Analyse discriminante multivoie sparse

Laurent Le Brusquet, Arthur Tenenhaus & Gisela Lechuga

Laboratoire des Signaux et Systèmes, CentraleSupelec - CNRS - Univ. Paris-Sud,
Université Paris-Saclay, 3 rue Joliot Curie 91192, Gif-sur-Yvette,
prenom.nom@centralesupelec.fr

Résumé. Une version « sparse » de l'analyse discriminante multivoie est ici présentée. À la manière des modèles de type group LASSO, les variables sont sélectionnées (ou non) par paquets ; ceci offre la possibilité d'injecter de l'a priori et de faciliter l'interprétation du classifieur obtenu. Par ailleurs, cette méthode présente l'avantage d'être peu gourmande en temps de calcul.

Mots-clés. Données multivoie, parcimonie, analyse discriminante.

Abstract. A sparse version of Fisher discriminant analysis for multiway data is presented. As for group LASSO, This method allows selecting (or not) jointly a set of variables. It gives the possibility to inject prior knowledge and yields model easier to interpret. In addition, this method is computationaly efficicient.

Keywords. Multiway data analysis, sparsity, Fisher discriminant analysis.

1 Introduction

L'intérêt pour les méthodes d'analyse statistique des données multivoie est croissant depuis quelques années. Cet engouement est amplifié par la nécessité de traiter des données volumineuses et structurées. En règle générale, une contrainte sur la structure du vecteur des paramètres cherchés est imposé afin de tenir explicitement compte de la structure tensorielle des données. Ce type de contrainte présente l'avantage de diminuer la taille du vecteur des paramètres à estimer, permettant une estimation en un temps raisonnable et une interprétation facilitée par le nombre restreint de paramètres.

Par ailleurs, l'utilisation de pénalité ℓ_1 permet d'injecter de la parcimonie à dessein d'une interprétation facilitée.

Ce papier propose de combiner ces 2 techniques en proposant une version sparse de l'anayse discriminante multivoie. L'analyse proposée est particulièrement adaptée aux données multivoie pour lesquelles on souhaite un classifieur facile à interpréter puisque l'interprétation des différents axes se fait séparément (intérêt de la modélisation multivoie), et que pour chaque axe, seulement une partie des variables intervient (intérêt de la pénalité ℓ_1). La section 2 résume les différentes versions de l'analyse discriminante à l'origine de ce travail. La version sparse de l'analyse discriminante mutivoie est présentée section 3 : le critère utilisé ainsi que la stratégie développée pour minimiser ce critère y sont présentés. Enfin, la méthode est testée sur un exemple simulé.

2 Différentes versions de l'analyse discriminante

En analyse discriminante multivoie, les données ne sont pas représentées par une matrice, comme c'est le cas en analyse standard, mais par un tenseur : les variables explicatives sont ainsi observées selon plusieurs modalités. Afin d'alléger les explications, le papier se concentre sur les tenseurs d'ordre 3 bien que la méthode proposée puisse s'appliquer aux tenseurs d'ordre quelconque.

Soit $\{\underline{\mathbf{X}}_{ijk}\}_{1\leq i\leq I, 1\leq j\leq J, 1\leq k\leq K}$ un tenseur d'ordre 3 de dimension $I\times J\times K$ où I désigne le nombre d'individus, J le nombre de variables et K le nombre de modalités. Soit $\mathbf{X}$ la matrice de taille $I\times JK$ où chaque ligne $\mathbf{x}_i = \text{vec}\,(\underline{\mathbf{X}}_{i..})^\top$. Soit $\mathbf{y}$ le vecteur de longueur I contenant la classe de chaque individu.

Analyse discriminante. L'analyse factorielle discriminante consiste à rechercher des projections de la forme $g(\mathbf{x}) = \boldsymbol{\beta}^\top\mathbf{x}$. Les vecteurs de poids $\boldsymbol{\beta}$ sont choisis de sorte à maximiser le rapport variance interclasse / variance intraclasse. Ce rapport de variance s'écrit (voir Hastie et al (2009)) :

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

$\mathbf{M}_{\text{Between}}$ et $\mathbf{M}_{\text{Within}}$ sont des matrices $I\times I$ semi-définies positives ne dépendant que du vecteur $\mathbf{y}$. L'analyse discriminante régularisée fait intervenir le terme $\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}$ afin de pallier les problèmes numériques et contrer le phénomène de sur-apprentissage.

