Une pénalité de groupe pour des données multivoie de grande dimension

Laurent Le Brusquet [1], Arthur Tenenhaus [1,2], Gisela Lechuga [1], Vincent Perlbarg [2], Louis Puybasset [3] & Damien Galanaud [4].

[1] *Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506) CentraleSupelec - CNRS - Université Paris-Sud 3, rue Joliot Curie 91192, Gif-sur-Yvette , prenom.nom@centralesupelec.fr*

[2] *Bioinformatics/Biostatistics Platform IHU-A-ICM, Brain and Spine Institute 47-83, Bd de l'hôpital; Paris, France, vperlbar@imed.jussieu.fr*

[3] *AP-HP, Pitié-Salpêtrière Hospital, Surgical Neuro-Intensive Care Unit, Paris, France*

[4] *AP-HP, Pitié-Salpêtrière Hospital, Department of Neuroradiology, Paris, France*

Résumé. Le problème de la classification supervisée de données multivoie de grande dimension avec un a priori de structure de groupes sur les variables est étudié. Plus précisément une pénalité adaptée à cette structure de données est proposée. Sans surcoût calculatoire notable, cette pénalité favorise l'interprétabilité des modèles obtenus. La pénalité est ici développée pour l'analyse discriminante et la régression logistique. Une application à l'analyse de données de neuroimagerie multimodale est présentée.

Mots-clés. Pénalité structurée, analyse multivoie, analyse discriminante, régression logistique.

Abstract. Supervised classification of multiway data with group structure at the level of the variables is considered in this paper. More specifically, a penalty that fits the natural structure of the data is presented. This penalty promotes a better interpretablity of the resulting model. The proposed penalty is plugged to discriminant analysis and logistic regression and tested on a real multimodal neuroimaging dataset.

Keywords. Structured penality, multiway analysis, Fisher discriminant analysis, logistic regression

1 Introduction

Ce papier s'inscrit dans le contexte de classification supervisée où les variables explicatives sont organisées en groupes de variables observées suivant plusieurs modalités. Ce type de structure se retrouve par exemple dans le cas de données spatio-temporelles où chaque modalité correspond à un instant d'observation et où les groupes correspondent à une partition de l'espace (par exemple le découpage d'un territoire en différentes régions).
De cette structure peut découler un nombre total de variables important, et donc un potentiel problème de sur-apprentissage, Pour éviter ce type de problème, il est usuel de pénaliser les modèles. Ce papier présente l'utilisation d'une pénalité quadratique qui tient compte à la fois de la structure multivoie et de la structure par groupe des variables. La figure 1 illustre la structure des données : $\{\mathbf{X}_{ijk}\}_{1\leq i\leq n, 1\leq j\leq J, 1\leq k\leq K}$ est un tenseur d'ordre 3 avec n le nombre d'individus, J le nombre de variables (spatiales lorsque les groupes représentent des régions) et K le nombre de modalités. L'ensemble des G groupes forment une partition de l'ensemble des J variables. Pour $1 \leq g \leq G$, on note J_g l'ensemble des variables du groupe g.

Le papier est décomposé comme suit : la section 2 présente une application de neuroimagerie multimodale permettant de motiver la pénalité présentée. La section 3 rappelle brièvement l'analyse discriminante multivoie (Lechuga et al (2014)) et la régression logistique multivoie (Le Brusquet et al (2014)) et la section 4 discute leur version avec pénalité structurante. La section 5 présente les résultats obtenus sur les données de neuroimagerie multimodale.

2 Les données COMA

Les méthodes proposées ont été testées sur des données de neuroimagerie récoltées sur des personnes victimes d'un coma et pour lesquelles on souhaite prédire leur capacité de récupération à partir d'images IRM multimodales. Pour chaque individu, $K = 4$ modalités d'imagerie ont été enregistrées (images de $91\times$ 109 $\times$ 91 voxels). Les voxels situés dans la matière blanche, principale siège des phénomènes explicatifs (Lechuga et al (2014)), ont été sélectionnés et constituent un ensemble de $J = 20764$ variables. Au sein de la matière blanche, les médecins ont identifié $G = 17$ régions (voir figure 2). Cette partition de l'espace va être utilisée comme a priori pour construire la pénalité. Enfin, les données ont été collectées pour $n = 143$ individus.

