MATHÉMATIQUES ET SCIENCES HUMAINES

VLADIMIR MAKARENKOV
PIERRE LEGENDRE

Une méthode d'analyse canonique non-linéaire et son application à des données biologiques

Mathématiques et sciences humaines, tome 147 (1999), p. 135-147
<http://www.numdam.org/item?id=MSH_1999__147__135_0>

Math. Inf. Sci. Hum., (37e année, n°147, *1999*, pp. 135-147)

UNE MÉTHODE D'ANALYSE CANONIQUE NON-LINÉAIRE ET SON APPLICATION À DES DONNÉES BIOLOGIQUES

Vladimir MAKARENKOV[1] et Pierre LEGENDRE[2]

RÉSUMÉ – *Parmi les méthodes d'ordination proposées dans la littérature statistique, l'ACR (analyse canonique de redondance) est devenue l'une des plus employées par les écologistes. En ACR, deux tableaux des données sont considérés. Le premier tableau (**Y**) contient les variables-réponse (e.g. les abondances des espèces étudiées) alors que le second (**X**) contient les variables explicatives (e.g. les variables environnementales). L'ACR classique impose des contraintes linéaires entre les variables **X** et **Y**, ce qui reflète rarement les processus naturels. Nous proposons une nouvelle méthode d'ordination, l'ACR polynomiale, qui permet de modéliser des relations linéaires ou non. Cette méthode est basée sur un algorithme empirique de régression qui permet de chercher la forme des relations polynomiales entre les variables en **X** et **Y** ainsi que de prendre en compte les interrelations entre variables explicatives.*

MOTS-CLÉS – Analyse de redondance, régression linéaire multiple, régression polynomiale.

SUMMARY – A method of nonlinear canonical analysis and its application to biological data.
*Among the various forms of canonical analysis available in the statistical literature, RDA (redundancy analysis) has become an instrument of choice for ecological analysis. A first data table (**Y**) contains the response variables (e.g. species data) whereas the second table (**X**) contains the explanatory variables (e.g. environmental variables). Classical RDA assumes that the relationships between variables in **X** and **Y** are linear ; this is unrealistic in most cases. We propose a new ordination method, called polynomial RDA, to do away with the constraints of linearity in these relationships. Polynomial RDA is based on an empirical regression algorithm which allows polynomial relationships to be modelled between the variables in **X** and **Y** ; it also takes into account the relationships among the explanatory variables.*

KEYWORDS – Redundancy analysis, multiple linear regression, polynomial regression.

[1] Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada et Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia, e-mail : makarenv@ere.umontreal.ca

[2] Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada, e-mail : legendre@ere.umontreal.ca

1. DESCRIPTION DU PROBLÈME

Le développement de l'analyse canonique de redondance (ACR) est l'œuvre de C. R. Rao [7]. L'analyse canonique de redondance est aussi désignée dans la littérature française sous le nom *d'analyse en composantes principales avec variables instrumentale*s (ACPVI*)*. L'ACR est une extension directe de l'analyse en composantes principales (ACP) de **Y** effectuée sous les contraintes imposées par **X**. L'ACR trouve les axes d'ordination des nuages de points qui sont le plus fortement linéairement liés à l'ensemble des variables explicatives. À cause de ces contraintes imposées par les variables explicatives, l'ACR est souvent appelée *ordination sous contrainte*.

Deux façons d'effectuer une ACR sont connues dans la littérature ; elles conduisent à des résultats identiques. La première stratégie, proposée par ter Braak [9], utilise l'algorithme itératif de calcul des axes principaux de Hill [5]. Cette méthode est incorporée entre autres dans le programme populaire CANOCO (ter Braak [10]) qui permet de mettre en œuvre plusieurs méthodes d'ordination. Dans notre étude, nous avons plutôt employé une stratégie de calcul de l'ACR qui consiste en une série de régressions linéaires multiples des variables-réponse **Y** sur les variables explicatives **X**, suivies d'une analyse en composantes principales de la matrice des valeurs ajustées. Cette stratégie est décrite, par exemple, par Legendre et Legendre [6].

