La Revue de Modulad

Analyse Discriminante appliquée à l'étude du rythme cardiaque : développements méthodologiques

Gilles Celeux, Jean Clairambault
INRIA Rocquencourt F78153 Le Chesnay

Résumé

L'analyse statistique du rythme cardiaque du nouveau-né nous a conduits à illustrer différents aspects relativement complexes de l'analyse discriminante : l'analyse factorielle discriminante dans un cadre plurifactoriel et l'influence des probabilités a priori dans une phase décisionnelle. Ces aspects méthodologiques sont regroupés dans cet article.

Mots clés *: Analyse statistique du rythme cardiaque, analyse discriminante dans un cadre plurifactoriel, probabilités a priori.*

1 Introduction

Dans cet article, nous considérons le même problème d'analyse du rythme cardiaque que dans l'article précédent (Clairambault, Celeux (1991)) mais nous concentrons notre attention sur les particularités statistiques de ce problème. Notre souci est ici de mettre en évidence des points délicats, parfois négligés, de l'analyse discriminante à l'occasion de cette étude du rythme cardiaque chez les nouveau-nés. Contrairement à l'article précédent, nous négligerons l'interprétation médicale au profit de considérations de méthodologie statistique. Il est tout d'abord important de noter que les unités statistiques, qui sont les périodes de 512 battements, ne sont pas a priori indépendantes. La prise en compte des liaisons entre battements successifs est difficile et, dans cette étude, nous avons fait l'hypothèse d'indépendance de ces périodes entre elles. Cette approximation se justifie par le fait que les données ont été moyennées sur une période assez longue (512 battements). Nous nous autorisons ainsi à utiliser les programmes d'analyse statistique multidimensionnelle. Il convient, bien sûr, de garder à l'esprit que les unités statistiques ne sont en fait pas nécessairement indépendantes lors de l'interprétation des analyses. Une autre particularité importante de ces données réside dans le fait que les unités statistiques (les périodes de 512 battements) appartiennent à une entité autonome (le bébé ...). Il est fort possible que chaque nouveau-né présente un effet propre qu'il conviendrait de prendre en compte lors des analyses discriminantes sous peine d'obtenir des résultats tout à fait illusoires (cf. Kobilinsky 1990). L'analyse et l'élimination de cet éventuel "effet bébé" constitue le thème méthodologique principal de cet article et fait l'objet de la Section 2. Un aspect important de l'analyse discriminante concerne le choix des probabilités a priori des classes à reconnaître. Ce point fait l'objet de la Section 3. La Section 4 est consacrée à la présentation des résultats des analyses discriminantes prenant en compte l'effet bébé et incluant des variations sur les probabilités a priori des classes. Les résultats sont comparés à ceux obtenus dans l'article précédent.

2 Analyse factorielle discriminante dans un cadre plurifactoriel

Les 3 variables fréquentielles qui décrivent le signal RR dépendent de 3 facteurs qui sont le stade du sommeil, l'âge (conceptionnel) du nouveau né et l'effet propre du bébé. Les deux premiers facteurs sont précisément ceux dont on veut mesurer l'importance ; le troisième ne nous intéresse pas directement, mais doit être considéré car des variations importantes entre bébés risquent d'occulter les deux premiers facteurs.

Notons que nous pourrions présenter l'analyse discriminante des stades du sommeil (resp. des âges de conception) tous âges (resp. tous stades du sommeil) confondus. Mais, une telle présentation ne permet pas d'analyser l'influence de l'âge sur le stade du sommeil ou vice-versa. La bonne stratégie consiste ici à conditionner suivant l'âge (resp. le stade du sommeil) pour

discriminer les stades du sommeil (resp les âges). Aussi, dans la suite nous travaillons à âge ou à stade du sommeil fixé

La première des choses à faire est certainement d'évaluer l'importance relative du facteur bébé par rapport aux 2 autres. Pour ce faire, on utilise les modèles d'analyse de variance (cf. Kobilinsky, 1988)

$$\begin{array}{lcl} Y & = & X_{be}\alpha_{be} + X_{so}\alpha_{so} + \varepsilon \\ et & & \\ Y & = & X_{be}\beta_{be} + X_{ag}\beta_{ag} + \varepsilon' \end{array} \qquad (1)$$

où Y est la matrice de dimension (n, 3) donnant les valeurs des n périodes retenues sur les 3 variables HF, MF et BF. Cette matrice est supposée centrée en colonne.

