# Réduction de dimension dans les modèles linéaires généralisés : application à la classification supervisée de données issues des biopuces

Gersende Fort[*], Sophie Lambert-Lacroix et Julie Peyre [†]

RÉSUMÉ. Une des principales caractéristiques des données issues des biopuces réside dans le fait qu'on dispose d'un nombre d'individus $n$ très faible devant un grand nombre de gènes $p$. Un des enjeux majeurs du traitement statistique de ces données est l'analyse discriminante à but décisionnel. D'un point de vue statistique, ce grand nombre de covariables devant un petit nombre d'observations rend l'analyse discriminante difficile. Une façon de contourner ce "fléau de la dimension" consiste à réduire cette dimension. Dans ce papier, la classification supervisée est vue comme un problème de régression avec peu d'observations et beaucoup de covariables. Nous proposons deux approches de réduction de dimension. La première, de type paramétrique, est une extension de la méthode "Partial Least Squares" aux modèles linéaires généralisés. La deuxième, de type semi-paramétrique, utilise des estimateurs par vraisemblance locale dans les modèles linéaires généralisés en indice simple. Nous illustrons la pertinence de ces approches lorsqu'elles sont appliquées à l'analyse de deux jeux de données réels : Colon et Leukemia.

ABSTRACT. A major characteristic of microarray data consists in the small number of observations n and a very large number of genes p. An important aspect of data-mining of microarray data is the discriminant analysis for classification. This "large p, small n" paradigm makes the statistical analysis difficult. To answer this curse of dimensionnality, a solution consists in reducing the dimension. In this paper, supervised classification is understood as a regression problem with a small number of observations and a large number of covariables. We give two approaches for dimension reduction. The first one, based on a parametric approach, is an extension of the "Partial Least Squares" method to the generalized linear models [Algorithm /Ridge-PLS/]. The second one, based on a semi-parametric approach, uses local likelihood estimates for single index generalized linear models [Algorithm /GSIM/]. We consider applications of our methods when applied to binary and multi-class classification of two real data sets : Colon and Leukemia.

**Mots clés :** Modèles linéaires généralisés, Modèles linéaires généralisés en indice simple, Réduction de dimension, Partial Least Squares, Vraisemblance locale, Analyse discriminante, Données issues des biopuces.

[*]CNRS/LTCI, 46, rue Barrault, 75634 Paris Cedex 13. gfort@tsi.enst.fr

[†]LMC-IMAG, 51, rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9 ; Sophie.Lambert, Julie.Peyre@imag.fr

# 1 Introduction

Grâce à la technologie des biopuces, nous disposons d'une foule de données obtenues en mesurant par le processus d'hybridation, le niveau d'expression des gènes d'une entité biologique. On peut espérer que cette technologie permette des avancées spectaculaires dans des domaines aussi variés que l'agro-alimentaire, l'environnement, la médecine. Dans ce dernier domaine en particulier, les biopuces sont utilisées pour le séquençage par hybridation comme alternative à la méthode enzymatique classique ; pour l'identification de cible pour la recherche thérapeutique, comme outil de compréhension plus fine du génome et de sa régulation ; pour la pharmacogénomique, comme façon de montrer qu'une molécule a sur une cible une action variable et en déduire des traitements adaptés aux patients ; pour le diagnostic de maladies infectieuses et génétiques. C'est pour ce dernier domaine d'applications que sont construites les biopuces dont nous voulons faire l'analyse statistique : le but de l'analyse de ces données est celui de la classification d'individus selon leur pathologie. Par exemple, Golub et al. (1999) utilisent ces données pour classer des individus selon différents types de leucémie. Alon et al. (1999) considèrent la classification supervisée de tissus de colon en fonction de leur état : tumoraux ou sains. Dans ces papiers, les approches utilisées font appel à des techniques de type analyse discriminante ou "machine learning" (cf. Dudoit et al. (2002) pour une étude comparative).

Une des principales caractéristiques de l'analyse des données issues des biopuces réside dans le fait que le nombre de sujets $n$ est relativement faible (de l'ordre de la centaine) par rapport au nombre de gènes $p$ (de l'ordre de quelques milliers). Sans préalablement réduire la dimension, les méthodes standards de statistique en classification supervisée, même utilisables, ne sont pas très performantes. En particulier, un des problèmes est lié à une très forte multicolinéarité des $p$ régresseurs. Les solutions des équations intervenant dans les méthodes traditionnelles peuvent ne plus être uniques et être instables. Par exemple dans le cas de deux classes, la matrice de variance intraclasse dans la fonction de Fisher est singulière lorsque $n < p + 2$ (Saporta (1990)). Même si on peut utiliser tous les gènes, il ne semble pas raisonnable de le faire. En effet, utiliser tous les gènes introduit du bruit via des gènes qui ont un faible pouvoir discriminant, dégradant ainsi les performances des méthodes utilisées. Dans ce cas, les méthodes de réduction de dimension peuvent s'avérer utiles. Dans les papiers cités ci-dessus, les auteurs utilisent des techniques unidimensionnelles (*i.e.* appliquées gène par gène) pour réduire le nombre de gènes.

On trouve plusieurs approches alternatives ; en particulier, des méthodes paramétriques basées sur la méthode "Partial Least Squares" (PLS). Cette méthode a d'abord été introduite dans le domaine de la chimiométrie (Cf. l'article de synthèse par Frank et Friedman (1993), et les références citées). Dans le cadre des biopuces, PLS produit des combinaisons linéaires orthogonales de gènes qui permettent de réduire la dimension à quelques "super-gènes". A la différence de l'analyse en composantes principales (ACP), PLS utilise la variable réponse à travers un critère basé sur la covariance empirique, et semble donc plus adapté dans les problèmes de régression. Nguyen et Rocke (2002b,a) réduisent la dimension via PLS, avant d'appliquer des techniques standards comme la régression logistique, l'analyse discriminante linéaire ou l'analyse discriminante quadratique. Cependant, PLS est introduit pour des variables continues homoscédastiques et l'utiliser directement avec des variables réponses catégorielles (Bernoulli ou Multinomial) ne semble pas d'un point de vue intuitif très approprié. De plus, en pratique, on observe des problèmes de convergence dans l'algorithme des moindres carrés itératif (IRLS) utilisé pour résoudre l'inférence par maximum de vraisemblance dans un modèle de régression logistique. Ce problème nommé problème de séparation ou quasi-séparation, est en fait lié à la configuration des $n$ vecteurs de régresseurs dans l'espace des covariables (voir par exemple Albert et Anderson (1984)).

Plus récemment, Ding et Gentleman (2005) ont proposé d'appliquer aux données issues des biopuces, une approche basée sur une extension de PLS aux modèles linéaires généralisés (GLM). Cette extension consiste d'une part, à remplacer dans l'étape de régression de l'algorithme IRLS, les moindres carrés par une régression type PLS et d'autre part, à introduire une pénalité de type Firth pour éviter les problèmes de (quasi-) séparation. Cependant, on peut également observer des problèmes de convergence comme des comportements cycliques (Fort et Lambert-Lacroix (2005); Fort (2005)). Enfin, Fort et Lambert-Lacroix (2005) ont proposé une extension de PLS pour des variables réponses de type Bernoulli, basée sur une pénalité Ridge et l'ont appliquée à la classification supervisée de biopuces.

Il existe des approches alternatives semi-paramétriques. Antoniadis et al. (2003) proposent d'utiliser la méthode MAVE (acronyme de Minimum Average Variance Estimation, Xia et al. (2002)) pour réduire la dimension avant d'appliquer une régression logistique paramétrique ou non paramétrique. La procédure MAVE est basée sur un critère de moindres carrés local combiné avec une estimation non paramétrique par polynômes locaux de la fonction de régression. Bien qu'applicable aux modèles linéaires généralisés, cette méthode n'exploite ni la structure particulière de ces modèles, à savoir la relation entre espérance et variance, ni le fait que dans les GLM on utilise des critères basés sur la vraisemblance (et qui ne coïncident avec les critères basés sur les moindres carrés que dans les modèles gaussiens).

Dans ce papier, nous proposons d'étudier deux approches qui permettent de réduire la dimension dans les modèles linéaires généralisés. La première, de type paramétrique, est l'extension de la méthode PLS proposée dans le cas de variables de réponse de Bernoulli par Fort et Lambert-Lacroix (2005). La deuxième approche est de type semi-paramétrique et utilise des estimateurs par vraisemblance locale dans les modèles linéaires généralisés en indice simple. Cette dernière est comparable à la méthode MAVE : le critère basé sur les moindres carrés est remplacé par un critère basé sur la vraisemblance locale, exploitant ainsi la structure particulière des modèles linéaires généralisés. Cette deuxième approche fera l'objet d'une publication ultérieure en cours de préparation (Lambert-Lacroix et Peyre (2005)).

Dans toutes les approches de réduction de dimension citées précédemment, on utilise une méthode de compression de variables afin d'éviter le fléau de la dimension : la dimension est réduite en passant d'un grand nombre de covariables à quelques combinaisons linéaires de ces covariables. L'objectif de la compression est de rechercher des directions informatives, et d'éliminer des directions qui ne contiennent que du bruit. Il existe une autre famille de méthodes, celles de sélection de variables, qui consiste à exhiber quelques gènes informatifs. Il semble effectivement intéressant de sélectionner des gènes plutôt que des combinaisons linéaires (faisant a priori intervenir tous les gènes) pour par exemple établir un diagnostic. Les techniques de sélection de variables se divisent en deux groupes : les approches purement scalaires (*i.e.* gènes par gènes) et celles multivariées. Dans Kuo et al. (2004), les auteurs indiquent que le deuxième groupe est à privilégier. Ils précisent également que ces méthodes commencent à être développées et qu'il serait utile d'en développer de nouvelles. Cependant, l'approche compression de variables dans les GLM, n'est pas incompatible avec une approche sélection de variables. Les méthodes de régression proposées ici peuvent être exploitées pour la sélection de variables, en adaptant par exemple le principe de l'algorithme *Recursive Feature Elimination* (RFE) proposé par Guyon et al. (2002). On trouvera dans Fort (2005) le développement de cette approche pour des méthodes type PLS. Les résultats pour la méthode semi-paramétrique font l'objet d'études en cours et ce point, bien que très intéressant ne sera pas traité ici.

Le papier est organisé de la manière suivante. La section 2 est consacrée aux modèles linéaires généralisés canoniques. En particulier, nous rappelons les méthodes d'estimation classiques, paramétrique et non paramétrique, et exposons les problèmes rencontrés en grande dimension. Dans la section 3, nous présentons l'extension de PLS dans le cas général des GLM. La quatrième partie est consacrée à l'approche semi-paramétrique basée sur les modèles linéaires généralisés en indice simple. Enfin dans la dernière section, nous étudions l'intérêt de ces approches appliquées à la classification supervisée des puces à ADN. Nous considérons pour ce faire deux jeux de données : Colon et Leukemia.

# 2 Inférence dans les modèles linéaires généralisés canoniques

Nous consacrons cette section aux modèles linéaires généralisés canoniques : après une brève définition de ces modèles, nous présentons dans une approche paramétrique, la méthode d'estimation des coefficients de régression par maximum de vraisemblance. Nous discutons particulièrement de l'existence et de l'unicité de cet estimateur dans les modèles de régression logistique di/polychotomique. Nous rappelons ensuite dans une approche semi-paramétrique, la méthode des polynômes locaux. Dans les deux cas, nous soulignons les limitations de ces approches en grande dimension $n << p$.

## 2.1 Notations

Par convention, les vecteurs sont des vecteurs colonne. Étant donnés deux entiers $a < b$, $a : b$ désigne le vecteur de coordonnées $(a, a+1, \cdots, b)$. Soit $\underline{k} = (k_1, \ldots, k_p)$ un vecteur de $\mathbb{N}^p$, on introduit les notations suivantes :

$$|\underline{k}| = \sum_{j=1}^{p} k_j, \quad \underline{k}! = \prod_{j=1}^{p} k_j!, \quad \forall x \in \mathbb{R}^p \ , \ x^{\underline{k}} = \prod_{j=1}^{p} x_j^{k_j}.$$

Pour une matrice $A$, $A_{i,j}$ désigne l'élément $(i, j)$ (si $A$ est un vecteur, on écrira simplement $A_i$). $A_{i:j,:}$ est la matrice extraite de $A$ constituée des lignes $i$ à $j$. $\mathbb{1}_n$ est le vecteur de $\mathbb{R}^n$ dont toutes les coordonnées sont égales à 1 ; $(e_1, \cdots, e_n)$ désigne la base canonique de $\mathbb{R}^n$ ; et $\mathrm{Id}_n$

est la matrice identité d'ordre $n \times n$. Enfin, $^T$ est l'opérateur de transposition et $\|\cdot\|$ la norme euclidienne.

## 2.2 Modèles linéaires généralisés canoniques

Pour une présentation détaillée des modèles linéaires généralisés, on pourra consulter Fahrmeir et Tutz (2001).

### 2.2.1 Définition

On dispose de $n$ couples indépendants $(Y_i, X_i)$, $i = 1, \ldots, n$, de même loi que le couple $(Y, X)$ ; $Y_i$ est la réponse à valeur dans $\mathbb{R}^G$, $G \in \mathbb{N}^*$, associée aux covariables $p$-dimensionnelles $X_i$. La loi conditionnelle de $Y$ sachant $X$ est supposée appartenir à la famille des lois exponentielles, de densité conditionnelle donnée par

$$\exp\left(\frac{y^T\eta(x) - b(\eta(x))}{\phi} + c(y, \phi)\right). \tag{1}$$

Les fonctions $b$, $c$ et le paramètre de dispersion $\phi > 0$ sont connus. Soit $g\ : \mathbb{R}^G \to \mathbb{R}^G$ une fonction inversible, spécifiant la relation entre le prédicteur $\eta$ et la fonction de régression $\mu$ (*i.e.* l'espérance conditionnelle $\mu(x) = \mathbb{E}[Y|X = x]$) selon la relation

$$\eta = g(\mu). \tag{2}$$

$g$ est appelée la fonction de lien canonique. La donnée des réalisations des $n$ couples indépendants, de la modélisation de l'aléa (1) et de la fonction de lien (2) définissent un modèle canonique généralisé.

Dans les modèles linéaires généralisés paramétriques, le prédicteur $\eta$ est linéaire en les régresseurs $x$ :

$$\eta_i(x) = \theta_1^{(i)} + x^T\theta_{2:p+1}^{(i)} = z^T\theta^{(i)}, \qquad 1 \le i \le G,$$

avec $z = [1\ x^T]^T$. Le vecteur des coefficients de régression $\theta \in \mathbb{R}^{G(p+1)}$ est obtenu par concaténation des $G$ vecteurs $\theta^{(1)}, \cdots, \theta^{(G)}$ de $\mathbb{R}^{p+1}$ et est supposé inconnu.

Dans certaines situations, la relation paramétrique ci-dessus n'est pas garantie et le modèle linéaire peut ne pas être assez général. Une approche plus souple est de supposer que $\eta$ est une fonction non paramétrique des régresseurs $x$.