Analyse discriminante mutivoie (Multiway-FDA). Elle consiste à optimiser le critère (1) en imposant une structure de Kronecker au vecteur $\boldsymbol{\beta}$ cherché : $\boldsymbol{\beta} = \boldsymbol{\beta}^K\otimes\boldsymbol{\beta}^J$. Ainsi, au lieu de rechercher un poids $\boldsymbol{\beta}_{j,k}$ pondérant l'influence de la variable j pour la modalité k, on se restreint à une analyse séparée de l'influence de la variable j et de la modalité k. Les vecteurs $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ sont obtenus par l'Algorithme 1, de type directions alternées. Le lecteur est renvoyé à Lechuga et al (2015) pour plus de détails sur Multiway-FDA.

Analyse discriminante sparse (sparse-FDA). Soit $\mathbf{Y}$ le tableau disjonctif complet associé à $\mathbf{y}$ ($\mathbf{Y}_{i,c} = 1$ si l'individu i est de la classe c). Hastie et al (2009) ont montré que le critère de l'analyse discriminante, régularisée ou non, pouvait également s'écrire sous la forme d'un problème de régression. Supposons que $s-1$ vecteurs $\boldsymbol{\beta}_r$ aient déjà été calculés. Le $s^{\text{ième}}$ vecteur $\boldsymbol{\beta}_s$ est défini par :

$$\min_{\boldsymbol{\beta}_s,\boldsymbol{\theta}_s}\left\{\|\mathbf{Y}\boldsymbol{\theta}_s - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 + \lambda\|\boldsymbol{\beta}_s\|_2^2\right\} \quad \text{s.c} \quad \frac{1}{n}\boldsymbol{\theta}_r^\top\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\theta}_s = \delta_{rs}, \quad r\leq s \tag{2}$$

où $\boldsymbol{\theta}_s$ est un vecteur de longueur C (nombre de classes).

Algorithm 1 Calcul de l'axe principal d'une analyse Multiway-FDA

Require: $\epsilon > 0, \quad \boldsymbol{\beta}^{K(0)}, \quad \underline{\mathbf{X}}, \quad \mathbf{y}, \quad \lambda$

$q \leftarrow 0$

repeat

- $\mathbf{X}_K = \sum_{k=1}^{K} \boldsymbol{\beta}_k^{K(q)} \underline{\mathbf{X}}_{..k}, \qquad \lambda^K = \lambda \|\boldsymbol{\beta}^{K(q)}\|_2^2$

$$\boldsymbol{\beta}^{J(q+1)} \leftarrow \underset{\boldsymbol{\beta}^J, \|\boldsymbol{\beta}^J\|=1}{\operatorname{argmax}} \frac{(\boldsymbol{\beta}^J)^\top \mathbf{X}_K^\top \mathbf{M}_{\text{Between}} \mathbf{X}_K \boldsymbol{\beta}^J}{(\boldsymbol{\beta}^J)^\top \mathbf{X}_K^\top \mathbf{M}_{\text{Within}} \mathbf{X}_K \boldsymbol{\beta}^J + \lambda^K \|\boldsymbol{\beta}^J\|_2^2}$$

- $\mathbf{X}_J = \sum_{j=1}^{J} \boldsymbol{\beta}_j^{J(q+1)} \underline{\mathbf{X}}_{.j.}, \qquad \lambda^J = \lambda \|\boldsymbol{\beta}^{J(q+1)}\|_2^2$

$$\boldsymbol{\beta}^{K(q+1)} \leftarrow \underset{\boldsymbol{\beta}^K, \|\boldsymbol{\beta}^K\|=1}{\operatorname{argmax}} \frac{(\boldsymbol{\beta}^K)^\top \mathbf{X}_J^\top \mathbf{M}_{\text{Between}} \mathbf{X}_J \boldsymbol{\beta}^K}{(\boldsymbol{\beta}^K)^\top \mathbf{X}_J^\top \mathbf{M}_{\text{Within}} \mathbf{X}_J \mathbf{w}_K + \lambda^J \|\boldsymbol{\beta}^K\|_2^2}$$

- $q \leftarrow q + 1$

until $\|\boldsymbol{\beta}^{K(q-1)} - \boldsymbol{\beta}^{K(q)}\| < \epsilon$

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

L'optimisation du critère s'effectue à l'aide d'un algorithme de directions alternées. Les étapes élémentaires sont ici très simples puisque, que ce soit pour l'optimisation par rapport à $\boldsymbol{\beta}_s$ ou par rapport à $\boldsymbol{\theta}_s$, les optima ont des expressions analytiques. On aboutit ainsi à l'Algorithme 2.