$X_{be} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & \ldots \\ 0 & \ldots & \ldots & 1 \\ 0 & \ldots & \ldots & 1 \\ 0 & \ldots & \ldots & 1 \end{bmatrix}$ est la matrice indicatrice (connue) du numéro de bébé et a pour dimension (n, 24).

De même, X_{so} et X_{ag} sont les matrices indicatrices du stade de sommeil et de l'âge : elles ont pour dimensions respectives (n, 2) et (n, 3). Les matrices ε et ε' sont constituées de trois vecteurs aléatoires de $\mathbf{R}^n$, ε_i $(1 \leq i \leq 3)$, d'espérance nulle et qui, si l'on admet l'approximation d'indépendance, ont des matrices variance proportionnelles à l'identité. De plus, pour tout $i = 1, 2, 3$, les ε_i^j $(1 \leq j \leq n)$ sont supposés indépendants entre eux ainsi que les $\varepsilon_i'^j$ $(1 \leq j \leq n)$ Enfin, les paramètres α_{be} de dimension (24, 3) et α_{so} (2, 3) mesurent l'effet des facteurs "bébé" et "stade du sommeil" dans le premier modèle et les paramètres β_{be} et β_{ag} mesurent l'effet des facteurs "bébé" et "âge" dans le deuxième modèle.

Ces modèles ont été estimés à l'aide du programme MODLI de la bibliothèque Modulad. A titre d'exemple, nous donnons les résultats pour les deux situations "groupe à terme" et "sommeil calme". Nous ne reproduisons pas ici le détail des sorties de ce programme. On donne simplement, pour chaque variable, les valeurs des rapports F qui permettent de tester la nullité des différents effets étudiés.

Var	Effet	F	degrés de liberté	$F_{0.05}$
HF	bébé	19.9	(7, 122)	2.08
	sommeil	49.1	(1, 122)	3.92
MF	bébé	11.4	(7, 122)	2.08
	sommeil	4.0	(1, 122)	3.92
BF	bébé	8.7	(7, 122)	2.08
	sommeil	52.6	(1, 122)	3.92

Tableau 1. *Les rapports F associés aux effets bébé et stade du sommeil pour le groupe à terme*

Var	Effet	F	degrés de liberté	$F_{0.05}$
HF	bébé	24.4	(21,89)	1.72
	âge	154.1	(1,89)	3.96
MF	bébé	12.9	(21,89)	1.72
	âge	44.9	(1,89)	3.96
BF	bébé	7.1	(21,89)	1.72
	âge	19.6	(1,89)	3.96

Tableau 2. *Les rapports F associés aux effets bébé et âge en sommeil calme*

Le seuil $F_{0.05}$ représente le niveau de signification au seuil de 5 % d'une loi de Fisher pour le nombre de degrés de liberté considérés. Ces niveaux sont donnés à titre indicatif car, du fait d'éventuelles liaisons entre battements cardiaques successifs, les unités statistiques ne sont pas réellement indépendantes et le nombre de degrés de liberté associé au terme d'erreur est en fait inférieur ou égal à celui indiqué dans les tableaux 1 et 2.