### 2.2.2 Exemple : Modèle de régression logistique

Le modèle de régression logistique est pertinent pour l'analyse discriminante car il permet de modéliser une variable catégorielle. Le label à valeur dans $\{0, \cdots, G\}$ peut s'exprimer à l'aide de $G$ variables à valeur dans $\{0, 1\}$ selon la convention suivante : la $j$-ième variable vaut 1 et les autres sont nulles si et seulement si le label vaut $j$ ; le label 0 est codé par $G$ variables nulles. Ce

vecteur binaire de dimension $G$ suit un modèle multinomial d'ordre 1, de paramètres $(\mu_0, \cdots, \mu_G)$ où $\mu_j$ est la probabilité de la $j$-ième classe. Le cas $G = 1$ est le modèle de Bernoulli.
Il s'agit d'un modèle généralisé ; $Y \in \{0,1\}^G$ a une composante au plus égale à 1 selon la convention : sa $i$-ième composante vaut 1 si et seulement si la variable catégorielle prend la $i$-ième modalité. La fonction de lien est donnée par $\eta_i = \ln \mu_i - \ln \mu_0$ pour tout $1 \leq i \leq G$ et le modèle exponentiel est défini par $\phi = 1$, $b(\eta) = \ln(1 + \sum_{l=1}^{G} \exp(\eta_l))$ et $c = 0$. Par suite,

$$\mu_i = \frac{\exp(\eta_i)}{1 + \sum_{l=1}^{G} \exp(\eta_l)}, \qquad 1 \leq i \leq G. \tag{3}$$

## 2.3 Approche paramétrique

L'estimation du paramètre $\theta$ est obtenue par la méthode du maximum de vraisemblance. Pour ce faire, le support de la loi dominante est supposé indépendant du paramètre. Le modèle est de plus, supposé suffisamment régulier pour que la fonction de log-vraisemblance soit deux fois continûment différentiable, que l'estimateur du maximum de vraisemblance soit unique lorsqu'il existe et qu'il soit solution des équations de vraisemblance. Ces hypothèses sont en particulier vérifiées par les modèles de régression logistique.

### 2.3.1 Notations matricielles

Les observations $(y_i, x_i)$, $1 \leq i \leq n$, des $n$ couples $(Y_i, X_i)$ sont collectées dans un vecteur des variables de réponses $\mathbf{Y} \in \mathbb{R}^{Gn}$ et dans une matrice $\mathbf{X}^{(G)} \in \mathbb{R}^{Gn \times Gp}$. La réalisation $y_k$ de $Y_k$ constitue le $k$-ième bloc de $\mathbf{Y}$

$$\mathbf{Y}_{\iota_k+1:\iota_k+G} = y_k, \qquad \text{en ayant posé} \qquad \iota_k = (k-1)G, \qquad k = 1, \cdots, n. \tag{4}$$

Les lignes $\iota_k + 1$ à $\iota_k + G$ de $\mathbf{X}^{(G)}$ sont construites à l'aide de la réalisation $x_k$ de $X_k$, de la façon suivante

$$\mathbf{X}^{(G)}_{\iota_k+1:\iota_k+G,:} = \begin{bmatrix} x_k^T & 0 & \cdots & 0 \\ 0 & x_k^T & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & x_k^T \end{bmatrix}. \tag{5}$$

De manière analogue, on construit la matrice bloc $\mathbf{Z}^{(G)} \in \mathbb{R}^{Gn \times G(p+1)}$ à partir des réalisations $z_k = [1 \ x_k^T]^T$. Notons que lorsque $G = 1$, on retrouve la matrice de plan d'expérience $\mathbf{X}^{(1)}$ de taille $n \times p$, et que $\mathbf{Z}^{(1)} = [\mathbb{1}_n \ \mathbf{X}^{(1)}]$. On suppose que la matrice $\mathbf{X}^{(1)}$ est centrée en colonnes ($\mathbb{1}_n^T \mathbf{X}^{(1)}_{:,j} = 0$ pour tout $1 \leq j \leq p$), de sorte que les coefficients de régression $\theta_1^{(i)}$, $1 \leq i \leq G$, sont identifiables.
La fonction de log-vraisemblance $\ell(\theta)$ s'exprime donc, à un terme additif et un terme positif multiplicatif près indépendants de $\theta$, par

$$\ell(\theta) = \mathbf{Y}^T \mathbf{Z}^{(G)} \theta - \sum_{k=1}^{n} b\left( (\mathbf{Z}^{(G)} \theta)_{\iota_k+1:\iota_k+G} \right).$$

Les équations de vraisemblance s'écrivent $\mathbf{Z}^{(G)^T}(\mathbf{Y}-\boldsymbol{\mu}(\theta))=0$ et le vecteur des moyennes $\boldsymbol{\mu}$ est donné par

$$\forall 1 \leq k \leq n, \qquad \boldsymbol{\mu}_{\iota_k+1:\iota_k+G}(\theta) = \nabla b(\varepsilon)\Big|_{\varepsilon=(\mathbf{Z}^{(G)}\theta)_{\iota_k+1:\iota_k+G}}, \tag{6}$$

$\nabla$ désignant l'opérateur de différentiation.

### 2.3.2 $\mathbf{Z}^{(G)}$ est de rang plein en colonnes

$\ell(\theta)$ dépend de $\theta$ à travers le prédicteur linéaire $\mathbf{Z}^{(G)}\theta$ ; lorsque $\mathbf{Z}^{(G)}$ est de rang plein en colonnes, le paramètre est identifiable. Les équations de vraisemblance ne sont pas, en général, linéaires en $\theta$ et se résolvent par des méthodes itératives. Lorsque l'estimateur du maximum de vraisemblance $\hat{\theta}^{\text{MV}}$ existe et est unique, la méthode de Newton-Raphson est certainement la méthode la plus répandue pour le calculer ; sa mise en œuvre est connue sous le nom d'algorithme IRLS (acronyme de Iteratively Reweighted Least Squares, Green (1984)). Soit $\mathbf{W}(\theta) \in \mathbb{R}^{Gn\times Gn}$, une matrice bloc-diagonale, définie positive, de $k$-ième bloc égal à $\nabla^2 b\left((\mathbf{Z}^{(G)}\theta)_{\iota_k+1:\iota_k+G}\right) \in \mathbb{R}^{G\times G}$, pour $1 \leq k \leq n$. Rappelons que

$$\mathbb{E}\left[(Y-\mu)(Y-\mu)^T | X=x\right] = \phi\ \nabla^2 b\left(\eta(x)\right), \tag{7}$$

si bien que $\phi\ \mathbf{W}(\theta)$ est une matrice de dispersion conditionnelle. En remarquant que $\nabla^2\ell(\theta) = -\mathbf{Z}^{(G)^T}\mathbf{W}(\theta)\mathbf{Z}^{(G)}$, IRLS produit une suite $(\theta^t)_t$ selon la dynamique

$$\begin{aligned}\theta^{t+1} &= \theta^t + \left[\mathbf{Z}^{(G)^T}\ \mathbf{W}(\theta^t)\ \mathbf{Z}^{(G)}\right]^{-1} \mathbf{Z}^{(G)^T}\left(\mathbf{Y}-\boldsymbol{\mu}(\theta^t)\right)\\ &= \left[\mathbf{Z}^{(G)^T}\ \mathbf{W}(\theta^t)\ \mathbf{Z}^{(G)}\right]^{-1} \mathbf{Z}^{(G)^T}\ \mathbf{W}(\theta^t)\ \left\{\mathbf{Z}^{(G)}\theta^t + \left[\mathbf{W}(\theta^t)\right]^{-1}(\mathbf{Y}-\boldsymbol{\mu}(\theta^t))\right\}.\end{aligned}$$

En d'autres termes, $\theta^{t+1}$ est obtenu par régression pondérée d'une pseudo-variable

$$\psi^t = \mathbf{Z}^{(G)}\theta^t + \left[\mathbf{W}(\theta^t)\right]^{-1}(\mathbf{Y}-\boldsymbol{\mu}(\theta^t)) \tag{8}$$

sur les colonnes de $\mathbf{Z}^{(G)}$. Toute suite convergente converge vers $\lim_t \theta^t = \hat{\theta}^{\text{MV}}$.

### 2.3.3 $\mathbf{Z}^{(G)}$ n'est pas de rang plein en colonnes

Dans ce cas, le paramètre $\theta$ n'est plus identifiable. Néanmoins, on peut toujours se ramener, au moyen de calculs matriciels standards, au cas d'une matrice $\mathbf{Z}^{(\text{red},G)}$ de rang plein en colonnes. On obtient alors, quand il existe, un estimateur du maximum de vraisemblance $\hat{\gamma}^{\text{MV}} \in \mathbb{R}^{\text{rang}(\mathbf{Z}^{(\mathbf{G})})}$ et on définit $\hat{\theta}^{\text{MV}}$ comme le vecteur de norme minimale vérifiant la condition $\mathbf{Z}^{(G)}\theta = \mathbf{Z}^{(\text{red},G)}\hat{\gamma}^{\text{MV}}$.

### 2.3.4 Cas particulier du modèle de régression logistique paramétrique

Dans ce cas, $\mathbf{Y}$ est à valeur dans $\{0,1\}^{nG}$ et est tel que $\sum_{i=\iota_k+1}^{\iota_k+G} \mathbf{Y}_i \in \{0,1\}$. Lorsque $\mathbf{Z}^{(G)}$ est de rang plein en colonnes, l'estimateur du maximum de vraisemblance n'existe pas nécessairement : Albert et Anderson (1984); Santner et Duffy (1986); Lesaffre et Albert (1989) ont montré que l'existence dépendait de la configuration des $n$ vecteurs de régresseurs dans l'espace des covariables. On peut ainsi distinguer trois cas : le cas de séparation, le cas de quasi-séparation et le cas de mélange. Le cas "séparation" signifie qu'il existe $\theta$ tel que pour tout $1 \leq k \leq n$, $1 \leq i \leq G$,

$$\left[(\mathbf{Z}^{(1)}\theta^{(i)})_k > (\mathbf{Z}^{(1)}\theta^{(l)})_k, \quad \forall l \in \{0, \cdots, G\} \setminus \{i\}\right] \qquad \text{ssi} \qquad \mathbf{Y}_{\iota_k+i} = 1, \tag{9}$$

en ayant posé $\theta^{(0)} = 0$ par convention. Le cas "quasi-séparation" signifie que (9) est vraie avec des inégalités larges. Dans ces deux cas, la fonction $\ell$ est maximale quand $\|\theta\| \to +\infty$ et l'estimateur du maximum de vraisemblance n'existe pas. Dans le cas complémentaire dit de "mélange", $\hat{\theta}^{\mathrm{MV}}$ existe et est unique ; le choix $\boldsymbol{\mu}(\theta^0) = (G+3)^{-1}(3\mathbf{Y} + (\mathbb{1}_{Gn} - \mathbf{Y}))$ est une initialisation efficace pour obtenir une trajectoire convergente de l'algorithme IRLS (Fahrmeir et Tutz (2001)).
Les applications considérées en Section 5 relèvent du cadre statistique $n << p$, si bien que, en pratique, $\mathrm{rang}(\mathbf{Z}^{(\mathrm{G})}) = Gn$, et $\mathbf{Z}^{(G)}$ n'est donc pas de rang plein en colonnes. En reparamétrant le modèle, les équations de vraisemblance deviennent $\mathbf{Y} = \mu$, ce qui entraîne

$$(\mathbf{Z}^{(\mathrm{red},G)}\hat{\gamma})_{\iota_k+i} = \ln\left(\frac{\mathbf{Y}_{\iota_k+i}}{1 - \sum_{l=1}^{G} \mathbf{Y}_{\iota_k+l}}\right), \qquad \forall 1 \leq k \leq n, 1 \leq i \leq G.$$

On a donc $\|\hat{\gamma}\| = +\infty$ et, par suite, l'estimateur du maximum de vraisemblance ne peut jamais exister. Il est alors nécessaire de considérer des méthodes de réduction de dimension afin de ramener le problème de régression dans un espace de grande dimension à un problème de régression dans un sous-espace de dimension $\kappa$.

## 2.4 Approche non paramétrique

Nous rappelons ci-après la méthode des polynômes locaux introduite par Fan et Gijbels (1996). Par souci de simplicité nous nous plaçons dans le cas où $G = 1$, *i.e.* dans le cas où la variable réponse est scalaire.
L'idée est d'approximer la fonction $\eta$ localement par une fonction polynômiale d'ordre $q$,

$$\eta(u) \sim \sum_{\underline{k} \in \mathcal{A}_q} D^{\underline{k}}\eta(x)\ (u-x)^{\underline{k}} = \sum_{\underline{k} \in \mathcal{A}_q} a_{\underline{k}}\ (u-x)^{\underline{k}},$$

pour $u$ dans un voisinage de $x$. Par définition, $\mathcal{A}_q = \{\underline{k}, |\underline{k}| \leq q\}$ et $D^{\underline{k}}f(x)$ désigne la dérivée partielle $\partial^{|\underline{k}|}f(x)/\partial x^{\underline{k}}$.
Soient $K^p$ un noyau $p$-dimensionnel avec une matrice de taille de fenêtres de lissage $H$ et $K^p_H(\cdot) = \det(H)^{-1}K^p(H^{-1} \times \cdot)$ le changement d'échelle de $K^p$. La vraisemblance locale est une

vraisemblance pondérée, avec les poids $K_H^p(X_i - x)$. Notons $\mathcal{L}(u, Y)$ la fonction log-vraisemblance avec $\eta(x)$ remplacé par son approximation polynômiale locale ; la log-vraisemblance locale est définie par

$$\sum_{i=1}^{n} \mathcal{L}\left(\sum_{\underline{k}\in\mathcal{A}_q} a_{\underline{k}}\ (X_i - x)^{\underline{k}}, Y_i\right) \quad K_H^p(X_i - x). \tag{10}$$

La méthode des polynômes locaux conduit à $\widehat{D^{\underline{k}}\eta}(x) = \underline{k}!\hat{a}_{\underline{k}}(x)$, où $\{\hat{a}_{\underline{k}}(x), \underline{k} \in \mathcal{A}_q\}$ maximise le critère (10) comme fonction de $\{a_{\underline{k}}, \underline{k} \in \mathcal{A}_q\}$. En particulier, on a

$$\hat{\eta}(x) = \hat{a}_{(0,\ldots,0)}(x), \qquad \widehat{\nabla\eta}(x) = (\hat{a}_{e_1}(x), \ldots, \hat{a}_{e_p}(x))^T.$$

Les estimateurs $\hat{a}_{\underline{k}}$ sont déterminés par un algorithme type IRLS avec une matrice de plan d'expérience et une matrice de poids appropriée. Par ailleurs, Fan et Gijbels (1996) décrivent plusieurs méthodes pour déterminer la taille de fenêtre.
Le problème de la méthode des polynômes locaux même en petite dimension est celui du "fléau de la dimension". Ce problème se réfère au fait qu'un voisinage local en grande dimension n'est plus vraiment local. En effet un voisinage avec un pourcentage de points donné peut se révéler très grand en dimension supérieure à un. Une façon de contourner ce problème consiste à réduire la dimension. Nous verrons à la section 4, une approche basée sur le modèle en indice simple.

# 3 Partial Least Squares pour les modèles linéaires généralisés

La première méthode de réduction de dimension envisagée est basée sur PLS. Nous motivons tout d'abord ce choix et justifions la nécessité d'étendre PLS aux modèles linéaires généralisés ; nous proposons ensuite une extension particulièrement adaptée à l'analyse de données qui, de par leurs dimensions, sont caractérisées par une très forte multicolinéarité des régresseurs.
Dans cette section 3, la matrice $\mathbf{X}^{(1)}$ est supposée standardisée : chaque colonne est centrée de norme 1. Nous reviendrons en Section 3.2.6 sur les implications de ce choix.

