Modèles références de régression multinomiale. Propriétés et applications en classification supervisée.

Jean Peyhardi [1,3], Catherine Trottier [1,2] & Yann Guédon [3]

[1] *UM2, Institut de Mathématiques et Modélisation de Montpellier - jean.peyhardi@univ-montp2.fr*
[2] *UM3, Institut de Mathématiques et Modélisation de Montpellier - catherine.trottier@univ-montp3.fr*
[3] *CIRAD, AGAP et Inria, Virtual Plants, 34095 Montpellier - yann.guedon@cirad.fr*

Résumé. De nombreuses extensions du modèle logit ont été introduites dans le cas binomial, comme le modèle probit, mais aucune n'a été proposée dans le cas multinomial non-ordonné. Nous introduisons une nouvelle famille de modèles de régression pour variable réponse nominale construits à partir de fonctions de répartition autres que la logistique et décrivons leur estimation. Pour cela nous mettons en évidence, dans la fonction de lien, la probabilité de chaque catégorie conditionnée par la catégorie de référence. Au contraire du modèle logit multinomial, le choix de cette catégorie de référence affecte l'ajustement du modèle. Nous utilisons alors cette propriété afin de proposer un ensemble de nouveaux classifieurs supervisés, que nous testons sur trois jeux de données classiques.

Mots-clés. Régression logistique, fonction de lien, invariance sous permutation.

Abstract. Several extensions of the logit model have been introduced in the binomial case, such as the probit model, whereas none has been proposed in the non-ordered multinomial case. We here introduce a new family of regression models for nominal response variable incorporating cumulative distribution functions other than logistic and we describe their estimation. To this end, probability of each category conditioned on the reference category is then highlighted within the link function. The choice of the reference category has now an impact on model's fit, unlike the logit multinomial model case. Using this property, we propose new supervised classifiers and test them on three benchmark datasets.

Keywords. Logistic regression, link function, invariance under permutation.

1 Extensions du modèle logit multinomial

Le modèle logit multinomial

Soient Y la variable réponse catégorielle à valeurs dans $\{1, 2, \ldots, J\}$ et X le vecteur de variables explicatives. La probabilité de chaque catégorie $j = 1, \ldots, J-1$ (J est par

convention la catégorie de référence) est modélisée par

$$P(Y=j) = \frac{\exp(\alpha_j + x^t\delta_j)}{1+\sum_{k=1}^{J-1}\exp(\alpha_k + x^t\delta_k)}. \tag{1}$$

Estimation On estime généralement ce modèle grâce à l'algorithme des scores de Fisher, dont l'itération à l'étape k s'écrit :

$$\beta^{[k+1]} = \beta^{[k]} - \left\{ \mathrm{E}\left(\frac{\partial^2 l}{\partial\beta^t\partial\beta}\right)_{\beta=\beta^{[k]}} \right\}^{-1} \left(\frac{\partial l}{\partial\beta}\right)_{\beta=\beta^{[k]}},$$

où $\beta^t = (\alpha_1, \ldots, \alpha_{J-1}, \delta_1^t, \ldots, \delta_{J-1}^t)$ est le vecteur de paramètres et l la log-vraisemblance. Le modèle logit multinomial étant vu comme un GLM, le score (d'une seule observation (y,x) pour simplifier) se décompose comme suit (McCullagh et Nelder, 1989)

$$\frac{\partial l}{\partial\beta} = \frac{\partial\eta}{\partial\beta}\,\frac{\partial\pi}{\partial\eta}\,\frac{\partial\theta}{\partial\pi}\,\frac{\partial l}{\partial\theta}, \tag{2}$$

où η est le prédicteur linéaire, $\pi = \mathrm{E}(Y|x)$ et θ le paramètre naturel de la famille exponentielle. De plus ce modèle correspond au lien canonique ($\theta = \eta$), donc le calcul se simplifie

$$\frac{\partial l}{\partial\beta} = \frac{\partial\eta}{\partial\beta}\,\frac{\partial l}{\partial\eta} = Z^t\,[y-\pi]. \tag{3}$$

avec Z la matrice de design

$$Z = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix}.$$

Les modèles références

Le modèle logit binomial a été étendu en modifiant la fonction de répartition F. Nous proposons d'étendre de la même manière le modèle logit multinomial en remarquant que (1) est équivalent à

$$P(Y=j|\ Y\in\{j,J\}) = \frac{\exp(\alpha_j + x^t\delta_j)}{1+\exp(\alpha_j + x^t\delta_j)},$$

