} \tag{2}$$

β_0 et $\boldsymbol{\beta}$ étant les paramètres du modèle.

En utilisant ce modèle, il vient l'expression de la log-vraisemblance :

$$\sum_{i=1}^{n} y_i \left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - \log\left(1 + \exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right).$$

Maximiser cette log-vraisemblance peut conduire à des classifieurs peu performants en raison du phénomène de sur-apprentissage. Il est courant dans ce cas d'ajouter au critère précédent un terme de régularisation de la forme $\lambda_0 \beta_0^2 + \boldsymbol{\beta}^\top \mathbf{R} \boldsymbol{\beta}$ où la matrice $\mathbf{R}$ permet d'introduire de l'a priori et d'éviter les problèmes numériques. À cet égard, une matrice proportionnelle à l'identité est souvent considérée.

β_0 et $\boldsymbol{\beta}$ sont choisis de façon à optimiser le critère :

$$\mathcal{C}(\beta_0, \boldsymbol{\beta}, \mathbf{X}, \mathbf{y}, \lambda_0, \mathbf{R}) = \sum_{i=1}^{n} y_i \left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - \log\left(1 + \exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right) - \lambda_0 \beta_0^2 - \boldsymbol{\beta}^\top \mathbf{R} \boldsymbol{\beta} \tag{3}$$

La log-vraisemblance, dans sa version pénalisée ou non, est habituellement maximisée par l'algorithme de Newton-Raphson. À noter que la régression logistique à noyau existe et permet à la fois de faire intervenir les variables explicatives via une forme non-linéaire, mais aussi de reduire la complexité calculatoire des problèmes où le nombre de variables explicatives est très grand devant le nombre d'individus (pour plus de détails, se référer par exemple au papier de Cawley (2007)).

3 Régression logistique multivoie

La version multi-voie proposée dans cette section consiste à maximiser le critère (3) habituellement utilisé en régression logistique sour la contrainte que le vecteur $\boldsymbol{\beta}$ soit de la forme $\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$. Cette contrainte permet de limiter les KJ degrés de liberté du vecteur $\boldsymbol{\beta}$ cherché tout en lui imposant une struture cohérente avec la structuration tensorielle des données. L'ajout de la contrainte $||\boldsymbol{\beta}^K|| = 1$ permet de rendre identifiables les vecteurs $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$.

Bien que l'algorithme proposé s'applique à une matrice de régularisation $\mathbf{R}$ quelconque, nous traitons par la suite le cas où, comme $\boldsymbol{\beta}$, elle se factorise sous forme tensorielle $\mathbf{R} = \mathbf{R}^K \otimes \mathbf{R}^J$ afin que l'a priori imposé via la régularisation soit lui aussi cohérent avec la structure tensorielle des données. Les matrices $\mathbf{R}^K$ et $\mathbf{R}^J$ sont de tailles respectives $K \times K$ et $J \times J$.

L'algorithme proposé repose sur l'optimistion du critère du critère (3) par rapport à β_0, $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ par une méthode de directions alternées. En effet, l'optimisation par rapport à $(\beta_0, \boldsymbol{\beta}^K)$ (resp. $(\beta_0, \boldsymbol{\beta}^J)$) peut s'interpréter comme une régression logistique menée sur des données matricielles de taille $n \times K$ (resp. $n \times J$) car :

$$\left(\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J\right)^\top \mathbf{x} = \left(\boldsymbol{\beta}^K\right)^\top \left(\left(\quad_K \otimes \left(\boldsymbol{\beta}_J\right)^\top\right) \mathbf{x}\right) = \left(\boldsymbol{\beta}^J\right)^\top \left(\left(\left(\boldsymbol{\beta}_K\right)^\top \otimes \quad_J\right) \mathbf{x}\right) \tag{4}$$

Cette reformulation des produits scalaires, ainsi que l'égalité (5)

$$\left(\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J\right)^\top \left(\mathbf{R}^K \otimes \mathbf{R}^J\right) \left(\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J\right) = \left(\left(\boldsymbol{\beta}^K\right)^\top \mathbf{R}^K \boldsymbol{\beta}^K\right) \left(\left(\boldsymbol{\beta}^J\right)^\top \mathbf{R}^J \boldsymbol{\beta}^J\right) \tag{5}$$

conduit à l'Algorithme (1).