## 3.1 Partial Least Squares : PLS$[\Upsilon, \mathbf{Z}^{(G)}, \mathbf{W}, \kappa]$

### 3.1.1 L'usage classique de PLS

PLS est une technique de réduction de dimension et de régression dans les modèles linéaires (cf. Helland (1988) et références citées). Dans Stone et Brooks (1990), on trouve une approche qui englobe dans la même procédure nommée "continuum regression" les moindres carrés, PLS et ACP.
Soit un vecteur de variable de réponse $\Upsilon \in \mathbb{R}^n$, une matrice de plan d'expérience standardisée $\mathbf{X}^{(1)}$ de taille $n \times p$ et une matrice de poids $W$ définie positive de taille $n \times n$. Le but de PLS est d'abord d'exprimer la relation entre $\Upsilon$ et $\mathbf{X}^{(1)}$ à travers la définition de $\kappa$ vecteurs de $\mathbb{R}^n$, $(t_j)_{1\leq j\leq\kappa}$. Ces vecteurs sont choisis dans l'espace engendré par $\mathbb{I}_n$ et les colonnes de $\mathbf{X}^{(1)}$ (*i.e.*

engendré par les colonnes de $\mathbf{Z}^{(1)}$) ; l'introduction du vecteur $\mathbb{1}_n$ correspond à l'introduction d'un terme constant dans le modèle de régression linéaire. Par construction, $\mathbb{1}_n^T W t_j = 0$ pour tout $j$ et les scores PLS $(t_j)_j$ sont deux à deux $W$-orthogonaux. On réalise ensuite une régression $W$-pondérée de $\Upsilon$ sur $(\mathbb{1}_n, t_1, \ldots, t_\kappa)$, ce qui conduit à une décomposition de la forme $\Upsilon = q_0 \mathbb{1}_n + q_1 t_1 + \cdots + q_\kappa t_\kappa + f_{\kappa+1}$ où $(q_j)_{0 \leq j \leq \kappa}$ sont des réels, le reste $f_{\kappa+1}$ étant $W$-orthogonal aux vecteurs $(\mathbb{1}_n, t_1, \cdots, t_\kappa)$.
L'estimateur du coefficient de régression, $\hat{\theta}^{\text{PLS},\kappa}$ est défini comme le vecteur (de norme minimale) vérifiant $\Upsilon - f_{\kappa+1} = \mathbf{Z}^{(1)}\theta$ (Fort (2005)).
Contrairement à d'autres méthodes de réduction de dimension type ACP par exemple, les scores $t_j$ dépendent du vecteur réponse $\Upsilon$. En effet, étant donnés $(t_j)_{1 \leq j \leq k}$, $t_{k+1}$ est la combinaison linéaire des colonnes de $\mathbf{Z}^{(1)}$, $t_{k+1} = \mathbf{Z}^{(1)} \upsilon$, telle que

$$\upsilon = \text{argmax}_{\{\upsilon, \upsilon^T \upsilon = 1\}} \left| \text{Cov}(\sqrt{W} \mathbf{Z}^{(1)} \upsilon, \sqrt{W} f_{k+1}) \right| \ ; \tag{11}$$

*i.e.* $t_{k+1}$ maximise la covariance empirique pondérée sous la contrainte $\upsilon^T \upsilon = 1$ ($\sqrt{W}$ désigne la racine carrée de la matrice définie-positive $W$). En quelque sorte, $t_{k+1}$ est le vecteur qui contient le plus d'information sur $f_{k+1}$.
Le nombre maximal de scores PLS, $\kappa_{\max}$, est inférieur ou égal au rang de $\mathbf{X}^{(1)}$ (Helland (1990)). En pratique, il est bien souvent égal à ce rang. Lorsque $\kappa = \kappa_{\max}$, les vecteurs $(\mathbb{1}_n, t_1, \cdots, t_{\kappa_{\max}})$ forment une base de l'espace engendré par les colonnes de $\mathbf{Z}^{(1)}$ qui contient la $W$-projection de $\Upsilon$ sur les colonnes de $\mathbf{Z}^{(1)}$. Ainsi PLS appliqué avec $\kappa = \kappa_{\max}$ conduit à l'estimateur des moindres carrés $W$-pondéré.
Dans la littérature, PLS est usuellement introduit avec $W = \text{Id}_n$. Nous détaillons ci-après la procédure PLS dans le cas pondéré, procédure que nous noterons par la suite PLS$[\Upsilon, \mathbf{Z}^{(1)}, W, \kappa]$. L'introduction d'une matrice de poids $W$ généralise l'usage de PLS pour la régression dans les modèles hétéroscédastiques.

(i) $E_0 = \mathbf{X}^{(1)} - \mathbb{1}_n [\mathbb{1}_n^T W \mathbb{1}_n]^{-1} \mathbb{1}_n^T W \mathbf{X}^{(1)}$ ; $f_0 = \Upsilon - \mathbb{1}_n [\mathbb{1}_n^T W \mathbb{1}_n]^{-1} \mathbb{1}_n^T W \Upsilon$.

(ii) Pour $j = 1, \cdots, \kappa$,

$$t_j = E_{j-1} E_{j-1}^T W f_{j-1}, \qquad q_j = t_j^T W f_{j-1} / (t_j^T W t_j),$$
$$E_j = E_{j-1} - t_j t_j^T W E_{j-1} / (t_j^T W t_j), \qquad f_j = f_{j-1} - q_j t_j.$$

Observons que cet algorithme est inchangé si la matrice $W$ est remplacée par $sW$, pour un scalaire $s$ strictement positif.

Dans le cas de variables réponses vectorielles, deux approches sont possibles : passer à une représentation bloc de $\mathbf{Y}^{(G)}$, ou modifier PLS. Classiquement c'est la première approche qui est retenue et la procédure PLS consiste à trouver des combinaisons linéaires des colonnes de $\mathbf{Z}^{(1)}$, $\mathbf{Z}^{(1)} \upsilon$, telles que

$$(\upsilon, \varsigma) = \text{argmax}_{\{(\upsilon, \varsigma), \upsilon^T \upsilon = 1, \ \varsigma^T \varsigma = 1\}} \left| \text{Cov}(\mathbf{Z}^{(1)} \upsilon, \mathbf{Y}^{(G)} \varsigma) \right| \ ;$$

$\upsilon$ est un vecteur de $\mathbb{R}^G$ et $\mathbf{Y}^{(G)}$ est une matrice de taille $n \times G$ dont la $k$-ième ligne est donnée par le $k$-ième bloc de $\mathbf{Y}$ défini en (4).

Puisque
$$\mathrm{Cov}(\mathbf{Z}^{(1)}\upsilon, \mathbf{Y}^{(G)}\varsigma) = \mathrm{Var}(\mathbf{Z}^{(1)}\upsilon)\mathrm{Cor}(\mathbf{Z}^{(1)}\upsilon, \mathbf{Y}^{(G)}\varsigma)\mathrm{Var}(\mathbf{Y}^{(G)}\upsilon),$$
PLS peut être vu comme une analyse canonique "pénalisée". Barker et Rayens (2003) font état de ce fait et proposent une version différente de PLS en remplaçant dans le critère pour définir les composantes PLS, la quantité $\varsigma^T\varsigma$ par $\mathrm{Var}(\mathbf{Y}^{(G)}\varsigma)$.

Nous avons opté pour la seconde approche et décidé de modifier PLS : nous aurons en effet besoin de l'utiliser avec comme matrice de régresseurs, la matrice bloc $\mathbf{Z}^{(G)}$. La modification que nous proposons repose sur l'analyse de l'étape (i) de PLS : elle vise à soustraire les $W$-projections sur $\mathbb{I}_n$, vecteur lié au terme constant dans le modèle. Dans la matrice $\mathbf{Z}^{(G)}$, nous avons $G$ vecteurs colonnes liés aux constantes. Notons $C$ la matrice constituée de ces $G$ vecteurs. Nous soustrayons les $W$-projections sur l'espace engendré par ces vecteurs et pour ce faire, nous remplaçons dans l'étape (i) $\mathbb{I}_n$ par $C$ :
$$E_0 = \mathbf{X}^{(G)} - C[C^T\mathbf{W}C]^{-1}C^T\mathbf{W}\mathbf{X}^{(G)} \; ; \qquad f_0 = \Upsilon - C[C^T\mathbf{W}C]^{-1}C^T\mathbf{W}\Upsilon.$$

Cependant, toutes ces procédures type PLS (même dans le cas pondéré) ne restent développées que pour des variables réponses continues. Appliquer directement ces méthodes à des données binaires ou catégorielles ne semble pas raisonnable. Nous décrivons dans la sous-section qui suit quelques tentatives pour étendre PLS aux modèles linéaires généralisés.

### 3.1.2 L'usage de PLS dans les modèles linéaires généralisés

Marx (1996); Nguyen et Rocke (2002a,b, 2004) et Ding et Gentleman (2005) ont utilisé PLS comme technique de réduction de dimension dans des modèles de régression logistique en grande dimension $n << p$ ; Bastien et al. (2004) proposent une extension de PLS à la régression linéaire généralisée dans le cadre standard $n > p$ et Bastien (2004) applique cette extension au modèle de Cox, lorsque $n << p$. On trouvera aussi une description algorithmique de toutes ces méthodes dans Fort (2005).
Fort et Lambert-Lacroix (2005) ont montré que les méthodes de Marx (1996), Nguyen et Rocke (2002a) et Ding et Gentleman (2005) appliquées aux modèles de régression logistique dichotomiques reposent sur des algorithmes itératifs qui, très souvent, ne convergent pas. Ces analyses ont ensuite été détaillées, étendues au cas de l'application aux modèles de régression logistique polytomiques (Nguyen et Rocke (2002b); Ding et Gentleman (2005)) et complétées par l'analyse de l'algorithme de Bastien et al. (2004), par Fort (2005). Appliquées à l'inférence dans les modèles de régression logistique, toutes ces méthodes, à l'exception de celle de Ding et Gentleman (2005), se heurtent au problème de "séparation" ; en pratique, les algorithmes itératifs sont arrêtés dès que le phénomène de séparation est détecté, mais l'estimateur qui en résulte est alors dépendant du point initial de la procédure itérative. L'estimateur proposé par Ding et Gentleman (2005) dépend dans certains cas du nombre maximal d'itérations autorisé puiqu'après une période "de chauffe", l'algorithme itératif peut avoir un comportement cyclique.

Fort (2005) a étudié la robustesse de ces algorithmes lorsqu'ils sont appliqués à la classification (binaire et multiclasse) de puces à ADN par discrimination logistique et à la sélection de gènes par une approche type *Recursive Feature Elimination* (RFE, Guyon et al. (2002)). Les méthodes de Marx (1996); Bastien et al. (2004) sont apparues très instables : non convergence des étapes itératives les constituant, taux d'erreur de classification relativement élevés par rapport aux autres extensions de PLS. Les méthodes de Nguyen et Rocke (2002a,b) sont plus ou moins robustes à la non-convergence des étapes IRLS (cf. Section 2.3.2) le constituant : sur les jeux de données considérés, et pour une valeur $\kappa$ du nombre de composantes PLS fixée, les taux d'erreur de classification sont stables et dépendent peu de l'initialisation de IRLS ; tandis que la méthode de sélection de variables est elle très sensible à l'initialisation. Enfin, l'estimateur des coefficients de régression proposé par Ding et Gentleman (2005) n'est pas unique mais il est apparu que le taux d'erreur de classification et l'ensemble des gènes sélectionnés par la méthode RFE ne dépendaient pas du choix du vecteur dans l'ensemble des points limites.
Un des rapporteurs a porté à notre attention l'extension de Nguyen et Rocke (2004). Ces derniers ont noté que les composantes PLS sont des combinaisons linéaires des régresseurs initiaux, et que les coefficients de la combinaison linéaire s'expriment à l'aide du coefficient de régression (linéaire) de la variable de réponse sur chacun des régresseurs. Ils proposent donc de mimer la définition et de remplacer la régression linéaire par une régression logistique. Deux algorithmes sont alors possibles en grande dimension $n << p$ : appliquer la méthode sur la matrice de plan d'expérience initiale, ou appliquer la méthode sur une matrice de rang plein en colonnes, déduite de la décomposition en valeurs singulières de $\mathbf{X}^{(1)}$. Cette extension est très proche de celle de Bastien et al. (2004), et sa mise en œuvre se heurte aussi au problème de séparation.
Dans les applications considérées en Section 5, nous mettrons en œuvre une méthode de discrimination logistique lorsque le vecteur de régression est estimé par l'algorithme de Nguyen et Rocke (2002a) (cas binaire) et Nguyen et Rocke (2002b) (cas multiclasse), et par l'algorithme de Ding et Gentleman (2005) (cas binaire et multiclasse). On trouvera donc en Section 5.3.1 une description plus détaillée de ces algorithmes.

## 3.2 L'algorithme Ridge-PLS

### 3.2.1 Pénalisation et réduction de dimension

Les données de biopuces souffrent de deux fléaux : le fléau de la dimension et le fléau de la multicolinéarité des régresseurs. Ridge-PLS repose sur l'algorithme PLS pour pallier le premier fléau, et sur un critère de maximum de vraisemblance pénalisé pour répondre au second.
Le développement de cet algorithme repose sur la nécessité suivante : remplacer le vecteur des variables de réponses $\mathbf{Y}$ par un vecteur $\Upsilon$ qui *(i)* est à valeur continue et reprend l'information contenue dans $\mathbf{Y}$ et *(ii)* a une structure de modèle linéaire ; puis appliquer PLS avec cette nouvelle variable. Pour ce faire, observons que dans le cadre inférentiel classique, l'estimateur $\hat{\theta}^{\mathrm{MV}}$ est la régression pondérée au sens moindres carrés de $\psi^{\infty}$, la pseudo-variable à convergence de IRLS, sur les colonnes de $\mathbf{Z}^{(G)}$ (Section 2.3.2). De plus, $\psi^{\infty}$ est de la forme $\mathbf{Z}^{(G)}\theta^{\infty} + \epsilon$ avec $\epsilon = [\mathbf{W}(\theta^{\infty})]^{-1}(\mathbf{Y} - \boldsymbol{\mu}(\theta^{\infty}))$ ; d'après (7), ce terme peut s'interpréter comme un terme de bruit

de matrice de dispersion $\phi\ \mathbf{W}(\theta^\infty)^{-1}$ (cette lecture est néanmoins peu rigoureuse). En ce sens, $\psi^\infty$ possède les propriétés recherchées et l'idée de Ridge-PLS est d'appliquer PLS avec $\Upsilon = \psi^\infty$ et $\phi^{-1}\mathbf{W}(\theta^\infty)$ pour matrice de poids (ou tout simplement $\mathbf{W}(\theta^\infty)$, PLS étant inchangé si la matrice de poids est multipliée par un scalaire).
Cette heuristique doit néanmoins être robustifiée : nous avons en effet souligné en Section 2.3.4 que dans certains cas, et en particulier dans le cas de la régression logistique appliquée aux données de biopuces, la vraisemblance est maximale lorsque $\|\theta\| \to +\infty$. En conséquence, $\lim_t \|\psi^t\| = +\infty$ et $\psi^\infty$ n'existe pas. Nous remplaçons donc le critère de maximisation de vraisemblance, par un critère de maximisation sous contraintes, la contrainte portant sur la norme quadratique du paramètre - à l'exception des coefficients relatifs aux termes constants -

$$\text{argmax}\ \ell_*(\theta), \qquad \text{où} \quad \ell_*(\theta) = \ell(\theta) - 0.5\lambda\ \theta^T \mathbf{R}_{G,p+1}\theta; \tag{12}$$

$\mathbf{R}_{s_1,s_2}$ est la matrice diagonale d'ordre $s_1 s_2 \times s_1 s_2$ de diagonale égale à $s_1$ réplications du vecteur $(0, 1, \cdots, 1) \in \mathbb{R}^{s_2}$. Il est aisé de vérifier que la recherche du maximum de $\ell_*$ par la méthode de Newton-Raphson s'interprète comme une suite de Ridge-régressions pondérées sur les colonnes de $\mathbf{Z}^{(G)}$ d'une pseudo-variable $\psi_*^t$ de même expression que $\psi^t$ (8).
L'algorithme Ridge-PLS est donc un algorithme PLS pondéré appliqué avec $\Upsilon = \psi_*^\infty$, la pseudo-variable à convergence d'un algorithme de Newton-Raphson pour la maximisation de la vraisemblance pénalisée $\ell_*$.