Clemmensen et al (2011) propose l'ajout d'une pénalité ℓ_1 au critère précédent afin d'obtenir un vecteur $\boldsymbol{\beta}$ parcimonieux :

$$\min_{\boldsymbol{\beta}_s, \boldsymbol{\theta}_s} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 + \lambda \|\boldsymbol{\beta}_s\|_2^2 + \lambda_1 \|\boldsymbol{\beta}_s\|_1 \right\} \quad \text{s.c} \quad \frac{1}{n} \boldsymbol{\theta}_r^\top \mathbf{Y}^\top \mathbf{Y} \boldsymbol{\theta}_s = \delta_{rs}, \quad r \leq s \tag{3}$$

L'optimisation par rapport à $\boldsymbol{\beta}_s$ dans l'algorithme 2 se fait par un algorithme de type elastic-net, largement étudié dans la littérature.

3 Méthode proposée : Sparse Multiway-FDA

Cette méthode consiste à reprendre la version multivoie de l'analyse discriminante en formulant les étapes d'analyse discriminante comme des problèmes de régression et en ajoutant au critère une pénalité ℓ_1 :

$$\min_{\boldsymbol{\beta}_s, \boldsymbol{\theta}_s} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 + \lambda \|\boldsymbol{\beta}_s\|_2^2 + \lambda_1 \mathrm{P}(\boldsymbol{\beta}_s) \right\} \quad \text{s.c} \begin{cases} \frac{1}{n} \boldsymbol{\theta}_r^\top \mathbf{Y}^\top \mathbf{Y} \boldsymbol{\theta}_s = \delta_{rs}, \quad r \leq s \\ \boldsymbol{\beta}_s = \boldsymbol{\beta}_s^K \otimes \boldsymbol{\beta}_s^J \end{cases} \tag{4}$$

Algorithm 2 Analyse discriminante exprimée comme un problème de régression

Require: $\epsilon > 0, \quad \boldsymbol{\beta_s}^{(0)}, \quad \mathbf{X}, \quad \mathbf{y}, \quad \lambda$

$q \leftarrow 0$

repeat

$\boldsymbol{\theta}_s^{(q)} \leftarrow \arg\min_{\boldsymbol{\theta_s}} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s - \mathbf{X}\boldsymbol{\beta}_s^{(q)}\|_2^2 \right\} \quad \text{s.c} \quad \frac{1}{n}\boldsymbol{\theta}_r^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_s = \delta_{rs}, \quad r \leq s$

$\boldsymbol{\beta}_s^{(q+1)} \leftarrow \arg\min_{\boldsymbol{\beta}_s} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s^{(q)} - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 + \lambda\|\boldsymbol{\beta}_s\|_2^2 \right\}$

$q \leftarrow q + 1$

until $\|\boldsymbol{\beta}_s^{(q)} - \boldsymbol{\beta}_s^{(q-1)}\| < \epsilon$

return $\boldsymbol{\beta}_s^{(q)}$

Deux pénalités sont ici proposées :

1. $\mathrm{P}(\boldsymbol{\beta}_s) = \|\boldsymbol{\beta}_s\|_1 = \|\boldsymbol{\beta}_s^K\|_1 \|\boldsymbol{\beta}_s^J\|_1$. Il s'agit de la transposition immédiate de l'équation (3).
2. $\mathrm{P}(\boldsymbol{\beta}_s) = \alpha\|\boldsymbol{\beta}_s^K\|_1 + (1-\alpha)\|\boldsymbol{\beta}_s^J\|_1$. Cette contrainte permet de forcer la parcimonie sur un axe plutôt que sur un autre. Pour les cas extrêmes ($\alpha = 0$ ou $\alpha = 1$), la sparsité n'est imposée que sur l'un des deux axes. Cette stratégie est à rapprocher des pénalités de type group LASSO (sans recouvrement) pour lesquelles tout un ensemble de variables est sélectionné ou non.

La convergence de l'algorithme peut être accélérée en ne faisant qu'une itération dans l'Algorithme 1. On obtient ainsi l'Algorithme 3 présenté pour la pénalité $\mathrm{P}(\boldsymbol{\beta}_s) = \|\boldsymbol{\beta}_s\|_1$.