Les axes d'ordination indiquent les directions de plus grande variabilité de la matrice des variables-réponse ajustées $\hat{\mathbf{Y}}$. Cette matrice s'obtient par régression linéaire multiple de la matrice des variables-réponse **Y** sur la matrice des variables explicatives **X**. Par conséquent, les vecteurs d'ordination sont contraints d'être des combinaison linéaires des variables de la matrice **X**. Il n'y a pas, en fait, de raison particulière pour laquelle les changements en quantité des espèces seraient reliés *linéairement*, dans la nature, à ceux des variables environnementales. Supposer la linéarité n'est pas réaliste dans plusieurs modèles écologiques. On ne le fait que parce que des méthodes de calcul aux suppositions plus réalistes ne sont pas disponibles.

Parmi les premiers travaux s'intéressant à la modélisation des relations non-linéaires en ordination sous contrainte, nous devons mentionner Van der Burg et de Leeuw [13] qui ont proposé une méthode de recherche de transformations optimales («optimal scaling» en anglais) impliquant des transformations non-linéaires dans le cadre de l'analyse des corrélations canoniques. Une autre façon de prendre en compte les relations non-linéaire a été proposée par Breiman et Friedman [2] qui ont utilisé les transformations optimales de variables pour maximiser la corrélation multiple moyenne. Leur algorithme ACE génère ces transformations de manière itérative sous les contraintes définies par l'utilisateur (Breiman et Friedman [2] ; de Veaux [3]). Un effet similaire à l'algorithme ACE s'obtient par l'application de fonctions splines pour trouver les transformations optimales des variables (Ramsay [7] ; Winsberg et Ramsay [14]). Ces transformations ont été utilisées dans le contexte de la régression multiple impliquant soit toutes les variables de **X** et **Y**, soit certains sous-ensembles de variables. L'avantage d'une telle approche est que ces transformations optimales peuvent être générées à partir des données observées sans supposition quelconque à propos de la forme des relations entre les variables étudiées. Pour sa part, Durand [4] a proposé d'utiliser les splines additifs pour transformer les variables dans le cadre de l'ACR.

Nous proposons une méthode empirique d'ordination, l'ACR polynomiale basée sur un algorithme de régression polynomiale, qui permet de passer outre aux contraintes de linéarité dans la description des relations entre les variables de **X** et **Y**. Cette méthode

est une généralisation de l'ACR classique. Notre ACR polynomiale utilise une nouvelle forme de régression polynomiale pour chacune des variables de **Y** en lieu et place de la régression linéaire multiple. Par rapport au modèle linéaire classique, cette approche permet souvent d'augmenter considérablement le pourcentage de variance de **Y** expliqué par le modèle.

La signification d'un modèle d'ordination canonique peut être estimée à l'aide d'une procédure de test par permutation (ter Braak [11] ; Legendre et Legendre [6]). Il est également possible de vérifier lequel des deux modèles, linéaire ou polynomial, correspond le mieux aux données observées, en testant la signification de la *quantité additionnelle de variance* expliquée par le modèle polynomial par rapport au modèle linéaire.

Notre algorithme empirique de régression polynomiale permet d'établir les relations polynomiales entre les variables-réponse et les variables explicatives. Les variables-réponse de la matrice des valeurs ajustées $\hat{\mathbf{Y}}$, obtenues à l'aide de cet algorithme et utilisées dans la suite de l'analyse, ne sont plus des combinaisons linéaires des variables de **X**, mais des combinaisons polynomiales. Dans cette étude, nous n'avons considéré que les polynômes où le degré de chaque variable est au maximum 2, certains termes étant sélectionnés durant la procédure de régression. Notre algorithme de régression n'est pas conçu pour chercher un polynôme optimal avec un nombre fixe de termes. Nous tentons simplement d'expliquer une part de la variance de **Y** qui n'a pas été expliquée par la régression linéaire multiple.