### 3.2.2 Ridge-PLS : RPLS$[\mathbf{Y}, \mathbf{Z}^{(G)}, \mathbf{W}, \kappa, \lambda]$

Soient $\lambda$ un réel strictement positif et $\kappa$ un entier naturel.

Étape A : Maximisation itérative de $\theta \mapsto \ell_*(\theta)$

Initialisation : $\theta^0$.

Tant que non-convergence,

$$\psi_*^t = \mathbf{Z}^{(G)}\theta^t + \mathbf{W}(\theta^t)^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}(\theta^t)\right),$$

$$\theta^{t+1} = \left(\mathbf{Z}^{(G)T}\mathbf{W}(\theta^t)\mathbf{Z}^{(G)} + \lambda \mathbf{R}_{G,p+1}\right)^{-1} \mathbf{Z}^{(G)T}\mathbf{W}(\theta^t)\psi_*^t.$$

Fin.

Étape B : Réduction de dimension

Appeler PLS $[\psi_*^\infty, \mathbf{Z}^{(G)}, \mathbf{W}(\theta^\infty), \kappa]$ où $\psi_*^\infty = \lim_t \psi_*^t$ et $\theta^\infty = \lim_t \theta^t$.

Retourner $\hat{\theta}^{\text{FL},\lambda,\kappa} = \hat{\theta}^{\text{PLS},\kappa}$.

Cet algorithme est invariant par reparamétrisation : les composantes de $\theta$ ont une interprétation physique et en ce sens, $\theta$ est le paramètre "naturel" du problème. Néanmoins, lorsque la matrice $\mathbf{X}^{(1)}$ n'est pas de rang plein en colonnes, le modèle de régression logistique peut être paramétré par $\gamma \in \mathbb{R}^{G(\text{rang}(\mathbf{X}^{(1)})+1)}$ de sorte que la matrice de plan d'expérience est de rang plein en colonnes ; RPLS peut alors être implanté dans un espace de dimension $G(\text{rang}(\mathbf{X}^{(1)}) + 1)$ ($<< G(p+1)$) pour produire un estimateur $\hat{\gamma}^{\text{FL},\lambda,\kappa}$, lié, de façon unique à $\hat{\theta}^{\text{FL},\lambda,\kappa}$. Cette reparamétrisation, qui permet une accélération drastique de l'algorithme, est détaillée dans Fort et

Lambert-Lacroix (2005).
Nous suggérons de fixer $\theta^0$ à la valeur initiale habituellement conseillée pour l'initialisation de IRLS (Cf. section 2.3.4).

### 3.2.3 Existence et unicité de l'estimateur $\hat{\theta}^{\mathrm{FL},\lambda,\kappa}$

La fonction $\ell_*$ est coercive et son hessien est partout défini-négatif, ce qui garantit l'existence et l'unicité d'un maximum global. Toutes les suites convergentes $(\theta^t)_t$ ont donc même point limite $\theta^\infty$ et par suite, les quantités $\psi_*^\infty$ et $\mathbf{W}(\theta^\infty)$ existent et sont uniques. Enfin, l'unicité de l'estimateur PLS garantit l'existence et l'unicité de l'estimateur $\hat{\theta}^{\mathrm{FL},\lambda,\kappa}$.

### 3.2.4 Choix du paramètre de régularisation $\lambda$

Lorsque $\lambda = 0$, $\ell_* = \ell$ et l'étape A peut ne jamais converger ; ce qui explique la condition $\lambda > 0$.
Le cas $\lambda \to +\infty$ revient à maximiser $\ell$ sous la contrainte que les coefficients de régression autres que ceux liés aux termes constants sont nuls : on nie donc la notion de régression et $\lim_t \theta^t$ est l'estimateur du maximum de vraisemblance dans un modèle exponentiel. Dans le cas particulier de la régression logistique dichotomique, on peut vérifier que $\hat{\theta}_{2:p+1}^{\mathrm{FL},+\infty,\kappa}$ est, à une constante positive multiplicative près, égal à l'estimateur $\hat{\theta}_{2:p+1}^{\mathrm{PLS},\kappa}$ obtenu en appliquant PLS avec $\Upsilon = \mathbf{Y}$ (Fort (2005)). En conséquence, les valeurs extrêmes $\lambda = 0$ et $\lambda \to +\infty$ ne sont pas recommandées.
On peut distinguer deux familles de méthodes pour déterminer le paramètre de régularisation : les méthodes par validation croisée, relativement coûteuses en temps de calcul, et les méthodes basées sur des critères d'information, plus simples à mettre en œuvre. Nous proposons d'exploiter la structure de Ridge-régression de l'étape A et de fixer $\lambda$ à la valeur minimisant le critère :

$$-2\ell(\theta) + \log(n)\ \mathrm{Trace}\left(\sqrt{\mathbf{W}}(\theta)\mathbf{Z}^{(G)}({\mathbf{Z}^{(G)}}^T\mathbf{W}(\theta)\mathbf{Z}^{(G)} + \lambda\mathbf{R}_{G,p+1})^{-1}{\mathbf{Z}^{(G)}}^T\sqrt{\mathbf{W}}(\theta)\right), \qquad (13)$$

calculé en la limite $\theta = \theta^\infty$, une limite qui dépend de $\lambda$. Ce critère mime le critère BIC (Schwarz (1978)) et évalue le degré de liberté du modèle par une mesure de la sensibilité de la prédiction aux (pseudo)-observations : à convergence de l'étape A, $\mathbf{Z}^{(G)}\theta^\infty = \mathcal{S}\psi^\infty$ et Trace($\mathcal{S}$) fournit une mesure de cette sensibilité (Hastie et Tibshirani (1990)). Nous avons choisi un critère type BIC plutôt que le critère type AIC par exemple, car BIC est consistant en situation asymptotique.
Dans la méthode que nous préconisons ici, $\lambda$ est choisi avant l'étape de réduction de dimension par un critère qui ne fait intervenir que $\theta^\infty$ ; l'étape de régularisation est donc indépendante de l'étape de réduction de dimension et donc de $\kappa$, et cela facilite la détermination de $\lambda$. Les fonctions de coût utilisées habituellement en validation croisée nécessitent la connaissance $\hat{\theta}^{\mathrm{FL},\lambda,\kappa}$ si bien que $\lambda$ dépend de $\kappa$ ; cela rend la mise en œuvre de la validation croisée beaucoup plus lourde puisqu'elle doit être couplée avec la méthode de détermination de $\kappa$.
Ridge-PLS appliqué avec cette valeur optimale de $\lambda$ définit un estimateur noté $\hat{\theta}^{\mathrm{FL},\lambda_*,\kappa}$.

### 3.2.5 Choix du nombre de composantes PLS, $\kappa$

La dépendance de l'estimateur $\hat{\theta}^{\mathrm{PLS},\kappa}$ en les variables de réponse $\Upsilon$ n'est pas linéaire et est relativement complexe, ce qui rend l'étude de cet estimateur difficile. A notre connaissance, il n'existe pas de résultats exacts et exploitables sur l'évolution du biais, de la variance, de l'erreur quadratique moyenne de prédiction, $\cdots$, en fonction de $\kappa$. Citons néanmoins quelques travaux, établis dans le cadre des modèles linéaires lorsque $W = \mathrm{Id}_n$ et la matrice de plan d'expérience est de rang plein en colonnes. Goutis (1996); De Jong (1995) ont montré que $\kappa \mapsto \|\hat{\theta}^{\mathrm{PLS},\kappa}\|$ est une application croissante et donc que pour tout $\kappa \leq \kappa_{\max}$, $\|\hat{\theta}^{\mathrm{PLS},\kappa}\| \leq \|\hat{\theta}^{\mathrm{PLS},\kappa_{\max}}\|$ où $\hat{\theta}^{\mathrm{PLS},\kappa_{\max}}$ coïncide avec l'estimateur des moindres carrés. Lingjaerde et Christophersen (2000) donnent une expression des composantes de $\hat{\theta}^{\mathrm{PLS},\kappa}$ sur la base des vecteurs propres de la matrice de corrélation des régresseurs, en fonction des facteurs de Ritz ; ils en déduisent des propriétés de contraction de cet estimateur dans ces directions propres. Stoica et Söderström (1998) ont étudié le biais et la variance de l'estimateur dans un cadre asymptotique défini par un terme de bruit devenant négligeable devant le prédicteur linéaire. Phatak et al. (2002) ont écrit l'expression de la matrice de covariance de l'estimateur. Enfin, Denham (2000) développe différentes approximations de l'erreur quadratique moyenne de prédiction et en déduit une méthode de choix du nombre de composantes PLS. Néanmoins, de l'aveu même de l'auteur, ces méthodes se dégradent lorsque $n << p$ et il conseille de choisir $\kappa$ par validation croisée ou par bootstrapping.
Dans les applications considérées ci-après, nous choisissons $\kappa$ par validation croisée (cf. section 5.3.3).

### 3.2.6 Standardisation de la matrice $\mathbf{X}^{(1)}$

Nous avons supposé la matrice $\mathbf{X}^{(1)}$ standardisée et par définition, $\hat{\theta}^{\mathrm{FL},\lambda,\kappa}$ est l'estimateur des coefficients de régression par rapport aux colonnes de $\mathbf{X}^{(1)}$ standardisée. L'estimateur $\breve{\theta}^{\mathrm{FL},\lambda,\kappa}$ des coefficients de régression relatifs aux $p$ covariables initiales est donc donné par $\breve{\theta}^{\mathrm{FL},\lambda,\kappa}_{k+1} = s_k^{-1}\ \hat{\theta}^{\mathrm{FL},\lambda,\kappa}_{k+1}$ où $s_k^2$ est la variance empirique du $k$-ième régresseur.
Le calcul de l'estimateur du maximum de vraisemblance dans les modèles (linéaires) généralisés est stable par standardisation - lorsque cet estimateur existe - : l'estimateur $\hat{\theta}^{\mathrm{MV}}$ obtenu en appliquant l'algorithme IRLS avec la matrice $\mathbf{X}^{(1)}$ standardisée est lié à $\breve{\theta}^{\mathrm{MV}}$ obtenu avec la matrice de plan d'expérience non standardisée par la relation $\breve{\theta}^{\mathrm{MV}}_{k+1} = s_k^{-1}\ \hat{\theta}^{\mathrm{MV}}_{k+1}$, si bien que l'estimateur du prédicteur linéaire est invariant par standardisation.
Aucune des deux étapes de l'algorithme Ridge-PLS n'est stable par standardisation de la matrice de plan d'expérience, et opter ou pas pour la standardisation dépend donc du problème que l'on cherche à résoudre. Pour Ridge-PLS, discuter de la standardisation de la matrice de plan d'expérience est équivalent à discuter *(i)* du choix de la norme dans le critère quadratique de pénalisation Ridge (12), et *(ii)* de la contrainte quadratique dans la construction des composantes PLS (11). Considérons le cas extrême où il existe deux régresseurs $(i, j)$ dont les réalisations diffèrent d'un même facteur d'échelle $a$ *i.e.* $\mathbf{X}^{(1)}_{:,i} = a\mathbf{X}^{(1)}_{:,j}$. Le pouvoir discriminant de ces deux covariables, mesuré par exemple par la statistique ANOVA, est le même. L'algorithme Ridge-PLS tel que décliné ci-dessus donne à ces deux variables la même importance dans l'expression du

prédicteur linéaire et en particulier, donne à ces deux variables le même poids dans un processus de discrimination logistique.

# 4 Approche semi-paramétrique

La deuxième méthode de réduction de dimension envisagée est basée sur une approche semi-paramétrique utilisant les modèles linéaires généralisés en indice simple. Après avoir introduit ces modèles, nous donnons la méthode d'estimation utilisant la vraisemblance locale. Nous rappelons que dans cette section, $G = 1$. Nous supposons de plus que le modèle est suffisamment régulier pour que la fonction de lien $g$ soit un $C^1$-difféomorphisme.

## 4.1 Modèles linéaires généralisés en indice simple

Une méthode répandue pour venir à bout du problème de dimension consiste à projeter le vecteur de covariables $X$ sur un espace linéaire engendré par les covariables et ensuite à ajuster une courbe non paramétrique à ces combinaisons linéaires. Ce principe conduit aux modèles en indice simple :

$$Y = \tilde{\mu}(\beta^T X) + \epsilon$$

où $\mathbb{E}[\epsilon|X] = 0$ presque sûrement. Par définition des modèles linéaires généralisés, $Y = \mu(X) + \epsilon$ ; les modèles en indice simple supposent l'existence d'une direction $\beta \in \mathbb{R}^p$ et d'une fonction $\tilde{\mu} : \mathbb{R} \mapsto \mathbb{R}^G$ telle que $\mu(X) = \tilde{\mu}(\beta^T X)$.

Notons que l'échelle de $\beta^T X$ dans $\tilde{\mu}(\beta^T X)$ peut être choisie de façon arbitraire : pour tout $u > 0$, $(\beta, \tilde{\mu}(\cdot))$ et $(u\beta, \tilde{\mu}(\cdot/u))$ conduisent à la même fonction de régression. Pour lever cette indétermination, posons $\beta = \mathbb{E}[\nabla\mu(X)]$ le vecteur des dérivées moyennes. Nous avons :

$$\mathbb{E}[\nabla\mu(X)] = \mathbb{E}\left[\nabla\{\tilde{\mu}(\beta^T X)\}\right] = \mathbb{E}\left[\{\nabla\tilde{\mu}\}(\beta^T X)\right]\beta = c\beta.$$

## 4.2 Méthode d'estimation : Algorithme GSIM

On veut donc estimer $\beta$ et $\tilde{\eta} = g(\tilde{\mu})$. Notons que, puisque $\beta = \mathbb{E}[\nabla\mu(X)]$ et $\mu = g^{-1}(\eta)$, nous avons

$$\beta = \mathbb{E}\left[(g^{-1})'(\eta(X))\ \nabla\eta(X)\right].$$

L'idée développée ici est d'estimer $\eta$ et $\nabla\eta$ par la méthode des polynômes locaux à partir de la fonction de vraisemblance conditionnelle (1). A ce stade nous n'utilisons pas la structure en indice simple de la fonction de régression. On estime alors $\beta$ par la moyenne empirique des variables $(g^{-1})'(\hat{\eta}(X_i))\ \widehat{\nabla\eta}(X_i)$. Pour finir, on régresse $Y_i$ sur $\hat{\beta}^T X_i$ par polynômes locaux pour obtenir $\widehat{\tilde{\eta}}$ et $\hat{\mu}(x) = g^{-1}(\widehat{\tilde{\eta}}(\hat{\beta}^T x))$. Plus précisément la procédure GSIM pour estimer $\beta$ et $\tilde{\eta}$ est décrite par l'algorithme suivant.