Exemple illustratif. L'algorithme proposé a été appliqué à des données simulées : $K = 7$ spectres calculés pour $J = 750$ longueurs d'ondes ont été simulés pour $I = 26$ individus . Les 7 modalités obtenues correspondent à 7 profondeurs différentes. Les 26 individus sont répartis en 2 classes. La Figure 1 donne un exemple de quelques spectres obtenus à un même instant pour deux individus de classes différentes.
Sparse Multiway FDA a été comparée à (i) la version sparse de l'analyse discriminante (sparse-FDA), (ii) la version sparse de l'analyse discriminante avec une pénalité de type group LASSO, chaque spectre constituant un groupe de variables.

Tous les algorithmes testés nécessitent l'optimisation de critères de type elastic-net, avec pour (ii) la contrainte supplémentaire de constituer des groupes de variables. Pour cela, les scripts fournis par Boyd et al (2011) ont été utilisés. Sparse Multiway-FDA a été appliqué avec la pénalité $\mathrm{P}(\boldsymbol{\beta}_s) = \|\boldsymbol{\beta}_s\|_1$. Les poids $\boldsymbol{\beta}^J$ et $\boldsymbol{\beta}^K$ sont donnés Figure 2 et Table 1 : l'interprétation séparée des vecteurs de poids permet une interprétation facile plus facile qu'avec sparse-FDA (voir Figure 3) ou la technique group LASSO (voir Figure 4). En outre, la Table 2 reporte les temps de calcul et montre la rapidité de l'algorithme proposé.

Algorithm 3 Sparse Multiway-FDA

Require: $\epsilon > 0, \quad {\boldsymbol{\beta}_{\boldsymbol{s}}}^{K(0)}, \quad {\boldsymbol{\beta}_{\boldsymbol{s}}}^{J(0)}, \quad \mathbf{X}, \quad \mathbf{y}, \quad \lambda$

$q \leftarrow 0$

repeat

$\boldsymbol{\beta}_s^{(q)} \leftarrow \boldsymbol{\beta}_s^{K(q)} \otimes \boldsymbol{\beta}_s^{J(q)}$

- $\boldsymbol{\theta}_s^{(q)} \leftarrow \arg\min_{\boldsymbol{\theta_s}} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s - \mathbf{X}\boldsymbol{\beta}_s^{(q)}\|_2^2 \right\} \quad \text{s.c} \quad \frac{1}{n}\boldsymbol{\theta}_r^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_s = \delta_{rs}, \quad r \leq s$
- $\mathbf{X}_K = \sum_{k=1}^{K} \boldsymbol{\beta}_k^{K(q)} \underline{\mathbf{X}}_{..k}, \qquad \lambda_K = \lambda \|\boldsymbol{\beta}^{K(q)}\|_2^2, \qquad \lambda_1^K = \lambda_1 \|\boldsymbol{\beta}^{K(q)}\|_1$

$\boldsymbol{\beta}_s^{J(q+1)} \leftarrow \arg\min_{\boldsymbol{\beta}_s^J} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s^{(q)} - \mathbf{X}_K\boldsymbol{\beta}_s^J\|_2^2 + \lambda_K \|\boldsymbol{\beta}_s^J\|_2^2 + \lambda_1^K \|\boldsymbol{\beta}_s^J\|_1 \right\}$

- $\mathbf{X}_J = \sum_{j=1}^{J} \boldsymbol{\beta}_j^{J(q+1)} \underline{\mathbf{X}}_{.j.}, \qquad \lambda_J = \lambda \|\boldsymbol{\beta}^{j(q+1)}\|_2^2, \qquad \lambda_1^J = \lambda_1 \|\boldsymbol{\beta}^{J(q+1)}\|_1$

$\boldsymbol{\beta}_s^{K(q+1)} \leftarrow \arg\min_{\boldsymbol{\beta}_s^K} \left\{ \|\mathbf{Y}\boldsymbol{\theta}_s^{(q)} - \mathbf{X}_K\boldsymbol{\beta}_s^K\|_2^2 + \lambda_J \|\boldsymbol{\beta}_s^K\|_2^2 + \lambda_1^J \|\boldsymbol{\beta}_s^K\|_1 \right\}$

- $q \leftarrow q + 1$

until $\|\boldsymbol{\beta}_s^{K(q)} - \boldsymbol{\beta}_s^{K(q-1)}\| < \epsilon$

return $\left(\boldsymbol{\beta}_s^{K(q)}, \quad \boldsymbol{\beta}_s^{J(q)}\right)$

	prof.1	prof. 2	prof. 3	prof. 4	prof. 5	prof. 6	prof. 7
$\boldsymbol{\beta}^K$	0	0	0	0	0.183	0.467	0.865

TABLE 1 – Sparse Multiway-FDA : vecteur $\boldsymbol{\beta}^K$ pondérant l'influence des profondeurs.

	taille optimisation ℓ_1	temps CPU (s)
sparse Multiway-FDA	J et K	1.20
sparse-FDA	$J \times K$	19.67
sparse-FDA group LASSO	$J \times K$	25.83

TABLE 2 – Comparaison des temps de calcul.