pour $j = 1, \ldots, J-1$. On reconnaît la fonction de répartition logistique dans la partie droite de l'équation. Nous proposons alors de la remplacer par n'importe quelle fonction de répartition F strictement croissante sur $\mathbb{R}$. Nous obtenons alors la forme plus générale:

$$P(Y=j|\ Y\in\{j,J\}) = F(\alpha_j + x^t\delta_j),$$

pour $j = 1, \ldots, J-1$. Puisque la probabilité d'une catégorie j est conditionnée par rapport à elle même et à la catégorie de référence J, nous appelons ces modèles les modèles références. Comme l'a remarqué Tutz (1991), deux fonctions de répartition liées linéairement génèrent le même modèle. Pour la loi normale, par exemple, il y a donc un seul représentant ($\mathcal{N}(0,1)$ par exemple) tandis que pour la loi de Student chaque degré de liberté $d \in \mathbb{N}^*$ génère un modèle différent.

Estimation Pour un GLM multinomial quelconque, la décomposition (2) du score devient

$$\frac{\partial l}{\partial \beta} = Z^t \, \frac{\partial \pi}{\partial \eta} \, \mathrm{Cov}(Y|x)^{-1} \, [y - \pi].$$

Pour chaque GLM, ce calcul se différencie par celui de la matrice Jacobienne $\partial\pi/\partial\eta$ (qui dépend de la fonction de lien). Pour les modèles références nous montrons que

$$\frac{\partial \pi_j}{\partial \eta_i} = \frac{f(\eta_i)}{F(\eta_i)[1 - F(\eta_i)]} \, \mathrm{Cov}(Y_i, Y_j),$$

où f est la densité associée à F (Peyhardi, 2013). Si l'on prend la fonction de répartition logistique on remarque que $f = F(1-F)$ et l'on retrouve bien le calcul du score dans le cas canonique (3).

2 Propriété d'invariance sous permutation

Pour le modèle logit multinomial, le choix de la catégorie de référence n'a pas d'impact sur l'ajustement du modèle. Plus généralement on a :

Propriété 1 *Le modèle logit multinomial est invariant sous toutes les permutations des catégories* $\{1, \ldots, J\}$.

En fait ce modèle ne tient pas compte de l'ordre sur les catégories. Pour les modèles références, la catégorie de référence joue un rôle particulier, et nous remarquons que :

Propriété 2 *Un modèle référence est invariant sous les permutations de* $\{1, \ldots, J\}$ *qui fixent la catégorie de référence* J.

Mais qu'en est-il si l'on change la catégorie de référence ? On peut montrer qu'un modèle référence n'est pas invariant si l'on transpose la catégorie de référence pour des fonctions de répartitions F analytiques telles que Gumbel min, Gumbel max et exponentielle (Peyhardi, 2013). Cela se complique pour les fonctions de répartitions F non-analytiques telles que la loi normale ou la loi de Cauchy. Nous introduisons alors la conjecture suivante :

Conjecture 1 *Un modèle référence autre que logistique est invariant* ***uniquement*** *sous les permutations de* $\{1, \ldots, J\}$ *qui fixent la catégorie de référence* J.

Nous proposons de s'en convaincre empiriquement avec le jeu de données de la littérature *boys disturbed dreams* qui comprend $J = 4$ catégories réponses et une variable explicative continue x. Les modèles références avec fonction de répartition normale (respectivement logistique, Laplace, Cauchy, Gumbel min et max) sont alors estimés pour chacune des $J! = 24$ permutations. Ils sont ensuite ordonnés selon la log-vraisemblance afin de mettre en évidence les équivalences (chaque plateau correspond à une invariance particulière).

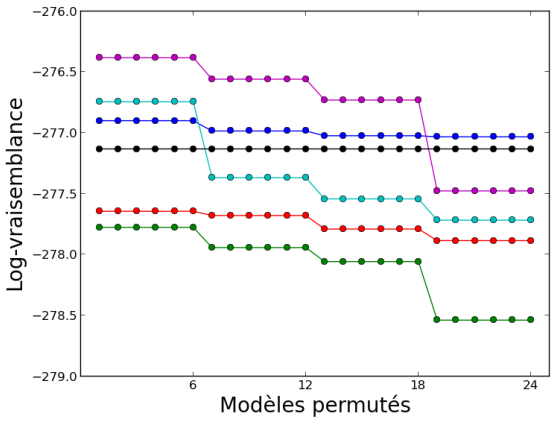

Figure 1: Log-vraisemblance des modèles références normal (bleu), logistique (noir), Laplace (rouge), Cauchy (vert), Gumbel min (magenta), Gumbel max (cyan) pour chacune des $J! = 24$ permutations.