Le tableau 1 fait apparaître que pour toutes les variables les trois effets sont bien présents. Mais l'effet bébé n'est jamais massif et est moins net que l'effet stade de sommeil pour les variables HF et BF. Le tableau 2 montre un effet âge nettement supérieur à l'effet bébé, en particulier en basse fréquence. En conclusion, on peut espérer discriminer les stades du sommeil ou les groupes d'âge, mais il serait dangereux de négliger la présence de l'effet bébé lors des analyses factorielles discriminantes

Nous allons présenter les équations de l'analyse factorielle discriminante éliminant l'effet bébé pour la discrimination des 2 stades du sommeil conditionnellement à l'âge. Le modèle d'analyse de variance à considérer est alors

$$Y = X_{be}a_{be} + X_{so}a_{so} + \varepsilon \tag{2}$$

On cherche une combinaison linéaire des variables HF, MF et BF qui sépare au mieux les stades du sommeil en étant le moins possible influencée par le numéro de bébé. On pourra trouver une analyse semblable de Kobilinsky dans (Chesneaux et Kobilinsky, 1982). Le principe de résolution consiste à se placer dans le sous-espace S de $\mathbf{R}^n$ (n étant le nombres d'unités statistiques considérées) orthogonal à l'espace engendré par les combinaisons linéaires des vecteurs colonnes de la matrice X_{be}. L'opérateur de projection sur S s'écrit

$$Q_{be} = I - X_{be}(X_{be}^t X b_{be})^{-1} X_{be}^t, \tag{3}$$

X^t désignant la matrice transposée de la matrice X.

De l'équation (2), on tire l'équation d'analyse de variance dans S

$$Q_{be}Y = Q_{be}X_{so}a_{so} + \varepsilon'. \tag{4}$$

On peut alors définir (cf. Kobilinsky 1990, formule (3.1)) les sommes de produits d'écarts totale, inter et intra, y étant un vecteur de $\mathbf{R}^n$:

$$\begin{aligned} T(y,y) &= y^t Q_{be}\, y \\ B(y,y) &= y^t P_{so\,be} y \\ W(y,y) &= y^t (Q_{be} - P_{so\,be}) y \end{aligned} \tag{5}$$

avec

$$P_{so\,be} = Q_{be} X_{so} (X_{so}^t Q_{be} X_{so})^{-1} X_{so}^t Q_{be} \tag{6}$$

On a la relation

$$T(y,y) = B(y,y) + W(y,y) \tag{7}$$

Ainsi, l'analyse factorielle discriminante éliminant l'effet bébé consiste à maximiser le rapport

$$\frac{y^t P_{so\,be} y}{y^t Q_{be} y} \tag{8}$$

Le premier et unique axe factoriel pour discriminer les 2 stades du sommeil est dirigé par le vecteur propre associé à la valeur propre non nulle de $Q_{be}^{-1} P_{so\,be} = X_{so}(X_{so}^t Q_{be} X_{so})^{-1} X_{so}^t Q_{be}$

3 Influence des probabilités a priori

Les probabilités a priori des groupes à reconnaître influent surtout dans la phase décisionnelle de la discrimination. Ainsi (cf. Celeux 1990) un individu x_i est affecté au groupe d'indice ℓ^* qui maximise $p_\ell f_\ell(x_i)$ où $f_\ell(x_i)$ est la densité de probabilité du groupe ℓ évaluée au point x_i et p_ℓ est la probabilité a priori du groupe ℓ ($p_\ell > 0$ pour tout $\ell = 1, \ldots, k$ et $\Sigma_\ell p_\ell = 1$). Souvent, l'utilisateur n'a pas une idée très précise des probabilités a priori. Dans ce cas, deux attitudes sont courantes

Soit on fait l'hypothèse que les données ont été générées suivant la loi du mélange $\Sigma_\ell p_\ell f_\ell(x)$, auquel cas les p_ℓ sont estimés par n_ℓ/n où n_ℓ est l'effectif du groupe ℓ dans l'ensemble d'apprentissage de taille n. C'est cette hypothèse qui a été faite dans l'article précédent.