4 Exemple illustratif - Discussion

L'extension multivoie proposée est ici testée sur des données simulées représentant, pour chacun des $n = 26$ individus, $K = 7$ spectres mesurés à différentes profondeurs. Chaque spectre est constitué de $J = 750$ variables. Le jeu de données comporte 13 individus dans chacune des 2 classes. La figure 2 illustre des exemples de spectres pour 2 individus.

Algorithm 1 Régression logistique multivoie

Require: $\epsilon > 0$, $\beta_0^{(0)}$, $\boldsymbol{\beta}^{K(0)}$

$q \leftarrow 0$

repeat

$\mathbf{Z}^J = \sum_{k=1}^{K} \left(\boldsymbol{\beta}^{K(q)}\right)_k \mathbf{X}_{..k}$

$\left(\beta_0^{(q)}, \boldsymbol{\beta}^{J(q)}\right) \leftarrow \arg\max_{\beta_0,\boldsymbol{\beta}} \mathcal{C}\left(\beta_0, \boldsymbol{\beta}, \mathbf{Z}^J, \mathbf{y}, \lambda_0, \left(\left(\boldsymbol{\beta}^J\right)^\top \mathbf{R}^J \boldsymbol{\beta}^J\right) \mathbf{R}^K\right)$

$\mathbf{Z}^K = \sum_{j=1}^{J} \left(\boldsymbol{\beta}^{J(q)}\right)_j \mathbf{X}_{.j.}$

$\left(\beta_0^{(q)}, \boldsymbol{\beta}^{K(q)}\right) \leftarrow \arg\max_{\beta_0,\boldsymbol{\beta}} \mathcal{C}\left(\beta_0, \boldsymbol{\beta}, \mathbf{Z}^K, \mathbf{y}, \lambda_0, \left(\left(\boldsymbol{\beta}^K\right)^\top \mathbf{R}^K \boldsymbol{\beta}^K\right) \mathbf{R}^J\right)$

$\boldsymbol{\beta}^{K(q)} \leftarrow \dfrac{\boldsymbol{\beta}^{K(q)}}{||\boldsymbol{\beta}^{K(q)}||}$

$q \leftarrow q + 1$

until $||\mathbf{w}^{K(q)} - \mathbf{w}^{K(q-1)}|| < \epsilon$

return $(\mathbf{w}^{K(q)}, \mathbf{w}^{J(q)})$

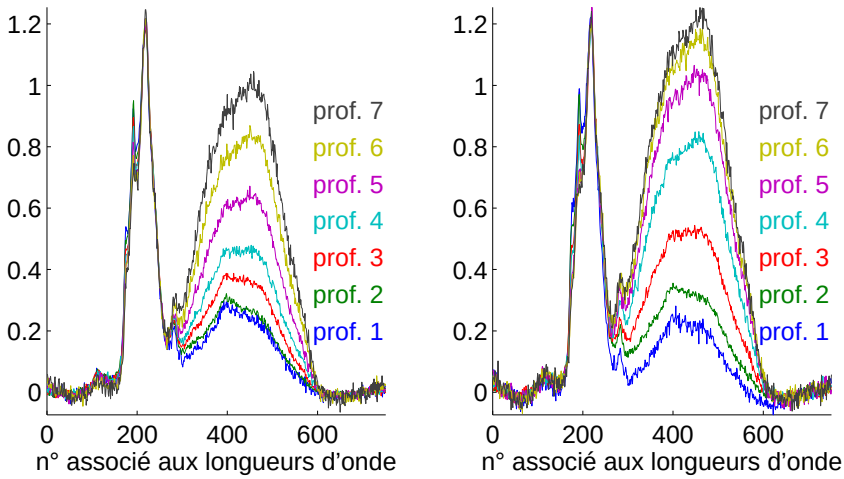

Figure 2: Données simulées pour un individu de chaque classe : pour chaque individu, 7 spectres ont été mesurés.