L'unique plateau de la courbe noire confirme bien la propriété 1 tandis que les $J = 4$ plateaux pour les trois lois autres que logistique correspondent aux 4 choix de catégorie de référence possibles. Les résultats obtenus montrent que les modèles références autre que logistique ne sont pas invariants si l'on modifie la catégorie de référence.

3 Application en classification supervisée

La régression logistique donne souvent de meilleurs résultats que d'autres méthodes classiques en classification supervisée (Lim et al., 2000); comme l'analyse discriminante linéaire ou quadratique par exemple. Notre étude comparative, sur trois jeux de données classiques (disponibles sur UCI et KEEL), se fera donc uniquement par rapport à la régression logistique. Nous proposons de comparer, dans un premier temps, l'efficacité de 10 modèles références (correspondant aux dix lois : normal, Laplace, Gumbel min, Gumbel max, Student$_1$, ..., Student$_6$ et définissant l'ensemble de classifieurs noté $\mathfrak{C}^*$) à celle de la régression logistique. Pour cela nous comparons les taux d'erreur de classification sur les trois jeux de données en utilisant la validation croisée (avec 10% pour l'échantillon test). Pour chaque classifieur, le taux d'erreur est moyenné sur les dix sous-échantillons

d'apprentissage et comparé au taux d'erreur moyen obtenu avec la régression logistique (indiqué en bleu dans les figures 2, 3 et 4). Dans un deuxième temps nous modifions la catégorie de référence et obtenons ainsi un ensemble de $10 \times J$ classifieurs noté $\mathfrak{C}$ (avec $\mathfrak{C}^* \subset \mathfrak{C}$). Le meilleur taux moyen d'erreur obtenu pour les classifieurs de $\mathfrak{C}^*$ (resp. $\mathfrak{C}$) est indiqué en vert (resp. rouge).

Thyroïde L'objectif de l'étude est de détecter si un patient n'a pas de problème de thyroïde (1), ou bien s'il souffre d'hyperthyroïdie (2) ou d'hypothyroïdie (3). Ce jeu de données contient $n = 7200$ individus et 21 variables explicatives continues.

Véhicule L'objectif de l'étude est de classer une silhouette donnée comme étant un certain type de véhicule : bus (1), opel (2), saab (3) ou van (4). Le véhicule peut être vu sous différents angles. Ce jeu de données contient $n = 846$ individus et 18 variables explicatives continues.

Blocs de pages L'objectif de l'étude est de classer tous les blocs de pages détectés par segmentation d'un document. C'est une étape essentielle dans l'analyse de document pour séparer le texte des graphiques. Les cinq classes sont : texte (1), ligne horizontale (2), image (3), ligne verticale (4) et graphique (5). Les $n = 5473$ blocs sont extraits de 54 documents. Les variables explicatives sont continues.

Au vu des résultats, la loi Gumbel min donne le plus mauvais classifieur (problème de convergence). Les lois normal, Laplace, Gumbel max sont comparables à la loi logistique. Enfin les lois de Student donnent de meilleurs résultats, certainement grâce à leurs queues de distribution plus lourdes. De plus le gain dû au changement de la fonction de répartition F est plus important que celui dû au changement de la catégorie de référence.

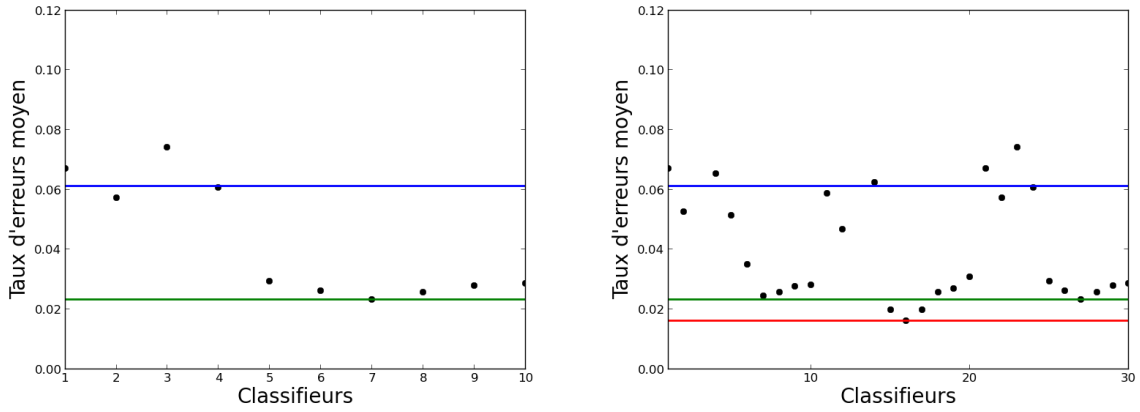

Figure 2: Taux d'erreur des classifieurs de $\mathfrak{C}^*$ et $\mathfrak{C}$ sur le jeu de données *thyroïde*.