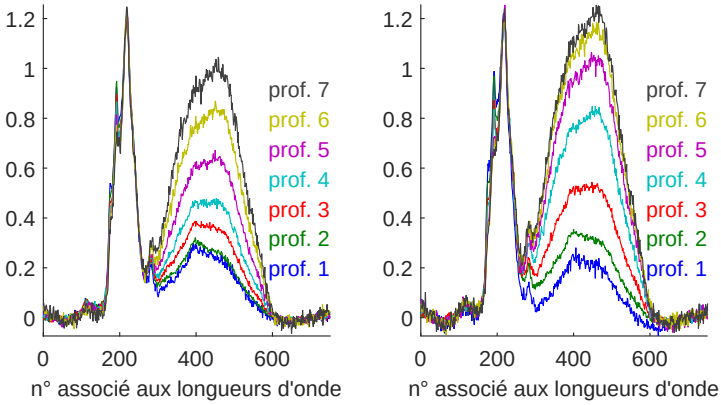

FIGURE 1 – Données simulées pour deux individus : pour chaque individu, 7 spectres ont été mesurés.

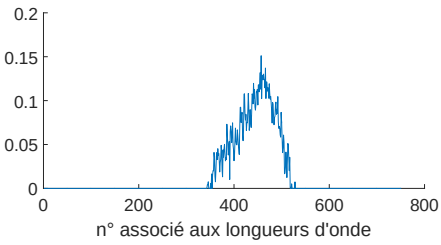

FIGURE 2 – Sparse MFDA : vecteur $\boldsymbol{\beta}^J$ obtenu.

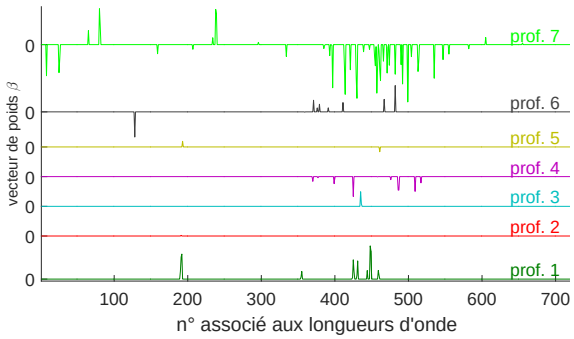

FIGURE 3 – Sparse-FDA : vecteur $\boldsymbol{\beta}$ obtenu.

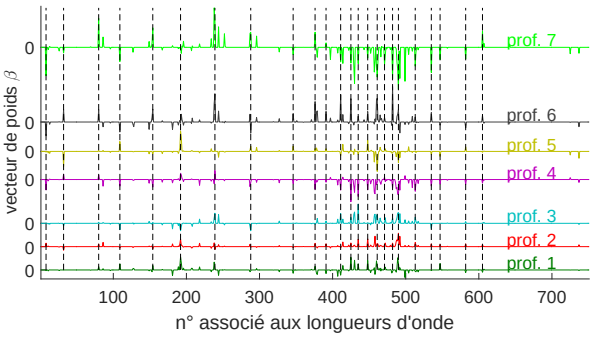

FIGURE 4 – FDA avec pénalité group LASSO : vecteur $\boldsymbol{\beta}$ obtenu

Bibliographie

[1] Lechuga G., Le Brusquet L., Perlbarg V., Puybasset L., Galanaud D., Tenenhaus A. (2015), Proceedings in Mathematics and Statistics, chapter Discriminant Analysis for Multiway Data. Springer Verlag.

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009), The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Springer.

[3] Clemmensen, L., Hastie, T., Witten, D. and Ersbøll B. (2011), Sparse discriminant analysis, Technometrics, 53(4) : 406-413.

[4] Boyd S., Parikh N., Chu E., Peleato B., and Eckstein J. (2011), Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Foundations and Trends in Machine Learning.