Étape A : Pour $j = 1, \dots, n$, calculer

$$\hat{\eta}(X_j) = \hat{a}_{(0,\dots,0)}(X_j), \quad \widehat{\nabla\eta}(X_j) = (\hat{a}_{e_1}(X_j) \dots, \hat{a}_{e_p}(X_j))^T,$$

obtenus en maximisant

$$\sum_{i=1}^{n} \mathcal{L}\left(\sum_{\underline{k}\in\mathcal{A}_q} a_{\underline{k}}\ (X_i - X_j)^{\underline{k}}, Y_i\right)\ K_H^p(X_i - X_j), \tag{14}$$

comme fonction de $a_{\underline{k}}$, $\underline{k} \in \mathcal{A}_q$. Poser

$$\hat{\beta} = n^{-1}\ \sum_{i=1}^{n} (g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i).$$

Étape B : Déterminer $\widehat{\eta}(x) = \hat{a}_0$ en maximisant

$$\sum_{i=1}^{n} \mathcal{L}\left(a_0 + a_1\hat{\beta}^T(X_i - x), Y_i\right)\ K_{h_B}^1\left(\hat{\beta}^T(X_i - x)\right),$$

comme fonction de $a_0$ et $a_1$.

La maximisation du critère (14) peut être lue comme la recherche d'un maximum de vraisemblance pondérée, de poids $\left(K_H^p(X_i - X_j)\right)_i$ et de matrice de plan d'expérience particulière, dont le nombre de colonnes augmente avec $q$. En pratique, la maximisation est résolue par un algorithme de type IRLS (Section 2.3.2). Nous avons rappelé en Section 2.3.2 que le maximum n'existe pas nécessairement. En particulier, dans le cadre des modèles de régression logistique, le maximum de vraisemblance ne peut jamais exister si le rang de la matrice de plan d'expérience est égal à $n$ (Section 2.3.4). Dans les applications considérées dans cet article, le rang de la matrice de plan d'expérience est, en pratique, égal à $n$. Pour remédier à ce problème, deux approches sont envisageables. La première consiste à faire une pré-sélection de gènes. La seconde consiste à introduire un terme de régularisation dans le critère (14), par exemple de type Ridge, comme introduit en Section 3. Dans l'application aux biopuces, la première solution conduit à sélectionner trop peu de gènes (au maximum autant que d'individus). Cette réduction drastique du nombre de covariables ne semble pas indiquée dans le cadre d'une approche compression de données.

Nous avons donc opté pour la deuxième solution qui permet de considérer tous les gènes. Par ailleurs, une pénalité de type Ridge dans l'estimation du gradient a été introduite par Seifert et Gasser (1996) dans le cas gaussien avec $p = 1$ et $q = 1$. Cette pénalité permet de résoudre les problèmes de regroupements (ou de zones creuses) dans l'espace des covariables. On s'attend justement à rencontrer des regroupements dans les problèmes de classification. C'est pourquoi même si on disposait d'un grand nombre d'individus, on pense qu'une telle pénalité serait toujours utile. On notera $\hat{\beta}^\lambda$ l'estimateur obtenu avec la procédure GSIM pénalisée.

Rappelons que cette procédure est donnée ici pour $G = 1$. L'extension au cas multiclasse n'est pas immédiate et fait l'objet de travaux en cours.

### 4.3 Implantation et choix des paramètres pour la procédure GSIM

Dans la première étape, il faut fixer l'ordre $q$ de l'approximation polynômiale. En pratique, on choisit un ajustement linéaire ($q = 1$) comme cela est fait dans la méthode OPG proposée par Xia et al. (2002). D'autre part pour réduire les effets du fléau de la dimension, nous utilisons dans l'étape A un noyau produit, ce qui conduit à choisir une matrice $H$ diagonale. Les noyaux considérés sont gaussiens.

Nous rappelons qu'ici $\mathbf{X}^{(1)}$ n'est pas standardisée. Notons $\Sigma^2$ la matrice diagonale dont le $i$-ème terme diagonal est égal à la variance empirique associée à la $i$-ème covariable. Comme il est d'usage dans une pénalité de type Ridge, on utilise la norme induite par $\Sigma^2$ pour les coefficients relevant du gradient. Cela revient à pénaliser le gradient fortement dans les directions les plus variables. La matrice de plan d'expérience standardisée est donnée par $\mathbf{X}_s^{(1)} = (\mathbf{X}^{(1)} - \mathbb{1}_n \mathbb{1}_n^T \mathbf{X}^{(1)}/n)\Sigma^{-1}$. Soit $\hat{\beta}_s^\lambda$ l'estimateur correspondant à $\mathbf{X}_s^{(1)}$ avec une taille de fenêtre $H_s$ et avec dans le terme de pénalité la norme induite par $\mathbf{R}_{1,p+1}$ ; on peut montrer que, pour $H_s = H\Sigma^{-1}$, on a $\hat{\beta}^\lambda = \Sigma^{-1}\hat{\beta}_s^\lambda$. Pour des raisons de stabilité numérique, on effectuera les calculs avec la matrice standardisée. Il est alors naturel de choisir $H_s = h_A \mathrm{Id}_p$, ce qui réduit le nombre d'hyper-paramètres à déterminer.

La procédure GSIM nécessite donc le choix d'un paramètre de lissage à deux niveaux différents. Dans la première étape, on cherche à estimer $\eta$ ainsi que son gradient et la taille de fenêtre $h_A$ doit être optimale pour cet objectif. On doit également déterminer l'hyper-paramètre $\lambda$. On choisit une approche validation croisée simultanément en $h_A$ et $\lambda$. Dans la seconde étape, on cherche à estimer $\tilde{\eta}$ et $h_B$ doit être optimal pour cette tâche. Pour ce choix plus standard on utilisera la méthode de “plug-in” de Fan et Gijbels (1996).

Notons enfin que, comme pour Ridge-PLS, la procédure GSIM est invariante par reparamétrisation utilisant la décomposition en valeurs singulières, permettant ainsi de réduire considérablement le temps de calcul.

## 5 Application à la classification de puces à ADN

Nous considérons l'application des méthodes d'estimation Ridge-PLS et GSIM pour l'analyse discriminante à but décisionnel de données de puces à ADN. Les puces à ADN permettent de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément ; un des enjeux de l'étude statistique de ces données est de classer les profils en fonction de leur différence d'expression.

Les méthodes de régression Ridge-PLS et GSIM peuvent être exploitées à d'autres fins que celle de la classification ; par exemple, celle de la sélection de variables. L'objectif de la sélection de variables est à la fois d'améliorer les performances de prédiction du prédicteur (dans le cas présent de la régression logistique, améliorer les performances du classifieur) et de fournir une meilleure compréhension du processus qui a engendré les observations.

Via l'estimateur du coefficient de régression dans les approches de régression paramétrique, ou la direction $\beta$ dans les approches semi-paramétriques, et via l'identification de super-gènes (les scores PLS ou le vecteur $\beta^T X$), les méthodes Ridge-PLS et GSIM quantifient l'importance des gènes dans le processus de discrimination. Cette information peut être exploitée pour exhiber un modèle parcimonieux, moins complexe et plus facilement interprétable, en adaptant par exemple le principe de l'algorithme *Recursive Feature Elimination* (RFE) proposé par Guyon et al. (2002). RFE est une méthode de sélection de variables qui exploite la forme de la règle de décision d'un classifieur SVM linéaire (cas binaire) ; le vecteur $\theta$ définissant l'hyperplan séparateur quantifie le rôle de chaque gène dans la règle de décision. Tous les gènes disponibles sont introduits dans le modèle ; une analyse SVM linéaire est menée sur ce modèle et les gènes dont le poids $\theta_j$ associé est le plus grand sont retenus pour former le modèle suivant. Ce processus est répété jusqu'à élimination de tous les gènes. Les différents modèles sont ordonnés selon une métrique (Guyon et al., 2002, section 4.2) qui tient compte, entre autre, de la qualité de prédiction du modèle. Cette technique peut être étendue au cadre de l'analyse discriminante par régression paramétrique, en prenant pour vecteur de poids $\theta$, l'estimateur du coefficient de régression. Cette approche a été exploitée par Zhu et Hastie (2004), l'estimateur étant celui du maximum de vraisemblance Ridge-pénalisé ; et par Fort (2005), qui envisage plusieurs estimateurs : l'estimateur Ridge-PLS, l'estimateur NR (Nguyen et Rocke (2002a)) et l'estimateur IRPLSF (Ding et Gentleman (2005)).
Parce que la qualité de prédiction du modèle joue un rôle essentiel dans les méthodes de sélection de variables *i.e.* dans le cadre présent, dans l'extraction des gènes influents dans l'identification de la pathologie, et parce que le problème de classification est un problème à part entière dans l'analyse des données de puces à ADN, nous avons choisi de montrer l'application des méthodes Ridge-PLS et GSIM à la classification de puces à ADN.

Notre objectif est d'illustrer l'intérêt des estimateurs Ridge-PLS et GSIM en étudiant la qualité de la règle de discrimination logistique lorsque le modèle est estimé par l'algorithme Ridge-PLS et par l'algorithme GSIM. En particulier, nous nous attachons à démontrer la robustesse de la méthode de classification induite par ces estimateurs lorsque le nombre de covariables est très grand devant le nombre d'observations. Nous montrerons qu'il n'est pas utile d'inclure une étape de pré-sélection des gènes pour mettre en œuvre nos algorithmes et obtenir des résultats "satisfaisants". Nous pensons que cette absence de pré-sélection rend notre approche particulièrement intéressante et originale *(i)* tout d'abord parce que notre approche peut être l'élément de base d'un classifieur dont la qualité n'est pas dépendante d'une règle de pré-sélection ad hoc et, le plus souvent, unidimensionnelle ; *(ii)* parce que contrairement à des méthodes plus classiques (discrimination linéaire diagonale, discrimination quadratique diagonale, $k$-plus proches voisins, $\cdots$), elle fournit une mesure du rôle de chaque gène du modèle - quelle que soit la taille de ce dernier -, information qui peut être exploitée par exemple pour la sélection de variables, et plus généralement, pour l'analyse discriminante à but explicatif ; *(iii)* et enfin, parce qu'elle répond pleinement au problème statistique posé : proposer une règle de classification qui prend en compte **tous** les gènes disponibles, et ne comporte pas une méthode de réduction (drastique) du nombre de gènes en première étape d'analyse.

## 5.1 Jeux de données

Nous envisageons deux jeux de données, le premier relevant de la discrimination entre deux groupes, et le second de la discrimination multi-groupes.
`Colon`[1] : Il est constitué de 62 profils d'expression issus de deux populations : 40 tissus tumoraux et 22 tissus sains. Chaque profil comporte 2000 niveaux d'expression de gènes. On trouvera dans Alon et al. (1999) une description complète de ces données.
`Leukemia`[2] : Il est constitué de 72 profils d'expressions issus de trois populations : 47 tissus atteints de Leucémie lymphoblastique aiguë (ALL) et 25 tissus atteints de Leucémie myéloïde aiguë (AML). Les 47 tissus ALL se subdivisent en deux populations selon que les cellules analysées sont de type $B$ (38 cas) ou de type $T$ (9 cas), de sorte que l'étude de ce jeu de données est un problème de discrimination multi-groupe ($G = 2$). Chaque profil comporte 7129 niveaux d'expression de gènes. On trouvera dans Golub et al. (1999) une description complète de ces données.

## 5.2 Algorithme de discrimination

### 5.2.1 Pré-traitement

Dans le cadre des données issues des biopuces, il est d'usage de pré-traiter les données "brutes" selon un protocole comportant une étape de seuillage, de filtrage et de transformation logarithmique (Dudoit et al. (2002)). *Seuillage :* les niveaux d'expression sont seuillés et restreints à l'intervalle [100, 16000]. *Filtrage :* les gènes dont le niveau d'expression est trop uniforme ne sont pas inclus dans le modèle ; pour chaque gène, on relève la valeur minimale $s_{\min}$ et maximale $s_{\max}$ du niveau d'expression parmi les tissus disponibles, et l'on ne garde que les gènes tels que $s_{\max}/s_{\min} > 5$ et $s_{\max} - s_{\min} > 500$. *Transformation logarithmique :* une transformation logarithmique en base 10 est enfin appliquée.
Le pré-traitement est partie intégrante du processus de discrimination : l'étape de *filtrage* dépend des échantillons disponibles *i.e.* constituant le fichier d'apprentissage. Les gènes retenus pour l'analyse discriminante dépendent donc de la subdivision du jeu de données en un fichier d'apprentissage et un fichier test. Ce nombre reste néanmoins très supérieur au nombre d'observations : dans les analyses envisagées ci-après, le nombre de régresseurs pour `Colon` est supérieur à 1000 ; pour `Leukemia`, il est supérieur à 3000.

Pour un même jeu de données, le pré-traitement diffère selon les contributions. Celui détaillé ci-dessus est celui préconisé par Dudoit et al. (2002) pour les données de biopuces obtenues selon la technologie *Affymetrix high-density oligonucleotide* (ce qui est le cas des deux jeux de données `Colon` et `Leukemia`). Dudoit et al. (2002) complètent ce pré-traitement par une standardisation des biopuces (chaque biopuce est centrée, de norme 1) et motivent cette étape par le fait que de nombreuses méthodes de mesure de similarité de profils d'expression reposent sur la corrélation.

[1] `http ://microarray.princeton.edu/oncology/affydata/index.html`
[2] `http ://www.broad.mit.edu/cancer/software/genepattern/datasets/`

Cette standardisation peut aussi être comprise comme une normalisation des biopuces, utile pour débiaiser les résultats. Elle est donc, par essence même, très différente de la standardisation en colonnes des biopuces, qui est elle motivée par des critères de bonne identifiabilité et de bon comportement de la régression.
Nous étudierons en section 5.3.5, la robustesse des classifieurs à cette standardisation des biopuces.

### 5.2.2 Discrimination logistique dicho/polytomique

Nous posons un modèle de régression logistique pour modéliser la dépendance entre la variable de classe et les gènes retenus à l'issue de l'étape précédente. Pour une estimation $\hat{\eta} \in \mathbb{R}^G$ du prédicteur, la règle de décision est d'affecter un individu caractérisé par un profil d'expression $x$ à la population la plus probable ce qui, d'après (3), est équivalent à la règle de décision

$$\text{Population } i \qquad \text{ssi} \qquad [\hat{\eta}_i(x) \geq \hat{\eta}_l(x), \quad \forall l \in \{0, \cdots, G\}] \,, \tag{15}$$

où par convention, $\hat{\eta}_0$ est le vecteur nul.