Soit on fait l'hypothèse que les probabilités a priori des k groupes sont égales ($p_\ell = 1/k$ pour tout ℓ) (cf. Celeux et Turlot 1990). Nous verrons dans la Section 4 que ces choix peuvent avoir des conséquences pratiques considérables

Notons, enfin, qu'il est possible d'adapter de manière directe les formules de l'analyse factorielle discriminante en tenant compte des probabilités a priori (cf. Huang et Li, 1991). La matrice d'inertie interclasse devient

$$B = \sum_{\ell=1}^{k} p_\ell \left[g_\ell - \sum_{\ell'=1}^{k} p_{\ell'} g_{\ell'} \right] \left[g_\ell - \sum_{\ell'=1}^{k} p_{\ell'} g_{\ell'} \right]^t \tag{9}$$

g_ℓ étant le centre de gravité du groupe ℓ ($\ell = 1, \ldots, k$). La matrice d'inertie intraclasse s'écrit

$$W = \sum_{\ell=1}^{k} p_\ell V_\ell \tag{10}$$

V_ℓ représentant la matrice variance du groupe ℓ. L'analyse factorielle discriminante consiste à rechercher les $k-1$ axes orthogonaux pour la métrique W qui maximisent

$$\frac{u^t B u}{u^t W u} \tag{11}$$

u étant un vecteur de $\mathbf{R}^p$ qui dirige l'axe.

4 Applications numériques

Dans cette section, nous comparons les résultats des analyses discriminantes selon que nous éliminons ou non l'effet bébé et selon que nous prenons des probabilité a priori des groupes proportionnelles aux effectifs ou égales. Notre but est d'illustrer les considérations des deux sections précédentes. Aussi, nous nous contentons de présenter le traitement de 2 analyses discriminantes où les différences selon la prise en compte ou non de l'effet bébé sont les plus marquées : la discrimination du stade du sommeil pour les bébés à terme, et la discrimination de l'âge en sommeil calme.

4.1 Discrimination du stade du sommeil

Le tableau 3 donne la corrélation totale de la Forme Linéaire Discriminante (FLD) avec les 3 variables fréquentielles pour discriminer SA et SC 'sans' prise en compte de l'effet bébé et 'avec' prise en compte de l'effet bébé pour les bébés à terme.

FLD	HF	MF	BF
sans	−0.80	0.14	0.68
avec	−0.24	0.42	0.99

Tableau 3 *Corrélation de la FLD avec les variables cardiaques sans et avec prise en compte de l'effet bébé pour discriminer les 2 stades du sommeil pour le groupe terme.*

Le tableau 4 donne les tableaux d'affectation obtenus par validation croisée issus de ces analyses où, de plus, l'on considère deux choix différents pour les probabilités a priori. Les variations sur les probabilités a priori ont une influence énorme au point de modifier la perception de la prise en compte de l'effet bébé. Lorsque les probabilités a priori des groupes sont prises proportionnelles aux effectifs, la prise en compte de l'effet bébé permet de diminuer légérement le taux d'erreur grâce à une amélioration de la reconnaissance du sommeil calme. En revanche, si les probabilités a priori sont prises égales, le taux d'erreur se détériore notablement, même si là encore le sommeil calme est mieux reconnu. Ce résultat surprenant pourrait surtout témoigner du caractère artificiel de la bonne reconnaissance du sommeil agité lorsqu'on le pondère fortement (ici, environ 3/4 du poids total).

	pas de prise en compte de l'effet bébé			prise en compte de l'effet bébé		
probabilités a priori proportionnelles aux effectifs	*g.o/g.a*	*SA*	*SC*	*g.o/g.a*	*SA*	*SC*
	SA	94	2	*SA*	93	3
	SC	21	14	*SC*	18	17
probabilités a priori égales	*g.o/g.a*	*SA*	*SC*	*g.o/g.a*	*SA*	*SC*
	SA	85	11	*SA*	63	33
	SC	14	21	*SC*	9	26

Tableau 4 *Evolution de la discrimination linéaire entre stades du sommeil, pour le groupe à terme, en fonction de la prise en compte ou non de l'effet bébé et du choix des probabilités a priori (g.a : groupe d'affectation g.o : groupe d'origine, SA : sommeil agité, SC : sommeil calme).*

4.2 Discrimination de l'âge

Pour cet exemple, nous avons supprimé le groupe intermédiaire difficile à distinguer des 2 autres et nous limitons l'analyse discriminante à la reconnaissance des groupes prématurés et 'à terme' en sommeil calme. Le tableau 5 résume les résultats de l'analyse de variance (équation (1)) pour cette population restreinte