Puisque les axes d'ordination obtenus après la décomposition en valeurs propres de la matrice de dispersion de $\hat{\mathbf{Y}}$ sont contraints d'être des combinaisons polynomiales des variables de **X**, nous suggérons une nouvelle façon de représenter les variables explicatives de **X** dans les diagrammes de double projection («biplot» en anglais) qui permettent d'interpréter les résultats de l'ACR en termes biologiques. Nous proposons d'utiliser la corrélation linéaire multiple au lieu de la corrélation linéaire simple pour calculer les coordonnées des variables explicatives dans ce diagramme.

Pour cette présentation, nous avons choisi d'analyser à l'aide de l'ACR linéaire et polynomiale des données recueillies à 20 stations d'échantillonnage de l'étang de Thau (côte méditerranéenne de la France) en octobre 1988. Les données comprennent deux variables bactériennes, deux variables chimiques et une variable spatiale.

Le logiciel réalisant l'ACR polynomiale ainsi que l'analyse canonique des correspondances (ACC) polynomiale est disponible sur le World Wide Web à l'adresse <http:/www.fas.umontreal.ca/BIOL/legendre/>. Le programme permet également de réaliser l'ACR classique, de même que l'ACC classique, basées sur la régression linéaire, et de comparer la signification de la différence entre les deux modèles à l'aide d'une procédure de test par permutation.

2. L'ALGORITHME DE RÉGRESSION POLYNOMIALE

Comme dans le cas de la régression linéaire, l'objectif principal de la procédure de régression que nous proposons est de trouver un modèle fonctionnel (ici polynomial) reliant les variables-réponse aux variables explicatives, ou encore d'essayer de prédire les valeurs des variables-réponse.

L'algorithme présenté dans ce paragraphe permet de décrire chacune des variables-réponse **y** comme une fonction polynomiale des variables explicatives. Notre algorithme procède par combinaisons successives des colonnes de la matrice **X** ; la valeur du coefficient de détermination R^2 pour la variable-réponse **y**, augmente à chaque itération. La procédure de réduction est effectuée de façon indépendante pour chacune des variables-réponse **y** de la matrice **Y**. La réduction de la matrice des variables explicatives est nécessaire pour éviter de surestimer les variables-réponse.

Soit une variable-réponse $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ et une matrice de variables explicatives **X** d'ordre $(n \times m)$. Notre algorithme comprend quatre étapes de base qui sont répétées $m - 1$ fois, ce qui permet de réduire la matrice **X** de m colonnes à un simple vecteur.

1. La première étape de l'algorithme est une régression multiple classique de **y** sur toutes les variables de **X**. Une colonne de 1 est ajoutée à la matrice **X** pour prendre en compte l'intercepte, ce qui, dans la régression linéaire classique, produit le même effet que le centrage du vecteur **y** ainsi que des variables de **X**. Le vecteur des valeurs ajustées $\hat{\mathbf{y}}$ se calcule selon la formule suivante, où **b** est le vecteur des coefficients de régression partiels :

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} \qquad (1)$$

2. La seconde étape permet d'obtenir la matrice des résidus de la régression linéaire multiple :

$$\mathbf{y}_{\mathbf{res}} = \mathbf{y} - \hat{\mathbf{y}} \qquad (2)$$

3. La troisième étape a pour but de sélectionner la paire de variables de **X** assurant la meilleure approximation quadratique du vecteur $\mathbf{y}_{res}$. Pour effectuer cette sélection, nous calculons, pour chaque paire de colonnes i et j de **X**, la régression linéaire multiple du vecteur $\mathbf{y}_{res}$ sur la matrice $\mathbf{X}^{ij}$ composée des colonnes comprenant les variables $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_i\mathbf{x}_j$, $\mathbf{x}_i^2$, $\mathbf{x}_j^2$ ainsi qu'une colonne de 1. Pour $i = 1$ et $j = 2$, l'équation est la suivante:

$$\mathbf{X}^{12} = \begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{21} & \mathbf{x}_{11}\mathbf{x}_{21} & \mathbf{x}_{11}^2 & \mathbf{x}_{21}^2 & 1 \\ \mathbf{x}_{12} & \mathbf{x}_{22} & \mathbf{x}_{12}\mathbf{x}_{22} & \mathbf{x}_{12}^2 & \mathbf{x}_{22}^2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{x}_{1n} & \mathbf{x}_{2n} & \mathbf{x}_{1n}\mathbf{x}_{2n} & \mathbf{x}_{1n}^2 & \mathbf{x}_{2n}^2 & 1 \end{bmatrix} \quad \text{et} \quad \hat{\mathbf{y}}_{\mathbf{res}}^{12} = \mathbf{X}^{12}\mathbf{c}^{12} \qquad (3)$$

où $\hat{\mathbf{y}}_{\mathbf{res}}^{12}$ est le vecteur des valeurs ajustées aux résidus obtenus de l'équation 2. Si la variable $\mathbf{x}_1$ est binaire, la quatrième colonne de $\mathbf{X}^{12}$ ne doit pas être incluse dans cette matrice ; de même, on ne doit pas inclure la cinquième colonne si la variable $\mathbf{x}_2$ est binaire. Cette opération s'impose parce que le carré d'une variable binaire est égal à cette même variable. Le vecteur des coefficients de régression $\mathbf{c}^{12}$ s'obtient par les moindres carrés, de la même façon que le vecteur **b** de l'équation 1. Le coefficient de détermination multiple $R^2(1,2)$ correspondant à cette régression est ensuite calculé. Cette procédure est répétée pour toutes les paires i,j de colonnes de **X** ; chaque fois, le

coefficient de détermination multiple $R^2(i,j)$ est calculé. La paire i,j correspondant au plus grand coefficient $R^2(i,j)$ est retenue pour être employée à la quatrième étape.

Durant cette étape, les deux colonnes i,j sélectionnées à l'étape précédente sont fusionnées ; elles forment alors une nouvelle colonne t de $\mathbf{X}$ qui remplacera i et j pour la suite de l'analyse. Nous utilisons la formule ci-dessous pour calculer les valeurs de la nouvelle variable t :

$$x_{k,t} = x_{k,i}\, b_i + x_{k,j}\, b_j + \hat{y}^{ij}_{res,k} \quad \text{pour tous les } k = 1, \ldots, n \tag{4}$$

Les coefficients b sont ceux qui ont été obtenus à partir de l'équation 1. Nous avons ainsi réduit la taille de la matrice $\mathbf{X}$, qui comprend maintenant une colonne de moins que précédemment. Cette nouvelle colonne combine les termes correspondant à la contribution linéaire de $\mathbf{x}_i$ et $\mathbf{x}_j$ à la régression de $\mathbf{y}$ sur $\mathbf{X}$, ainsi que les termes de la régression du vecteur des résidus $\mathbf{y}_{res}$ sur la matrice $\mathbf{X}^{ij}$.

La boucle en quatre étapes, décrites ci-dessus, doit être répétée $m - 1$ fois pendant lesquelles la matrice $\mathbf{X}(n \times m)$ se transforme en une matrice $\mathbf{X}(n \times 1)$ qui est un simple vecteur. Pour obtenir le vecteur final $\hat{\mathbf{y}}$ de notre procédure de régression multiple, nous effectuons une régression linéaire simple de $\mathbf{y}$ sur $\mathbf{X}(n \times 1)$. Il est évident que le vecteur $\hat{\mathbf{y}}$ ainsi obtenu est une fonction polynomiale des variables explicatives de la matrice initiale $\mathbf{X}(n \times m)$ considérée au début de la procédure de régression. Notons qu'il ne serait pas possible de contrôler le degré de ce polynôme si la forme matricielle de l'équation 3 était employée indistinctement pour chacune des $m - 1$ itérations en quatre étapes, présentées ci-dessus. Pour limiter le polynôme à une forme où le degré de chacune des variables de la matrice $\mathbf{X}$ initiale est égal à 2, la règle suivante de composition de la matrice $\mathbf{X}^{ij}$ est appliquée pour toute paire de variable (i,j), à partir de la deuxième itération de l'algorithme :

- si la colonne i est déjà une variable combinée, obtenue à l'aide de l'équation 4, la colonne $\mathbf{x}_i^2$ ne doit pas être incluse dans la matrice $\mathbf{X}^{ij}$;
- la même règle s'applique pour la variable j s'il s'agit d'une variable déjà combinée.