## 5.3 Étude de la qualité des classifieurs `RPLS` et `GSIM`

Nous étudions la qualité des classifieurs définis en Section 5.2 lorsque dans (15) $\hat{\eta}$ est l'estimateur obtenu par l'algorithme `RPLS` selon la relation $\hat{\eta}(x) = [1 \ x^T]\hat{\theta}^{\mathrm{PLS},\lambda_*,\kappa}$ ; et lorsque $\hat{\eta}$ est l'estimateur obtenu par l'algorithme `GSIM`. Comme la procédure `GSIM` n'est pas disponible pour le moment dans le cas multiclasse, elle ne sera pas appliquée au jeu de données `Leukemia`.

### 5.3.1 Autres classifieurs envisagés

Nous comparons tout d'abord ces méthodes à des règles de discrimination "classiques", telles que la règle de discrimination linéaire diagonale (`DLDA`) et quadratique diagonale (`DQDA`) ; et la règle des $k$-plus proches voisins pour la métrique euclidienne classique (`KNN`). Nous envisageons ensuite la méthode de discrimination logistique lorsque le prédicteur $\hat{\eta}$ est obtenu par d'autres méthodes de régression paramétriques et semi-paramétriques.
► *Classifieur* `Ridge` : Pour illustrer l'intérêt de la combinaison d'une étape de réduction de dimension et d'une étape de régularisation, fondements de notre algorithme Ridge-PLS, nous évaluons le classifieur `Ridge` proposé par Zhu et Hastie (2004) : il est défini par la règle (15) appliquée avec $\hat{\eta}(x) = [1 \ x^T]\theta^{\mathrm{R}}$, où $\theta^{\mathrm{R}}$ désigne l'estimateur du maximum de vraisemblance Ridge-pénalisé (et coïncide donc avec le vecteur $\theta^{\infty}$ obtenu en fin de l'étape A de l'algorithme Ridge-PLS, Section 3.2.2). Zhu et Hastie (2004) détermine le paramètre de régularisation $\lambda$ par validation croisée, avec un coût basé sur la déviance. Nous avons choisi de déterminer $\lambda$ comme indiqué en Section 3.2.4 de sorte que la comparaison de `Ridge` et `RPLS` permet d'illustrer l'apport de l'étape de réduction de dimension de `RPLS`. Nous commenterons au paragraphe 5.3.5 la

pertinence de la méthode de sélection de $\lambda$ proposée par Zhu et Hastie (2004).
► *Classifieur* `NR` *et* `MNR` : Nous évaluons la règle (15) appliquée avec $\hat{\eta}(x) = [1\ x^T]\hat{\theta}^{\mathrm{NR}}$, où $\hat{\theta}^{\mathrm{NR}}$ est l'estimateur `NR` (resp. `MNR`) proposé par Nguyen et Rocke (2002a) dans le cas dichotomique (resp. Nguyen et Rocke (2002b) dans le cas polytomique). Cet estimateur est obtenu par régression logistique du vecteur des observations à valeur dans $\{0,1\}^G$ sur $\kappa$ super-covariables définies comme les $\kappa$ premières composantes PLS. Dans le cas binaire ($G=1$), ces covariables sont obtenues en appliquant $\mathrm{PLS}[\mathbf{Y}, \mathrm{Id}_n, \mathbf{Z}^{(G)}, \kappa]$ ; dans le cas multiclasse ($G>1$), ces covariables sont obtenues en appliquant un algorithme PLS multivarié (Garthwaite (1994)). Comme nous l'avons mentionné en Section 3.1.2, $\hat{\theta}^{\mathrm{NR}}$ n'existe pas toujours car le cas de séparation est très fréquent dans l'estimation du modèle de régression logistique. Cf. Fort et Lambert-Lacroix (2005) (resp. Fort (2005)) pour une description de ces estimateurs dans le cas de la régression dichotomique (resp. dicho et polytomique) et l'étude de leur existence.
► *Classifieurs* `IRPLSF` *et* `MIRPLSF` : Nous évaluons la règle (15) appliquée avec $\hat{\eta}(x) = [1\ x^T]\hat{\theta}^{\mathrm{DG}}$, où $\hat{\theta}^{\mathrm{DG}}$ est l'estimateur `IRPLSF` (resp. `MIRPLSF`) proposé par Ding et Gentleman (2005) dans le cas dichotomique (resp. polytomique). Cet estimateur combine régularisation et réduction de dimension, respectivement basés sur une pénalité Firth et sur PLS ; Ding et Gentleman robustifient la méthode proposée par Bull et al. (2001) dans le cas $n > p$ et $n$ petit, pour la rendre applicable au cas $n << p$. La démarche algorithmique est la suivante : la maximisation du critère de maximum de vraisemblance Firth-pénalisé peut être résolue par un algorithme itératif type IRLS (section 2.3.2) en adaptant la définition de la pseudo-variable $\psi^t$. Ding et Gentleman (2005) proposent de substituer la régression par une régression type PLS avec $\kappa$ composantes ; leur estimateur est "la" limite de cette méthode itérative. Comme nous l'avons mentionné en section 3.1.2, la limite n'est pas unique. (Cf. Fort et Lambert-Lacroix (2005) et Fort (2005) pour une description des estimateurs `IRPLSF` (resp. `IRPLSF` et `MIRPLSF`), et l'étude de leur existence).
► *Classifieur* `OPG` : Xia et al. (2002) proposent la méthode OPG qui vise à estimer $\kappa$ ($\kappa << p$) directions orthonormales engendrant l'espace estimé de *effective dimension reduction* (EDR). La procédure OPG avec $\kappa = 1$, est analogue à l'étape A de GSIM, mais avec un critère de moindres carrés à la place de la log-vraisemblance (*i.e.* ce qui, conceptuellement, revient à considérer que les observations sont gaussiennes). Pour cette procédure, nous avons appliqué, après l'étape correspondant à l'estimation de la direction de projection $\beta$, la même étape $B$ que l'algorithme `GSIM` correspondant à l'estimation de $\widehat{\eta}$. Seule la méthode d'estimation de la direction de projection diffère par rapport à `GSIM`, permettant ainsi de mesurer l'intérêt de la prise en compte de la relation entre espérance et variance dans les GLM. Xia et al. (2002) définissent une version raffinée (procédure `rOPG`) qui revient à ré-itérer l'étape A en modifiant la définition des poids. Comme nous l'expliquerons en section 5.3.5, les procédures `OPG` et `rOPG` coïncident sur le jeu de données `Colon`. Elles coïncident aussi avec la procédure `MAVE-LD` proposée par Antoniadis et al. (2003).

► *Codes disponibles* : Les algorithmes `OPG` et `GSIM` sont disponibles en R[3]. Les algorithmes `Ridge`, `NR`, `MNR`, `RPLS`, `IRPLSF` et `MIRPLSF` sont disponibles en MATLAB[4]. On trouvera aussi, en

[3] `http ://www-lmc.imag.fr/lmc-sms/Julie.Peyre`
[4] `http ://www.tsi.enst.fr/~gfort/GLM`

MATLAB, deux démonstrations de l'utilisation de ces algorithmes pour la classification binaire et multi-groupes sur les jeux `Colon` et `Leukemia`.

### 5.3.2 Estimation du taux d'erreur

Les performances des règles de décision sont évaluées par le taux d'erreur produit lorsqu'elles sont appliquées à de nouveaux cas. Le taux d'erreur est estimé par la valeur moyenne du taux d'erreur sur le fichier test ; la moyenne étant calculée sur 100 subdivisions aléatoires des données en un fichier d'apprentissage et un fichier test. Le fichier d'apprentissage est constitué de 41 échantillons pour le jeu `Colon` et chaque sous-population y est représentée dans la même proportion que dans la population totale. Le fichier d'apprentissage est constitué de 38 échantillons pour le jeu `Leukemia`, se répartissant en 19, 8 et 11 échantillons des classes ALL-Bcell, ALL-Tcell et AML (pour des raisons liées à l'origine du jeu de données).
Sur ces 100 subdivisions, la valeur moyenne du nombre de gènes à l'issue du pré-traitement est de 1157 pour `Colon` et 3070 pour `Leukemia`.

### 5.3.3 Paramètres d'implantation

Pour `KNN`, le nombre de voisins est déterminé par validation croisée sur l'ensemble d'apprentissage dans la grille des entiers $\{1, 2, 3, \cdots, 20\}$.
Pour `RPLS, NR, MNR, IRPLSF` et `MIRPLSF`, le nombre $\kappa$ de composantes PLS est déterminé par validation croisée sur l'ensemble d'apprentissage, dans la grille $\{1, \cdots, 6\}$.
Pour `RPLS` et `Ridge`, le paramètre de régularisation $\lambda$ est déterminé par optimisation du critère type BIC (13) sur 61 points de l'intervalle $[10^{-3}, 10^{3}]$, $\log_{10}$-linéairement espacés.
Pour `GSIM`, le paramètre de lissage $h_A$ et le paramètre de régularisation $\lambda$ sont choisis simultanément par validation croisée sur 5 points de l'intervalle $[7, 90]$ pour $h_A$ et sur 5 points de $[0.01, 30]$ pour $\lambda$ ; dans les deux cas, les points sont log-linéairement espacés.

### 5.3.4 Études menées

► Nous étudions tout d'abord le comportement de méthodes de discrimination classiques (`DLDA`, `DQDA`, `KNN`) lorsque le nombre de covariables $p$ est égal au nombre de covariables à l'issue du pré-traitement et est donc très grand devant le nombre d'observations.
► Nous étudions ensuite, dans les mêmes conditions, les classifieurs décrits en section 5.3.1 afin de souligner l'intérêt de la réduction de dimension dans une approche par régression généralisée. En particulier, en comparant les classifieurs `Ridge` et `RPLS`, nous étudions la pertinence de la combinaison "régularisation/réduction de dimension" par rapport à la seule stratégie "régularisation". En comparant les classifieurs `NR, MNR, IRPLSF, MIRPLSF` et `RPLS`, nous situons notre extension de PLS aux modèles linéaires généraux parmi les précédentes extensions. Enfin, en comparant les classifieurs `OPG` et `GSIM`, nous étudions l'intérêt de la prise en compte de la relation entre espérance et variance dans les GLM.

▶ Nous étudions la robustesse de ces classifieurs à l'augmentation du "bruit" induit par le grand nombre de régresseurs. Cette dernière étude est menée sur le jeu de données `Leukemia` pour lequel le déséquilibre des dimensions est particulièrement important ($n = 38$, $p \sim 3100$) ; ce jeu de données étant un cas de classification "multi-classe", cette étude comparative n'est menée que sur les classifieurs "classiques" et ceux issus de l'approche par régression paramétrique. Pour ce faire, nous décidons de réduire le nombre de régresseurs et de ne garder successivement que les 500, les 1000 puis les 1500 gènes les plus discriminants au sens de la statistique ANOVA (Dudoit et al. (2002)). Augmenter ainsi le nombre de gènes introduit une information de plus en plus mineure, au sens où le pouvoir discriminant des nouvelles covariables est de moins en moins significatif, et contribue à bruiter l'information disponible en augmentant la redondance (multi-colinéarité des régresseurs).
▶ Enfin, nous étudions la robustesse des classifieurs à la standardisation en ligne. Jusqu'ici, les données ont été pré-traitées selon les trois étapes indiquées en section 5.2.1 ; nous complétons ce pré-traitement par une étape de standardisation des biopuces et étudions l'impact de ce nouveau pré-traitement sur le taux d'erreur de classification. Cette étude est menée sur le jeu de données `Colon`.

### 5.3.5 Résultats

Les 100 taux moyens d'erreur de classification sur les échantillons test sont représentés par un graphique type boxplot (Figures 1 et 2) ; la valeur médiane est représentée par un trait épais et se confond, dans certains cas, avec l'un des quartiles. Nous reportons aussi dans les tableaux 1 et 3, la moyenne et l'écart-type empiriques de ce taux d'erreur obtenus sur les 100 échantillons test, ainsi que la valeur moyenne de certains hyper-paramètres. Les tableaux 2 et 4 et la figure 3 sont relatifs aux études de la robustesse au bruit et de la robustesse à la standardisation des biopuces.

▶ *Méthodes classiques :* La méthode `DQDA` est la moins performante. Sur les jeux `Colon` et `Leukemia`, certaines covariables ont une variance nulle au sein d'une classe (par convention, la distribution de cette covariable est un Dirac en la valeur moyenne). Ce phénomène, particulièrement fréquent sur le jeu `Leukemia` du fait d'un très petit nombre d'observations par classe dans l'ensemble d'apprentissage, explique la faiblesse de `DQDA` par rapport à `DLDA`. Pour le jeu `Colon`, `DLDA` est nettement moins efficace que les méthodes basées sur l'approche par régression généralisée ; en revanche, sur le jeu `Leukemia`, le taux d'erreur de classification est assez faible bien que plus élevé que la meilleure des approches paramétriques par régression généralisée (7% pour DLDA contre 4% pour Ridge-PLS).
En appliquant `KNN` sur `Colon` pour un nombre pair de voisins $k$, nous avons observé un grand nombre de cas d'indécisions : le $k$-voisinage contient autant de voisins d'une classe que de l'autre. Ainsi, sur les 100 subdivisions envisagées, le taux moyen d'indécision est respectivement de 31.2%, 21.7%, 16% lorsque $k = 2, 4, 6$. Le paramètre $k$ est déterminé par validation croisée sur l'ensemble d'apprentissage et le taux moyen d'indécision est du même ordre.
Ce phénomène d'indécision est aussi présent dans l'analyse du jeu `Leukemia` ; en théorie, il peut survenir pour toutes les valeurs de $k$ (à l'exception de $k = 1$), et en pratique, nous l'avons en

effet observé sur toutes les valeurs de $k > 1$, avec une fréquence particulièrement élevée pour des valeurs paires du nombre de voisins ($19.8\%, 10.8\%, 8.2\%$ et $2.6\%$, $6.6\%$ pour $k = 2, 4, 6$ et $k = 3, 5$).
Nous présentons donc deux résultats pour la méthode `KNN` respectivement référencés `KNN(1)` et `KNN(2)`. Nous déterminons tout d'abord $k$ comme indiqué en Section 5.3.3 et dans les cas d'indécision, la classe du nouvel échantillon est tirée au hasard (Devroye et al. (1996)). Dans la seconde approche, pour chacune des 100 étapes de l'analyse, nous choisissons $k$ parmi l'ensemble des valeurs pour lesquelles il n'y a jamais d'indécisions ni lors de la validation croisée sur l'ensemble d'apprentissage, ni lors de la classification des échantillons de l'ensemble test. Sur le jeu `Colon`, cela revient à chercher $k$ parmi l'ensemble des entiers impairs ; sur le jeu `Leukemia`, cela revient à prendre $k = 1$ (94% des cas) ou $k = 3$ (4% des cas).
Les résultats de classification sont sensibles à la parité de $k$ ; par exemple, le taux moyen d'erreur de classification sur les 100 échantillons test lorsque le nombre de voisins est fixe et vaut 1 (resp. 2) est de 10.8% (resp. 14.8%) pour le jeu `Leukemia`. Outre ce phénomène d'indécision dont la fréquence relativement élevée traduit la non adéquation de la méthode `KNN` pour traiter les données de biopuces, nous constatons que les résultats obtenus sont supérieurs à ceux obtenus par les méthodes basées sur l'approche par régression généralisée (25% contre 15% par `IRPLSF` pour `Colon` ; 10% contre 4% par `RPLS` pour `Leukemia`).