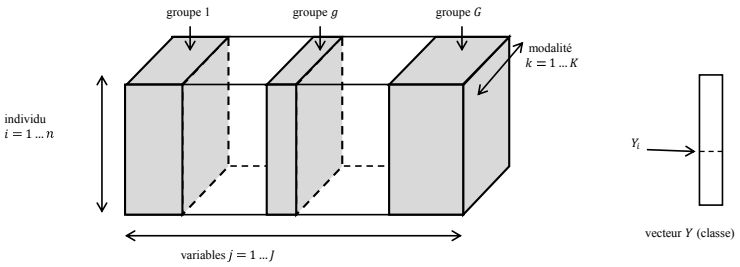

Figure 1: Tenseur $\mathbf{X}$ des variables explicatives et vecteur $\mathbf{Y}$ de la variable à prédire. Les données sont multivoie et les variables sont structurées en groupes disjoints : chaque individu est représenté par G groupes de variables observées selon K modalités.

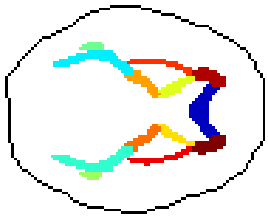

Figure 2: Coupe frontale de la matière blanche et partition en 17 régions distinctes. Les régions sont repérées par des couleurs différentes. Le contour de la boite crânienne est repéré par le trait noir.

3 Classifieurs pour données multivoie

Dans (Lechuga et al (2014) et Le Brusquet et al (2014)), l'analyse discriminante et la régression logistique binaire multiple ont été étendues aux données mutivoie. Cette section rappelle brièvement ces deux méthodes.

Soit $\mathbf{X}^{\mathrm{u}}$ ($^{\mathrm{u}}$ pour unfolded) la matrice de taille $n \times (JK)$ obtenue en "dépliant" le tenseur, c'est-à-dire en concaténant les K matrices $\mathbf{X}_{..k}$. La matrice $\mathbf{X}^{\mathrm{u}}$ est ainsi composée de n vecteurs $\mathbf{x}_i = \mathrm{vec}(\mathbf{X}_{i..})$ de longueur JK.

Analyse discriminante. L'analyse factorielle discriminante consiste à rechercher des projections de la forme $g(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$. Le vecteur de poids $\boldsymbol{\beta}$ est choisi de sorte à maximiser le rapport variance interclasse / variance intraclasse. Ce rapport de variance s'écrit (voir Hastie, Tibshirani et Friedman (2009)):

$$R(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^\top \left(\mathbf{X}^{\mathrm{u}}\right)^\top \mathbf{M}_{\mathrm{Between}} \mathbf{X}^{\mathrm{u}} \boldsymbol{\beta}}{\boldsymbol{\beta}^\top \left(\mathbf{X}^{\mathrm{u}}\right)^\top \mathbf{M}_{\mathrm{Within}} \mathbf{X}^{\mathrm{u}} \boldsymbol{\beta} + \mu \boldsymbol{\beta}^\top \boldsymbol{\beta}} \tag{1}$$

$\mathbf{M}_{\mathrm{Between}}$ et $\mathbf{M}_{\mathrm{Within}}$ sont des matrices $n \times n$ semi-définies positive ne dépendant que du vecteur $\mathbf{Y}$. L'analyse discriminante régularisée fait intervenir le terme $\mu \boldsymbol{\beta}^\top \boldsymbol{\beta}$ afin de palier les problèmes numériques et contrer le phénomène de sur-apprentissage.

Régression logistique. La régression logistique s'appuie sur la maximisation de la log-vraisemblance conditionnelle $\sum_{i=1\ldots n} \log \mathbb{P}(y_i/\mathbf{x}_i)$. $\mathbb{P}(y_i/\mathbf{x}_i)$ est modélisée en supposant

affine son log-ratio, c'est-à-dire $\log \dfrac{\mathbb{P}(y=1/\mathbf{x})}{1-\mathbb{P}(y=1/\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$. Après réécriture de la log-vraisemblance, le problème revient à choisir les paramètres β_0 et $\boldsymbol{\beta}$ qui maximisent :

$$\mathcal{C}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - \log\left(1 + \exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right) - \mu\left(\beta_0^2 + \boldsymbol{\beta}^\top \boldsymbol{\beta}\right) \qquad (2)$$

Comme pour l'analyse discriminante, le terme $\mu\beta_0^2 + \mu\boldsymbol{\beta}^\top\boldsymbol{\beta}$ sert à pénaliser la solution.