Il en résulte que la matrice $\mathbf{X}^{ij}$ peut comprendre de quatre à six colonnes, selon la nature des variables i et j qui la composent. La régression sans limitation du degré du polynôme peut aussi être considérée comme un cas particulier de la régression polynomiale. Si les relations entre $\mathbf{X}$ et $\mathbf{Y}$ sont purement linéaires, les coefficients des termes quadratiques des variables de $\mathbf{X}$ sont égaux a zéro pour chacune des variables de $\mathbf{Y}$. Deux exemples numériques d'utilisation de notre procédure empirique de régression polynomiale sont présentés à la section 5 de cet article.

Estimons maintenant la complexité algorithmique de cette procédure. Pour obtenir la matrice des valeurs ajustées $\hat{\mathbf{Y}}$ d'ordre $(n \times p)$, nous avons $m - 1$ itérations à effectuer indépendamment pour chacune des p variables-réponse $\mathbf{y}$ de $\mathbf{Y}$. La complexité de la première étape de la procédure est dans le pire des cas $O(nm^2) + O(m^3)$, alors que pour la seconde, la troisième et la quatrième étapes c'est $O(n)$. Puisque nous supposons toujours que $n>m$, la complexité totale de chaque itération comprenant les quatre étapes ci-dessus sera $O(nm^2)$. Nous concluons donc que toute la procédure de régression, exécutée sur les matrices $\mathbf{Y}(n \times p)$ et $\mathbf{X}(n \times m)$, exige un nombre d'opérations d'ordre pnm^3.

3. L'ANALYSE CANONIQUE DE REDONDANCE ET SA GÉNÉRALISATION POLYNOMIALE

Il existe un certain nombre de logiciels permettant de réaliser l'analyse canonique de redondance (ACR). Mentionnons à titre d'exemple CANOCO (ter Braak [10], [11]) et RDACCA (P. Legendre, 1998, disponible sur le World Wide Web à l'adresse <http:/www.fas.umontreal.ca/BIOL/legendre/>). Toutes différentes que soient les stratégies algorithmiques utilisées dans ces deux programmes, elles conduisent à des résultats identiques. Dans cet article, nous avons suivi la stratégie d'ordination décrite dans Legendre et Legendre [6] et qui est à la base du programme RDACCA. Les principales étapes de cette technique sont rappelées ci-dessous.

Soit $\mathbf{Y}$ une matrice de variables-réponse à n lignes, représentant les sites ou autres objets, et p colonnes correspondant aux espèces étudiées ou autres variables. Par exemple, $\mathbf{Y}$ peut être la matrice des abondances des espèces observées à chacun des sites. Soit $\mathbf{X}$ une matrice de variables explicatives à n lignes représentant les mêmes sites que dans la matrice $\mathbf{Y}$ et m colonnes correspondant aux variables explicatives – des variables environnementales, par exemple – qui ont été observées sur ces sites.

1. La première étape de l'ACR classique consiste en la régression linéaire multiple de chaque variable de $\mathbf{Y}$, à tour de rôle, sur toutes les variables de $\mathbf{X}$. On centre les variables de $\mathbf{Y}$ avant de procéder aux régressions. Les variables de la matrice $\mathbf{X}$ doivent aussi être centrées ; si elles ne le sont pas, une colonne de 1 doit être ajoutée à la matrice $\mathbf{X}$ avant les régressions. Les valeurs ajustées $\hat{\mathbf{Y}}$ utilisées plus loin dans l'analyse sont donc obtenues à l'aide de la formule suivante :

$$\hat{\mathbf{Y}} = \mathbf{XB} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y} \qquad (5)$$

2. À la seconde étape, on calcule la matrice de covariance $\mathbf{S}$ de la matrice des valeurs ajustées $\hat{\mathbf{Y}}$:

$$\mathbf{S} = [1/(n-1)]\,\hat{\mathbf{Y}}'\,\hat{\mathbf{Y}} = [1/(n-1)]\,\mathbf{Y}'\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y} \qquad (6)$$