Var	Effet	F	degrés de liberté	$F_{0.05}$
HF	bébé	37.9	(14, 58)	1.92
	âge	358.8	(1, 58)	4.07
MF	bébé	10.3	(14, 58)	1.92
	âge	83.1	(1, 58)	4.07
BF	bébé	5.8	(14, 58)	1.92
	âge	36.5	(1, 58)	4.07

Tableau 5 *Les rapports F associés aux effets bébé et âge en sommeil calme sans le groupe intermédiaire*

Ce tableau montre que la suppression du groupe intermédiaire conduit à une augmentation de l'effet âge par rapport à l'effet bébé. Ainsi, dans ce cas, la prise en compte de l'effet bébé dans la discrimination ne devrait pas modifier notablement les résultats.

FLD	HF	MF	BF
sans	0.90	0.79	0.67
avec	0.82	0.70	0.79

Tableau 6 *Corrélation de la FLD avec les variables cardiaques sans et avec prise en compte de l'effet bébé pour discriminer l'âge en sommeil calme*

Le tableau 6 donne la corrélation de la FLD avec les 3 variables pour discriminer les prématurés et le groupe à terme sans et avec prise en compte de l'effet bébé. Le tableau 7 donne les tableaux d'affectation issus de ces analyses où, de plus, l'on considère deux choix différents pour les probabilités a priori.

	pas de prise en compte de l'effet bébé			prise en compte de l'effet bébé		
probabilités a priori proportionnelles aux effectifs p	$g.o/g.a$	$t.$	$p.$	$g.o/g.a$	$t.$	$p.$
	$t.$	24	11	$t.$	25	10
	$p.$	6	33	$p.$	7	32
probabilités a priori égales	$g.o/g.a$	$t.$	$p.$	$g.o/g.a$	$t.$	$p.$
	$t.$	24	11	$t.$	26	9
	$p.$	8	31	$p.$	7	32

Tableau 7. ***Evolution de la discrimination linéaire entre âges (v prématuré, t à terme) en sommeil calme, en fonction de la prise en compte ou non de l'effet bébé et du choix des probabilités a priori*** *(g.a : groupe d'affectation, g.o : groupe d'origine).* Ici, les variations sur les probabilités a priori sont peu sensibles. Ce n'est pas étonnant du fait que les rapports n_i/n sont proches de 1/2. Comme l'examen du tableau 5 pouvait le laisser prévoir la prise en compte de l'effet bébé ne modifie guère les résultats. Notons que cette fois, elle permet d'améliorer la qualité globale de la discrimination.

Remerciements. Les auteurs remercient vivement André Kobilinsky pour ses conseils et son aide lors de cette étude.

Bibliographie

[1] Celeux G. (1990) : Règles statistiques de décision. In : Analyse discriminante sur variables continues. pp. 15-36. Sous la direction de G. Celeux. Collection Didactique 7, INRIA.

[2] Celeux G., Turlot J. C. (1990) : Estimation de la qualité d'une règle de décision. In : Analyse discriminante sur variables continues. pp. 37-49. Sous la direction de G. Celeux. Collection Didactique 7, INRIA.

[3] Chesnaux, M-T., Kobilinsky, A. (1982) : Possibilité d'identification variétale du maïs au stade plantule. Agronomie 2 (1). 45-54.

[4] Clairambault, J., Celeux, G. (1991) : Analyse discriminante appliquée à l'étude du rythme cardiaque. La Revue de Modulad 8. 61-72.

[5] Huang X-N., Li B-B. (1991) : A New Discriminant Technique: Bayes-Fisher Discrimination. Biometrics 47. 741-744.

[6] Kobilinsky, A. (1988) : Tactiques en analyse de variance et en Régression. La Revue de Modulad 1. 25-58.

[7] Kobilinsky, A. (1990) : Analyse Factorielle Discriminante. In : Analyse discriminante sur variables continues, pp. 65-80. Sous la direction de G. Celeux. Collection Didactique 7, INRIA.