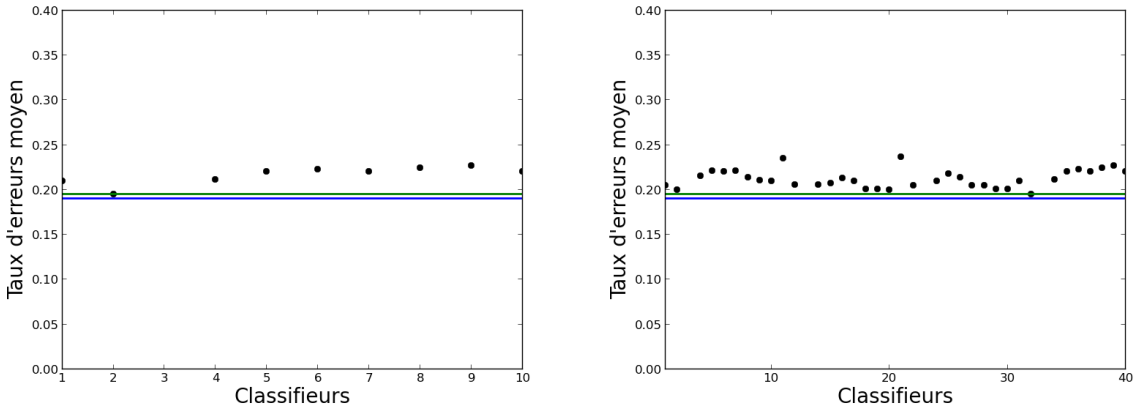

Figure 3: Taux d'erreur des classifieurs de $\mathfrak{C}^*$ et $\mathfrak{C}$ sur le jeu de données *véhicule*.

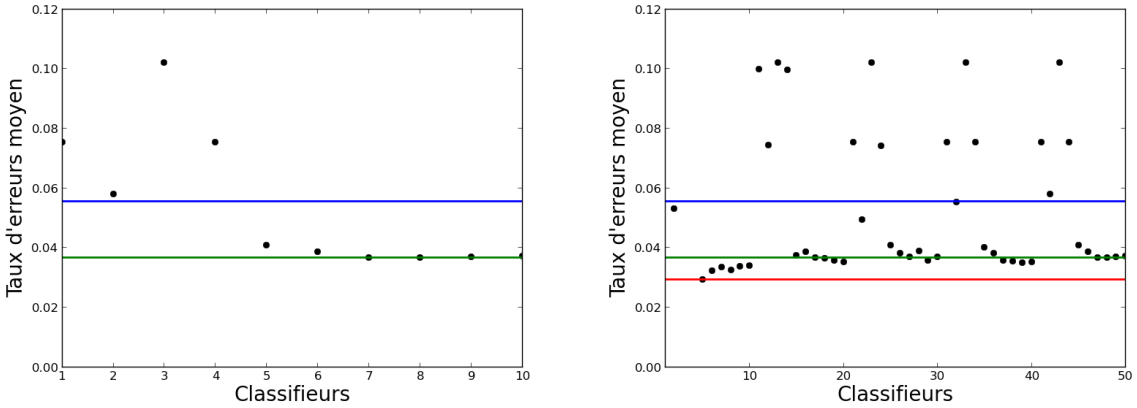

Figure 4: Taux d'erreur des classifieurs de $\mathfrak{C}^*$ et $\mathfrak{C}$ sur le jeu de données *blocs de pages*.

Bibliographie

[1] Lim, T-S., Loh, W-Y. et Shih, Y-S. (2000), *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*, Machine learning, 40(3), 203–228.

[2] McCullagh, P., et Nelder, J. A. (1989), *Generalized linear models*, 37, CRC press.

[3] Peyhardi, J. (2013), *A new GLM framework for analysing categorical data; application to plant structure and development.*, thèse de doctorat.

[4] Tutz, G. (1991), *Sequential models in categorical regression*, Computational Statistics & Data Analysis, 11(3):275–295.