3. À la troisième étape, on fait la décomposition en valeurs propres et en vecteurs propres de la matrice de covariance $\mathbf{S}$ de $\hat{\mathbf{Y}}$. Cela correspond à la résolution de l'équation matricielle suivante :

$$(\mathbf{S} - \lambda_k\mathbf{I})\mathbf{u}_k = \mathbf{0} \qquad (7)$$

où λ_k désigne une valeur propre canonique et $\mathbf{u}_k$ le vecteur propre canonique associé à λ_k. La matrice contenant les vecteurs propres canoniques normalisés est notée $\mathbf{U}$. Les vecteurs propres fournissent les contributions des descripteurs de $\mathbf{Y}$ aux différents axes canoniques. Dans l'ACR linéaire, la matrice $\mathbf{U}$ est d'ordre $(p \times min[p,m,n-1])$ puisque le nombre de vecteurs propres canoniques ne peut excéder le minimum de p, m et $(n-1)$.

4. L'ordination des objets (i.e. les lignes de $\mathbf{Y}$) dans l'espace des variables-réponse $\mathbf{Y}$ s'obtient directement à partir de la matrice $\mathbf{Y}$ centrée, au moyen de l'équation standard pour le calcul des composantes principales :

$$\mathbf{Ord}_{(\text{espace des variables-réponse } \mathbf{Y})k} = \mathbf{Yu}_k \qquad (8)$$

Les vecteurs d'ordination définis par l'équation (8) sont appelés 'coordonnées des sites'. Ces vecteurs ont des variances proches mais pas nécessairement égales aux valeurs propres correspondantes.

De la même façon, l'ordination des objets en espace **X** s'obtient à l'aide de la formule suivante :

$$\mathbf{Ord}_{(\text{espace des variables explicatives } \mathbf{X})k} = \hat{\mathbf{Y}}\mathbf{u}_k = \mathbf{XBu}_k \qquad (9)$$

Dans ce cas, les vecteurs d'ordination, appelés 'coordonnées ajustées des sites', sont les combinaisons linéaires des variables explicatives de **X**. Ces vecteurs ont des variances égales aux valeurs propres correspondantes.

Les 'coordonnées des sites' de l'équation (8) sont obtenues par la projection des données initiales de la matrice **Y** sur l'axe k ; elles offrent une approximation des donnée observées qui contiennent les résidus ($\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{Y}_{res}$). Par ailleurs, les 'coordonnées ajustées des sites' de l'équation (9) correspondent à la projection des valeurs ajustées de la matrice $\hat{\mathbf{Y}}$ sur l'axe k ; elles fournissent une approximation des données ajustées. Chacun des deux ensembles de coordonnées peut être utilisé dans les diagrammes de double projection.

5. Une autre information importante, nécessaire pour l'interprétation des relations entre les variables **X** et **Y**, est la contribution des variables explicatives aux différents axes canoniques. Pour estimer cette contribution, on peut calculer les corrélations linéaires entre les variables **X** et les axes d'ordination, soit dans l'espace **Y** (en employant les axes de l'équation (8) soit dans l'espace **X** (en employant ceux de l'équation (9)). Les corrélations entre les variables en **X** et les vecteurs d'ordination dans l'espace **X** sont utilisées pour représenter les variables explicatives dans les diagrammes de double projection.

6. En ACR, les diagrammes de double projection peuvent contenir trois ensembles de données : les coordonnées des sites obtenues des équations (8) ou (9), les variables-réponse de **Y** et les variables explicatives de **X**. Deux types de normalisation («scaling» en anglais) sont le plus souvent employés pour les diagrammes de double projection en ACR. Leur description détaillée se trouve dans ter Braak [12] et Legendre et Legendre [6]. Mentionnons seulement que la normalisation de type 1, qui produit un diagramme préservant les distances entre objets («distance biplot» en anglais), met en évidence les relations entre les sites (voir par exemple Legendre et Legendre [6]) ; la normalisation de type 2, qui produit un diagramme préservant les corrélations («correlation biplot» en anglais), met plutôt en évidence les relations entre les variables-réponse.