La régression logistique multivoie est comparée à la régression logistique standard. La régression logistique renvoie un vecteur $\boldsymbol{\beta}$ de longueur $K \times J$ représenté sur la Figure 3. La régression logistique multivoie renvoie un vecteur de longueur J pour décrire l'influence des longeurs d'onde (Figure 4) et un vecteur de longueur K pour décrire l'influence des profondeurs (Table 1) : l'interprétation séparée des 2 effets est ainsi facilitée. En outre, l'a priori structurant a permis d'améliorer les capacités de prédiction puisqu'une validation croisée par Leave-One-Out conduit à 2 individus mal classés avec la régression logistique

menée sur les données dépliées et aucun individu mal classé pour la version multivoie.

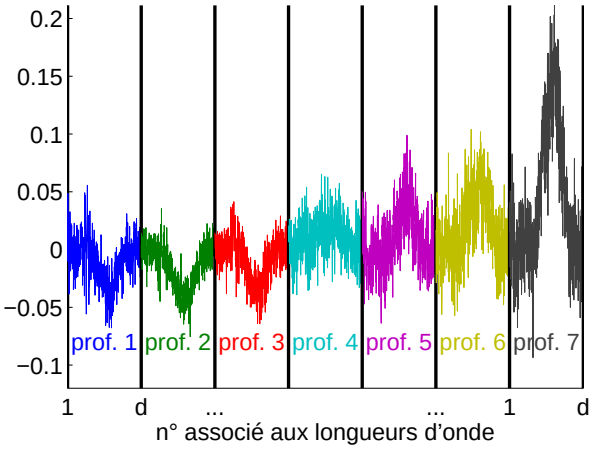

Figure 3: Régression logistique : vecteur β obtenu.

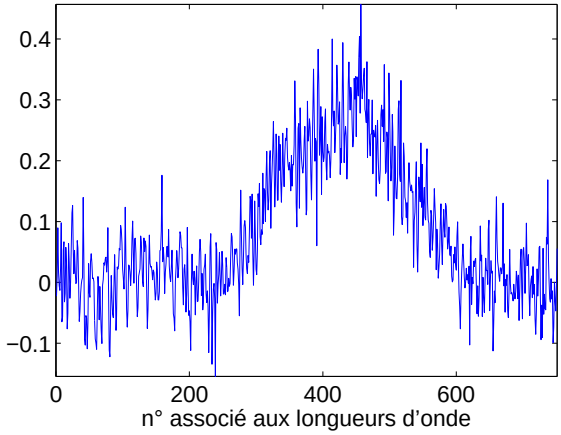

Figure 4: Régression logistique Multivoie : vecteur β_J obtenu.

	prof.1	prof. 2	prof. 3	prof. 4	prof. 5	prof. 6	prof. 7
$\boldsymbol{\beta}^K$	-0.1764	-0.2172	-0.1577	0.1304	0.1710	0.1723	0.5665

Table 1: Régression logistique multivoie : vecteur $\boldsymbol{\beta}^K$ pondérant l'influence des profondeurs.

Bibliographie

[1] Le Brusquet L. et Tenenhaus A. (2013), Analyse Factorielle Discriminante Multi-voie, 45ème Journée de Statistique.
[2] Bro, R. (2000), Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications *ICSLP Proceedings*.
[3] Cawley, G. C., Janacek, G. J., and Talbot, N. L. (2007), Generalised kernel machines. IEEE International Joint Conference on Neural Networks.