► *Approches paramétriques :* La comparaison de `Ridge` et `RPLS` illustre l'intérêt de la combinaison des étapes régularisation/réduction de dimension par rapport à une approche inférentielle incluant une étape de régularisation seule, du moins lorsque la régularisation est de type Ridge. La pénalité Ridge n'a pas d'effet sélectif : tous les régresseurs formant le modèle sont retenus dans la définition de l'estimateur et aucune réduction de dimension n'est appliquée. Il est manifeste, au vu des résultats, que l'approche `Ridge` n'est pas adaptée à l'analyse des données de biopuces et plus généralement, des données de grande dimension. Cette conclusion peut néanmoins dépendre de la méthode de sélection du paramètre de régularisation $\lambda$. Vues les applications que nous considérons ici, il pourrait sembler plus pertinent de choisir $\lambda$ par validation croisée avec une mesure de qualité basée sur le taux d'erreur de classification. On pourrait en particulier respecter la définition que Zhu et Hastie (2004) donnent de $\hat{\theta}^{\mathrm{R}}$ en déterminant $\lambda$ par validation croisée avec une mesure de qualité basée sur la déviance. Dans le cas de la discrimination dichotomique, il est facile de se convaincre que cette méthode de calcul de $\lambda$ coïncide avec la méthode par validation croisée avec une mesure de qualité basée sur le taux d'erreur de classification : dans les deux cas, le critère de validation croisée est minimal et vaut zéro pour $\lambda = 0$. En effet, quelle que soit la répartition de $n$ points dans $\mathbb{R}^n$ en deux classes, il existe toujours une direction $\hat{\theta}$ dont l'hyperplan associé sépare parfaitement les deux populations. Un tel vecteur (défini à une constante multiplicative près) rend alors la vraisemblance $\ell$ arbitrairement grande, et la déviance arbitrairement proche de zéro, de sorte que le critère de sélection du paramètre $\lambda$ proposé par Zhu et Hastie (2004) retourne $\lambda = 0$. Cette direction séparatrice n'est pas unique, et parmi les directions admissibles, il existe toujours une direction telle que l'échantillon de l'ensemble test est bien classé, de sorte que le critère de validation croisée basé sur le taux d'erreur de classification est minimal et vaut zéro pour $\lambda = 0$. La direction séparatrice n'étant pas unique, il n'y

a pas unicité de l'estimateur $\hat{\theta}^{\mathrm{R}}$ pour un ensemble d'apprentissage donné, et pour $\lambda = 0$ ; par suite, le taux d'erreur de classification sur les 100 échantillons test, du classifieur `Ridge` tel que défini par Zhu et Hastie (2004) est (très) sensible à la direction sélectionnée. Par exemple, sur le jeu `Colon`, le taux moyen d'erreur de classification est de 16.81% [écart type 0.048] (resp. 20.33% et écart-type 0.076 ou 44.20% et écart-type de 0.102) si $\hat{\theta}^{\mathrm{R}}$ est le vecteur de norme minimale vérifiant $\mathbf{Z}^{(G)}\theta = 2\mathbf{Y} - \mathbb{I}_n$ (resp. $\mathbf{Z}^{(G)}\theta = \Upsilon^{(i)}, i = 1, 2$ où $\Upsilon^{(i)}$ est choisi aléatoirement de façon à vérifier $\Upsilon^{(i)}_k \geq 0$ ssi $\mathbf{Y}_k = 1$).
Le calcul des estimateurs `NR` et `MNR` repose sur l'estimation d'un modèle de régression logistique à $\kappa$ régresseurs ; en pratique, pour `Colon` comme pour `Leukemia`, le phénomène de (quasi)-séparation est très fréquent lorsque $\kappa = 2$, et est presque systématique lorsque $\kappa \geq 3$. Dans ce cas, les estimateurs `NR` et `MNR` ne sont pas rigoureusement définis (cf. section 2.3.4). En conséquence, nous présentons deux résultats pour chacune des méthodes `NR` et `MNR`, respectivement référencés `NR(i)` et `MNR(i)`, $i = 1, 2$. Dans le premier cas, $\kappa$ est choisi comme indiqué en Section 5.3.3. Le phénomène de (quasi)-séparation survient dans 46% (resp. 98%) des résultats présentés en Figure 1 (resp. Figure 2) ; en pratique, l'algorithme itératif IRLS (section 2.3.2) est arrêté dès que la séparation est détectée, la valeur de l'estimateur est alors fixée à la valeur courante de la suite $(\theta^t)_t$ et elle devient donc dépendante de la valeur initiale de la procédure itérative. Dans le second cas, nous choisissons $\kappa$ par validation croisée parmi l'ensemble des valeurs pour lesquelles le phénomène de (quasi)-séparation ne survient jamais, ni lors de la validation croisée sur l'ensemble d'apprentissage, ni lors de la classification des échantillons de l'ensemble test. Les taux d'erreur obtenus pour chacun de ces deux résultats montrent la très grande sensibilité de la méthode au choix du nombre de composantes PLS (en moyenne, 16% contre 18% pour `Colon` et 5% contre 17% pour `Leukemia`). Plus fondamentalement, c'est le problème de l'existence même de l'estimateur qui est posé : Nguyen et Rocke (2002a,b) construisent un estimateur qui, très souvent sur les données de grande dimension telles que les biopuces, est de norme infinie. Aucun de leurs travaux ne mentionne ce fait ; nous avons décidé de pallier cette non-existence en choisissant pour estimateur le premier vecteur $\theta^t$ qui sépare parfaitement les échantillons de l'ensemble d'apprentissage. Cette frontière n'est pas unique et une utilisation honnête des méthodes `NR` et `MNR` nécessite de préciser le choix de l'estimateur parmi l'ensemble des vecteurs admissibles.
Le calcul de l'estimateur $\hat{\theta}^{\mathrm{DG}}$ repose sur un algorithme itératif "tronqué" dont il est difficile d'identifier les points limites (à notre connaissance, cela reste un problème ouvert). Nous avons observé que cet ensemble limite n'était pas un singleton : du choix du point limite dans l'ensemble peut dépendre le comportement du classifieur associé. En conséquence, nous montrons là encore deux résultats, référencés respectivement par `IRPLSF(i)` et `MIRPLSF(i)`, $i = 1, 2$. Dans le premier cas, $\kappa$ est choisi comme indiqué en Section 5.3.3. 16% (resp. 14%) des résultats présentés en Figure 1 (resp. Figure 2) ont été obtenus en arrêtant l'algorithme itératif IRPLSF (resp. MIRPLSF) après un nombre fixe d'itérations. Dans le second cas, nous choisissons $\kappa$ par validation croisée parmi l'ensemble des valeurs pour lesquelles les algorithmes IRPLSF / MIRPLSF convergent, lors de la validation croisée sur l'ensemble d'apprentissage et lors de la classification des échantillons de l'ensemble test. Les deux approches envisagées ont des performances similaires (12% contre 11.8% pour `Colon` ; 5.5% contre 5.2% pour `Leukemia`). Contrairement au cas de l'estimateur $\hat{\theta}^{\mathrm{NR}}$, le "problème" lié à l'estimateur $\hat{\theta}^{\mathrm{DG}}$ relève de sa non-unicité et non pas de

son existence. Sans caractérisation de l'ensemble des points limites (par exemple, caractérisation du type points critiques d'un critère à optimiser), il est difficile de justifier d'un choix particulier d'un point dans l'ensemble des points admissibles ; néanmoins les résultats présentés ici montrent que ce choix est sans (grande) conséquence sur les performances du classifieur, ce qui confirme des analyses antérieures relatives à l'utilisation de cet estimateur dans un processus de sélection de variables (Fort (2005)).

► *Approches semi-paramétriques :* La procédure `OPG` se simplifie dans le cas présent. Observons tout d'abord que l'on peut supposer, sans pertes de généralité, que la matrice $X$ est centrée en colonnes, à condition d'appliquer le même centrage à tout nouvel individu $x$ - et c'est ce que nous faisons -. L'analogue de l'étape A de `GSIM` pour l'algorithme `OPG` a alors une solution explicite donnée par $\hat{\eta}(X_j) = \mathbf{Y}_j$ et $\widehat{\nabla\eta}(X_j) = X^{+}\{\mathbf{Y} - \mathbb{1}_n \ \mathbb{1}_n' \mathbf{Y}/n\}$, pour tout $j \in \{1, \cdots, n\}$, et ce, quelle que soit la fenêtre de lissage $h_A$. Par suite, $\hat{\beta}$ colinéaire à $X^{+}\{\mathbf{Y} - \mathbb{1}_n \ \mathbb{1}_n' \mathbf{Y}/n\}$ où $X^{+}$ désigne la pseudo-inverse de la matrice $X$. Cette direction est aussi solution du critère MAVE (Xia et al., 2002, equation (2.7)). En conséquence, la direction $\hat{\beta}$ exhibée par la procédure `OPG`, par sa version raffinée `rOPG` et par la méthode `MAVE` coïncident. Dans `GSIM`, la fenêtre de lissage $h_B$ est déterminée par la méthode de plug-in, et c'est ce que nous faisons pour `OPG` afin de pouvoir véritablement comparer `GSIM` et `OPG/rOPG/MAVE`. Dans le cas présent, la valeur retournée par la méthode plug-in est $h_B = +\infty$ ; par convention, nous posons $K^1_{+\infty}(\cdot) = 1$. L'algorithme `OPG` (ou `rOPG` et `MAVE-LD`) est donc équivalent à la procédure suivante : **(a)** déterminer $\alpha_0, \alpha_1$, les coefficients de la régression logistique de $\mathbf{Y}$ sur la matrice de design $[\mathbb{1}_n \ X\hat{\beta}]$ ; **(b)** pour un nouvel individu $x$, estimer le prédicteur linéaire par $\widehat{\eta}(x) = \alpha_0 + \alpha_1 \ x\hat{\beta}$.
Cette simplification des méthodes `OPG`, `rOPG` et `MAVE-LD` lorsque la matrice de plan d'expérience est de rang plein en lignes souligne l'intérêt de l'algorithme `GSIM` : en ne tenant pas compte de la structure du bruit dans l'équation $Y = \mu(X) + \epsilon$ et en supposant implicitement que le modèle est gaussien (les critères sont de type moindres carrés), les méthodes `OPG`, `rOPG` et `MAVE-LD` se ramènent formellement au cas paramétrique. La méthode `GSIM` en revanche, prend en compte la structure des moments des modèles linéaires généralisés.

Inserer ici Figure 1 ; puis tableau 1 ; puis figure 2 et tableau 3.

► *Robustesse au bruit :* La figure 3 montre l'évolution de l'erreur moyenne de classification sur l'ensemble test lorsque le nombre de gènes formant le modèle augmente. Le tableau 4 précise l'évolution des hyper-paramètres des algorithmes. `KNN` souffre du fléau de la dimension : l'introduction du bruit augmente la fréquence du phénomène d'indécision, et les taux d'erreur de classification les plus bas sont atteints pour des valeurs de $k$ de plus en plus faibles. Les résultats montrent que l'instabilité de `MNR` commentée ci-dessus existe quelle que soit la valeur de $p$, le problème de l'existence de l'estimateur étant lié au déséquilibre des dimensions ($p > n$). Observons que pour `MNR(2)`, $\kappa$ est pratiquement toujours égal à 1 (100% des cas pour $p = 500$, 99% des cas pour $p = \text{pmax}$) : lorsque $\kappa \geq 2$, le phénomène de séparation est systématique. Cela signifie que le jeu `Leukemia` est un problème de classification relativement simple - du moins plus simple que celui de Colon - et explique l'efficacité de la méthode DLDA pour l'analyse du jeu Leukemia. Cette étude illustre enfin la stabilité des méthodes RPLS et MIRPLSF ; observons que pour pallier la difficulté croissante du problème de classification (liée à un nombre de régresseurs de plus en plus grand), les algorithmes doivent s'adapter : le nombre $\kappa$ de super-covariables nécessaires à la conservation des qualités prédictives du classifieur augmente, mais reste plus faible avec RPLS qu'avec MIRPLSF (en moyenne, 2.71 contre 4.34).

Inserer ici la figure 3 et le tableau 4.

► *Robustesse à la standardisation des biopuces :* Les résultats de cette étude sont présentés dans le tableau 2. Les méthodes DLDA et DQDA sont très sensibles à la standardisation des biopuces : le taux d'erreur de classification est pratiquement divisé par deux, et les performances de ces classifieurs deviennent comparables à celles des approches par régression. En revanche, toutes les méthodes par régression paramétrique comme non-paramétrique sont très stables, les fluctuations étant relativement faibles (comparées à celles de DLDA et DQDA). On constate que les résultats obtenus avec GSIM sont meilleurs que ceux obtenus avec OPG, utilisé ici en indice simple. Cela montre bien l'intérêt de la prise en compte de la relation entre espérance et variance dans les GLM.

Inserer ici le tableau 2.

► *Conclusion :* Nous avons étudié différentes méthodes de discrimination pour la classification de données de puces à ADN ; ces données relèvent du cadre statistique de la grande dimension : le nombre de variables décrivant chaque échantillon est très grand devant le nombre d'échantillons. Nous avons montré comment les méthodes Ridge-PLS et GSIM définies en Sections 3 et 4, pouvaient être utilisées avec succès pour répondre à ce problème de classification. L'originalité de ces approches réside dans la capacité des classifieurs associés à prendre en compte tous les gènes disponibles et donc à s'affranchir d'une étape de pré-sélection préliminaire ; cette robustesse des méthodes peut être exploitée à d'autres fins que celles présentées ici, comme par exemple, la sélection de variables et l'identification des variables discriminantes.
Nous avons montré que de toutes les méthodes classiques envisagées ici, DLDA semble la plus adaptée pour étudier les données de grande dimension comme les biopuces ; cette méthode apparaît néanmoins très sensible à la standardisation des biopuces. Bien que les taux d'erreur de classification de KNN soient relativement bas, nous voyons dans la fréquence du phénomène d'indécision, l'expression de l'instabilité de cette méthode lorsque l'information utile est très bruitée : les nuages de points regroupant les individus d'une même classe sont de plus en plus imbriqués et pour minimiser l'erreur de prédiction, il faut se contenter de la méthode du plus proche voisin. La mise en œuvre de ces classifieurs est certes simple, mais ces méthodes n'ont pas le potentiel des autres méthodes basées sur l'approche régression. Au prix certes, d'une plus grande complexité de mise en œuvre (par exemple, choix du paramètre de régularisation $\lambda$ et

du nombre de super-covariables $\kappa$ pour Ridge-PLS ; choix du paramètre de régularisation $\lambda$, de la fenêtre de lissage $h_A$ et de la fenêtre $h_B$ pour GSIM), les méthodes Ridge-PLS et GSIM *(i)* sont capables de traiter les données de grande dimension (nous avons vu par exemple que les méthodes `NR/MNR` n'étaient pas toujours applicables, l'estimateur du coefficient de régression n'étant pas toujours défini), *(ii)* apportent une réponse pertinente au problème statistique de la discrimination à but décisionnel (les taux d'erreur de classification sont relativement faibles), et *(iii)* permettent de quantifier le rôle joué par chaque variable, information qui pourra être exploitée pour répondre au problème statistique de la discrimination à but explicatif : fournir une meilleure compréhension du phénomène qui a engendré les observations.