Extension aux données multivoie. La version multivoie des deux précédentes méthodes consiste à maximiser les deux critères (1) ou (2) sous la contrainte que le vecteur $\boldsymbol{\beta}$ soit de la forme $\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$ où $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ sont deux vecteurs de taille K et J pondérant l'influence des modalités et des variables. Cette contrainte permet de réduire à $K + J$ les KJ degrés de liberté initiaux du vecteur $\boldsymbol{\beta}$ cherché tout en lui imposant une structure cohérente avec la structuration tensorielle des données. Un algorithme de type directions alternées permet d'optimiser les critères (1) et (2) par rapport à $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$.

4 Pénalité proposée

La pénalité proposée vise à modifier les termes de régularisation de la forme $\mu\boldsymbol{\beta}^\top\boldsymbol{\beta}$ pour tenir compte d'une part de la structure tensorielle et d'autre part le la structure de groupes de $\mathbf{X}$. S'imposant de ne pas dégrader le temps calculatoire (il faut garder à l'esprit que le vecteur $\boldsymbol{\beta}$ est de longueur KJ donc potentiellement grand), nous nous restreignons ici à des pénalités quadratiques, donc de la forme $\boldsymbol{\beta}^\top\mathbf{R}\boldsymbol{\beta}$. Le but cherché étant un gain en interprétabilité, la pénalité a été conçue de manière à : (i) Séparer l'influence des variables de l'influence des modalités et (ii) Homogénéiser les poids associés à des variables d'un même groupe, sans contraindre les variations entre groupes.

Séparer l'influence des variables de l'influence des modalités. Cet objectif a été atteint en imposant à $\mathbf{R}$ une structure tensorielle de la forme $\mathbf{R} = \mathbf{R}^K \otimes \mathbf{R}^J$. L'a priori imposé via la régularisation est ainsi cohérent avec la structure tensorielle des données et présente de plus l'avantage de ne pas alourdir le coût calculatoire puisque $\boldsymbol{\beta}^\top\mathbf{R}\boldsymbol{\beta} = (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J)^\top\mathbf{R}(\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J) = \left(\left(\boldsymbol{\beta}^K\right)^\top \mathbf{R}^K\boldsymbol{\beta}^K\right)\left(\left(\boldsymbol{\beta}^J\right)^\top \mathbf{R}^J\boldsymbol{\beta}^J\right)$.

Homogénéiser les poids associés à des variables d'un même groupe, sans contraindre les variations entre groupes. Le choix de $\mathbf{R}^J$ permet de gérer l'homogénéité par région. De nombreuses pénalités de groupe existent dont les plus récentes sont fondées sur une norme $\ell_{1/2}$ afin d'introduire de la parcimonie (voir par exemple Bach et al (2012)). Nous avons préféré conserver une pénalité quadratique en raison du grand nombre de variables (autant que de voxels dans notre application). Nous avons choisi $\mathbf{R}^J$ tel que :

$$\left(\boldsymbol{\beta}^J\right)^\top \mathbf{R}^J \boldsymbol{\beta}^J = \sum_{g=1}^{G} \sum_{v,v' \in J_g, v \text{ et } v' \text{ voisins}} \left(\boldsymbol{\beta}_v^J - \boldsymbol{\beta}_{v'}^J\right)^2 .$$

Cette pénalité a tendance à rendre homogènes les poids associés aux variables d'un même groupe puisqu'elle pénalise les variations à l'intérieur d'un groupe sans pénaliser les écarts entre voxels de deux groupes distincts. L'interprétabilité en est ainsi facilitée. À noter que la matrice $\mathbf{R}^J$ proposée a une structure bloc-diagonale (2 voxels de 2 groupes différents ne sont pas pénalisés pour leur écart entre poids) et que chaque bloc est lui même creux (2 voxels éloignés d'un même groupe ne sont pas pénalisés). Cela permet une manipulation aisée de la matrice $\mathbf{R}^J$ (définition, stockage, coût calculatoire de $\left(\boldsymbol{\beta}^J\right)^\top \mathbf{R}^J \boldsymbol{\beta}^J$).
D'autres choix de $\mathbf{R}^J$ sont possibles pour limiter les écarts entre variables issues d'un même groupe et proches géographiquement l'une de l'autre. On pourrait par exemple modéliser les variations entre variables par un processus gaussien et choisir une fonction de covariance traduisant la régularité entre variables proches l'une de l'autre. Avec cette modélisation, $\mathbf{R}^J$ aurait alors été l'inverse de la matrice de covariance du processus.