### 5.3.6 Perspectives

► *Approches paramétriques :* Ridge-PLS est une méthode de régularisation et de réduction de dimension dans les modèles linéaires généralisés, développée pour l'analyse de données de grande dimension. Nous avons déjà mentionné que l'introduction d'une pénalité plus sélective pourrait améliorer les propriétés de l'estimateur PLS-pénalisé en permettant de combiner en une même procédure, régularisation, sélection de variables et réduction de dimension. Par exemple, le calcul du maximum de vraisemblance pénalisé dans les modèles linéaires lorsque la pénalité est de type Lasso, est une procédure itérative qui peut être lue comme une méthode de sélection de variables "forward" (Tibshirani (1996); Efron et al. (2004)). L'extension de cette méthode aux modèles linéaires généralisés et aux modèles de grande dimension (on préférera dans ce cas une pénalité type Elastic Net (Zou et Hastie (2005))) devrait donc être par nature même, une méthode de régularisation (introduction de pénalité), une méthode de sélection de variables (à chaque itération, une nouvelle variable est introduite pour définir le nouveau modèle) et une méthode de réduction de dimension.
Un second axe de recherche consiste en l'application de ces méthodes PLS-pénalisé pour la sélection de variables afin de définir des modèles parcimonieux sans détériorer les qualités prédictives. Dans Fort (2005), les estimateurs `NR,` `RPLS` et `IRPLSF` sont insérés dans une procédure de sélection type RFE (Guyon et al. (2002)), et permettent d'exhiber de petits modèles (formés de 8 gènes par `RPLS` et de 12 par `IRPLSF`). On pourra donc là encore étudier l'impact de pénalités autres que la pénalité Ridge sur cette procédure de sélection de variables.
Enfin, on pourra exploiter les méthodes PLS-pénalisé pour mener une analyse discriminante à but descriptif. On pourra par exemple mimer l'approche de Hastie et al. (2000), qui exploitent les super-gènes, combinaisons linéaires des gènes initiaux obtenus par une ACP, pour identifier les gènes dont le niveau d'expression est cohérent au sein de chaque classe mais très variable entre les classes. Plus généralement, il s'agit d'analyser la structure des données en définissant des groupes de covariables (non nécessairement disjoints) homogènes. Une des motivations importante pour développer de telles méthodes est l'intérêt d'identifier des groupes de gènes qui séparent de manière optimale des individus de classe différente et qui pourraient être reliés à une propriété biologique.

► *Approches semi-paramétriques :* La méthode GSIM n'est pour le moment développée que dans le cas $G = 1$. L'extension au cas multiclasse n'est pas directe. En effet, s'offrent plusieurs possibilités. La plus naturelle serait de poser dans le modèle une direction par classe, c'est-à-dire $\eta_i(x) = \eta_i(\beta_i^T x)$ $i = 1, \dots, G$. Cependant cette approche conduit à l'estimation d'un nombre conséquent de paramètres. Une alternative consisterait à garder la même direction dans chaque classe. Indépendemment du nombre de groupes, on peut espérer de meilleurs résultats pour la méthode GSIM étendue à plusieurs directions. On pourra s'inspirer de ce qu'ont fait Xia et al. (2002). Enfin, il serait intéressant d'appliquer GSIM dans la procédure de sélection de type RFE. Tout ceci fait l'objet de travaux en cours.

**Remerciements** Nous remercions A. Antoniadis et les rapporteurs anonymes pour leurs suggestions et commentaires constructifs qui ont permis d'améliorer la présentation de ce papier. G. Fort a mené cette recherche au sein du laboratoire de Modélisation et Calcul (UMR 5523) à Grenoble.

## Références

Albert, A. et Anderson, J. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1) :1–10.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., et Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12) :6745–6750.

Antoniadis, A., Lambert-Lacroix, S., et Leblanc, F. (2003). Effective Dimension Reduction Methods for Tumor - Classification using gene Expression Data. *Bioinformatics*, 19(5) :563–570.

Barker, M. et Rayens, W. (2003). Partial least squares for discrimination. *J. Chemometrics*, 17 :166–173.

Bastien, P. (2004). PLS-Cox model : application to gene expression. Dans *Proceedings in Computational Statistics*, pages 655–662. Physica-Verlag, Springer.

Bastien, P., Esposito Vinzi, V., et Tenenhaus, M. (2004). PLS generalised linear regression. *Comput. Stat. Data Anal.*, 48(1) :17–46.

Bull, S., Mak, C., et Greenwood, C. (2001). A modified score function estimator for multinomial logistic regression is small samples. *Comput. Stat. Data Anal.*, 39 :57–74.

De Jong, S. (1995). PLS shrinks. *J. Chemometrics*, 9 :323–326.

Denham, M. (2000). Choosing the number of factors in partial least squares regression : estimating and minimizing the mean squared error of prediction. *J. Chemometrics*, 14 :351–361.

Devroye, L., Gyorfi, L., et Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New-York.

Ding, B. et Gentleman, R. (2005). Classification Using Generalized Partial Least Squares. *J. Comp. Graph. Stat.* A paraître.

Dudoit, S., Fridlyand, J., et Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97(457) :77–87.

Efron, B., Hastie, T., Johnstone, I., et Tibshirani, R. (2004). Least angle regression. *Ann. Stat.*, 32(2) :407–499.

Fahrmeir, L. et Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models. 2nd ed.* Springer Series in Statistics. New York.

Fan, J. et Gijbels, I. (1996). *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Fort, G. (2005). Partial Least Squares for Classification and Feature selection in Microarray gene expression data. Soumis.

Fort, G. et Lambert-Lacroix, S. (2005). Classification using Partial Least Squares with Penalized Logistic Regression. *Bioinformatics*, 21(7) :1104–1111.

Frank, I. et Friedman, J. (1993). A statistical view of some chemometrics regression tools, with discussion. *Technometrics*, 35(2) :109–148.

Garthwaite, P. (1994). An interpretation of partial least squares. *J. Am. Stat. Assoc.*, 89(425) :122–127.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., et Lander, E. (1999). Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439) :531–537.

Goutis, C. (1996). Partial Least Squares algorithm yields shrinkage estimators. *Ann. Stat.*, 24(2) :816–824.

Green, P. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *J. R. Stat. Soc., Ser. B*, 46(2) :149–192.

Guyon, I., Weston, J., Barnhill, S., et Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.*, 46(1-3) :389–422. Erratum : http ://clopinet.com/isabelle/Papers/index.html.

Hastie, T. et Tibshirani, R. (1990). *Generalized Additive Models.* Monographs on Statistics and Applied Probability. New York : Chapman and Hall.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., et Brown, P. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, 1.

Helland, I. (1988). On the structure of Partial Least Squares regression. *Commun. Stat. Simulation Comput.*, 17(2) :581–607.

Helland, I. (1990). Partial Least Squares Regression and Statistical Models. *Scand. J. Statist.*, 17(2) :97–114.

Kuo, W., Kim, E., Trimarchi, B., Jenssen, J., Vinterbo, T., et Ohno-Machado, L. (2004). Bayes factor. *J. Biomed. Inform.*, 37 :293–303.

Lambert-Lacroix, S. et Peyre, J. (2005). Local quasi-likelihood regression in generalized single-index models. Travaux en cours.

Lesaffre, E. et Albert, A. (1989). Partial separation in logistic discrimination. *J. R. Stat. Soc., Ser. B*, 51(1) :109–116.

Lingjaerde, O. et Christophersen, N. (2000). Shrinkage structure of Partial Least Squares. *Scand. J. Stat.*, 27 :459–473.

Marx, B. D. (1996). Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression. *Technometrics*, 38(4) :374–381.

Nguyen, D. et Rocke, D. (2002a). Multi-class cancer classification via Partial Least Squares with gene expression profiles. *Bioinformatics*, 18(9) :1116–1226.

Nguyen, D. et Rocke, D. (2002b). Tumor classification by Partial Least Squares using microarray gene expression data. *Bioinformatics*, 18(1) :39–50.

Nguyen, D. et Rocke, D. (2004). On partial least squares dimension reduction for microarray-based classification : a simulation study. *Comput. Stat. Data Anal.*, 46 :407–425.

Phatak, A., Reilly, P., et Penlidis, A. (2002). The asymptotic variance of the univariate PLS estimator. *Linear Algebra Appl.*, 354(1-3) :245–253.

Santner, T. et Duffy, D. (1986). A note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 73(3) :755–758.

Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Paris : Éditions Technip.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2) :461–464.

Seifert, B. et Gasser, T. (1996). Finite sample variance of local polynomials : Analysis and solutions. *J. Am. Stat. Assoc.*, 91(433) :267–275.

Stoica, P. et Söderström, T. (1998). Partial Least Squares : A First-order Analysis. *Scand. J. Stat.*, 25(1) :17–24.

Stone, M. et Brooks, R. (1990). Continuum regression : Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Stat. Soc., Ser. B*, 52(2) :237–269.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc., Ser. B*, 58(1) :267–288.

Xia, Y., Tong, H., Li, W., et Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc., Ser. B*, 64(3) :363–410.

Zhu, J. et Hastie, T. (2004). Classification of Gene Microarrays by Penalized Logistic Regression. *Biostatistics*, 5 :427–443.

Zou, H. et Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc., Ser. B*, 67(2) :301–320.

## Tableaux et Figures

FIG. 1 – `Colon.` Boxplot du taux d'erreur de classification moyen sur l'ensemble test.

| | DLDA | DQDA | KNN(2) | Ridge | RPLS | NR(1) | NR(2) | IRPLSF(2) | OPG | GSIM |
|---|---|---|---|---|---|---|---|---|---|---|
| moy | 0.286 | 0.345 | 0.252 | 0.342 | 0.154 | 0.163 | 0.180 | 0.146 | 0.164 | 0.167 |
| std | 0.140 | 0.125 | 0.080 | 0.061 | 0.060 | 0.064 | 0.095 | 0.048 | 0.065 | 0. 069 |
| $\kappa$ | - | - | 5.06 | - | 2.84 | 3.27 | 1.87 | 3.13 | - | - |

TAB. 1 – `Colon.` Moyenne et écart-type du taux d'erreur dans le fichier test. La dernière ligne précise la valeur moyenne de $\kappa$ (ou $k$ pour KNN).

| | DLDA | DQDA | KNN(2) | Ridge | RPLS | NR(1) | NR(2) | IRPLSF(2) | OPG | GSIM |
|---|---|---|---|---|---|---|---|---|---|---|
| moy | 0.144 | 0.173 | 0.248 | 0.352 | 0.160 | 0.158 | 0.173 | 0.155 | 0.158 | 0.145 |
| std | 0.057 | 0.072 | 0.067 | 0.054 | 0.060 | 0.056 | 0.073 | 0.063 | 0.049 | 0.056 |
| $\kappa$ | - | - | 4.44 | - | 2.11 | 2.73 | 1.07 | 1.95 | - | - |

TAB. 2 – `Colon.` Moyenne et écart-type du taux d'erreur dans le fichier test lorsque le prétraitement est suivi d'une étape de standardisation des biopuces. La dernière ligne précise la valeur moyenne de $\kappa$ (ou $k$ pour KNN).

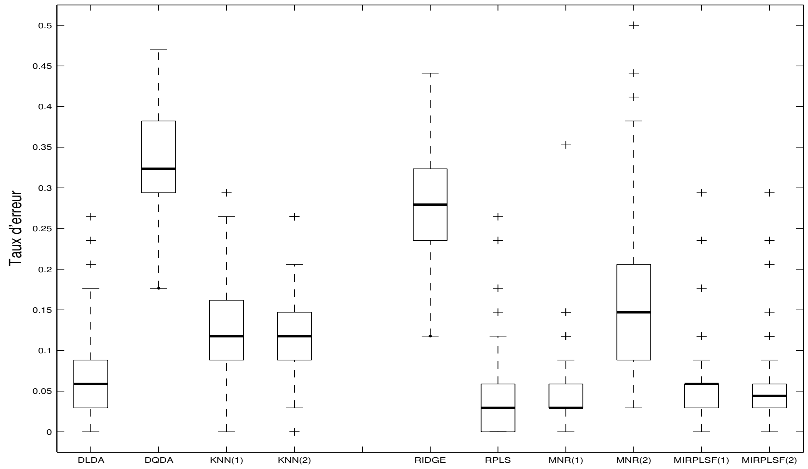


FIG. 2 – `Leukemia.` Boxplot du taux d'erreur de classification moyen dans le fichier test.

| | DLDA | DQDA | KNN(2) | Ridge | RPLS | MNR(1) | MNR(2) | MIRPLSF(2) |
|---|---|---|---|---|---|---|---|---|
| moy | 0.071 | 0.335 | 0.108 | 0.282 | 0.041 | 0.048 | 0.168 | 0.052 |
| std | 0.053 | 0.067 | 0.048 | 0.072 | 0.046 | 0.046 | 0.095 | 0.046 |
| $\kappa$ | - | - | 1.22 | - | 2.71 | 3.93 | 1.01 | 4.34 |

TAB. 3 – `Leukemia.` Moyenne et écart-type du taux d'erreur dans le fichier test. La dernière ligne précise la valeur moyenne de $\kappa$ (ou $k$ pour KNN).

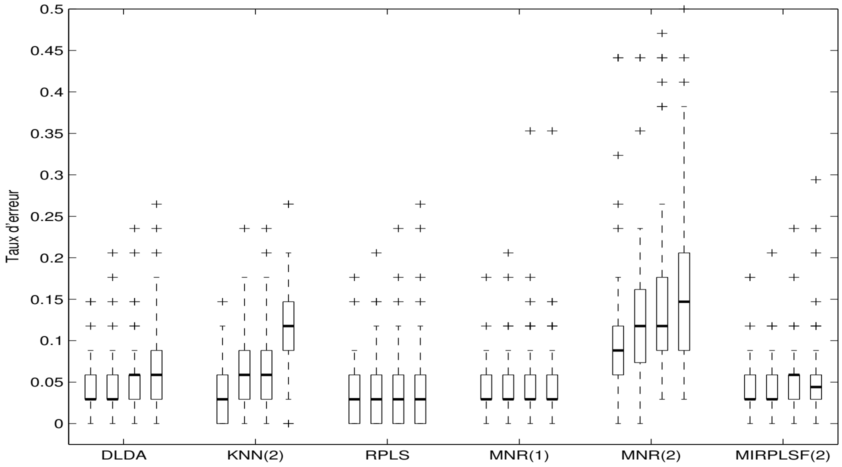


FIG. 3 – `Leukemia`. Boxplot du taux d'erreur de classification moyen dans le fichier test pour différentes valeurs de $p$ : 500, 1000, 1500, $p_{\max}$ (de gauche à droite).

| | KNN(2) | RPLS | MNR(1) | MNR(2) | MIRPLSF(2) |
|---|---|---|---|---|---|
| 500 | 2.34 | 1.62 | 3.39 | 1.00 | 3.02 |
| 1000 | 2.47 | 2.06 | 3.47 | 1.00 | 3.20 |
| 1500 | 1.61 | 2.38 | 3.74 | 1.11 | 3.41 |
| $p_{\max}$ | 1.22 | 2.71 | 3.93 | 1.01 | 4.34 |

TAB. 4 – `Leukemia`. Valeur moyenne de $\kappa$ (ou $k$ pour KNN) pour différentes valeurs de $p$.