5 Application aux données COMA

Seuls les résultats obtenus avec l'analyse discriminante sont présentés. Les résultats de l'analyse discriminante régularisée appliquée à $\mathbf{X}^{\mathrm{u}}$ sont comparés à ceux obtenus par analyse discriminante multivoie avec et sans pénalité de groupes. La figure 3 montre que la version standard de l'analyse discriminante conduit à des poids $\boldsymbol{\beta}$ difficilement interprétables en raison des fortes variations entre poids pour des voxels au sein d'une même région.

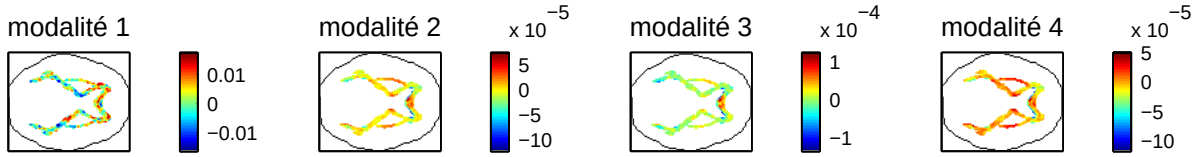

Figure 3: Analyse discriminante standard + pénalité standard : poids $\boldsymbol{\beta}$.

La version multivoie avec une pénalité constuite avec $\mathbf{R} = \mathbf{I}$ (résultats de la figure 4) permet de séparer l'influence des voxels de celle des 4 modalités. L'identification des régions de la matière blanche les plus discriminantes reste difficile en raison de la forte hétérogénéité au sein d'une même région. Les résultats de la figure 5 montre que la pénalité de groupe proposée permet de contourner cette difficulté : on reconnait le découpage en régions de la figure 2. De plus, les poids au sein d'une même région sont homogènes.

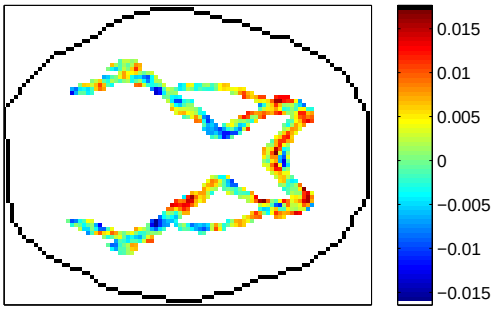

Figure 4: Analyse discriminante multivoie + pénalité standard : poids $\boldsymbol{\beta}^J$.

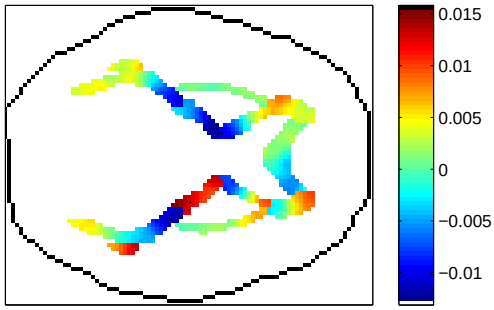

Figure 5: Analyse discriminante multivoie + pénalité proposée : poids $\boldsymbol{\beta}^J$.

Remarque. Les hyperparamètres μ ont été ajustés pour chacune des trois méthodes envisagées de sorte à fournir des taux de classification similaires.

6 Conclusion

En grande dimension, il apparaît indispensable de forcer l'interprétabilité des modèles obtenus en imposant des contraintes de régularité. Ce papier présente une pénalité volontairement simple pour qu'elle soit facilement utilisable sur des données de grande dimension.
Les cas de l'analyse discriminante et de la régression logistique sont ici détaillés mais la pénalité envisagée pourrait être très facilement adaptée à d'autres méthodes.
Le nombre de groupes impliqués dans la prédiction pouvant être limité, une extension parcimonieuse est envisagée comme perspective afin de pouvoir sélectionner les groupes de variables les plus discriminants.

Bibliographie

[1] Lechuga G., Le Brusquet L., Perlbarg V., Puybasset L., Galanaud D. et Tenenhaus A. (2014), Discriminant Analysis for Multi-way Data, *PLS'2014*.
[2] Le Brusquet L., Lechuga G., Tenenhaus A. (2014), Régression Logistique Multivoie, *46eme Journée de Statistique*.
[3] Bro, R. (2000), Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications *ICSLP Proceedings*.
[4] Hastie, T., Tibshirani, R. and Friedman, J. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.
[5] Bach F., Jenatton R., Mairal J., Obozinski G. (2012) Structured sparsity through convex optimization. Statistical Science, 27(4), 450–468.