Montrons maintenant comment la procédure de régression polynomiale décrite au paragraphe 2 peut être incorporée en ACR. L'utilisation de cette technique dans le cadre de l'ACR nous permet non seulement de modéliser les relations polynomiales entre les ensembles de variables-réponse et les variables explicatives, mais également d'augmenter, dans bien des cas, le pourcentage de variance de la matrice des variables-réponse **Y** qui est expliqué par l'analyse, par rapport à l'ACR linéaire classique. La différence principale avec l'ACR linéaire consiste en application de la procédure du paragraphe 2, au lieu de la régression linéaire multiple, à la première étape de l'ACR. Un autre changement important intervient à la cinquième étape de l'ACR où des corrélations linéaires multiples sont calculées, au lieu des corrélations linéaires simples employée

dans le cas linéaire. Les autres étapes de l'ACR décrites ci-dessus resteront sans modification.

Au paragraphe 2, nous avons établi que la matrice des valeurs ajustées $\hat{\mathbf{Y}}$ obtenue par la procédure de régression polynomiale ne produit pas une combinaison linéaire des variables explicatives de $\mathbf{X}$, mais plutôt une combinaison polynomiale où le degré de chacune des variable est limité à 2. Par conséquent, à la première étape de l'ACR polynomiale, l'équation (5) est remplacée par :

$$\hat{\mathbf{Y}} = P^2(\mathbf{X}) \qquad (5')$$

où P^2 désigne les polynômes où le degré maximal de chacune des variables est limité à 2 ; leur structure peut être différente pour chacune des variables-réponse $\mathbf{y}$ de $\mathbf{Y}$. À la seconde étape, la matrice de covariance $\mathbf{S}$ de $\hat{\mathbf{Y}}$ est calculée. La troisième étape produit la décomposition de $\mathbf{S}$ en valeurs et vecteurs propres ; les coordonnées des sites servant à représenter les variables $\mathbf{Y}$ dans les diagrammes de double projection sont calculées à l'aide des équations (8) et (9). En ACR polynomiale, la matrice des vecteurs propres canoniques $\mathbf{U}$ est d'ordre $(p \times l)$ où l peut varier de 1 à p.

L'ACR polynomiale utilise les *corrélations linéaires multiples* au lieu des *corrélations linéaires simples* pour déterminer les coordonnées des variables explicatives dans les diagrammes de double projection. Une telle approche est mise en œuvre parce que chacun des vecteurs canoniques d'ordination est maintenant une combinaison polynomiale des variables explicatives de $\mathbf{X}$:

$$\mathbf{Ord}_{(\text{espace des variables explicatives } \mathbf{X})k} = \hat{\mathbf{Y}}\,\mathbf{u}_k = P^2(\mathbf{X})\mathbf{u}_k \qquad (9')$$

Pour obtenir les coordonnées d'une variable explicative $\mathbf{x}$ sur un axe canonique dans un diagramme de double projection, on calcule d'abord la *corrélation linéaire multiple* $R_{\mathbf{ord},\{\mathbf{x},\mathbf{x}^2\}}$ entre le vecteur de coordonnées des sites **ord** (obtenu à partir de l'équation 9'), correspondant à l'axe canonique considéré, et les vecteurs $\mathbf{x}$ et $\mathbf{x}^2$. Le signe de la *corrélation linéaire simple* entre **ord** et $\mathbf{x}$ est attribué à $R_{\mathbf{ord},\{\mathbf{x},\mathbf{x}^2\}}$ ce qui donne la coordonnée de la variable explicative $\mathbf{x}$ sur cet axe canonique.

4. EXEMPLE NUMÉRIQUE D'UTILISATION DE L'ACR POLYNOMIALE

Ce paragraphe est consacré à l'étude d'un exemple illustrant la mise en œuvre de l'ACR polynomiale sur un ensemble des données réelles. Nous avons choisi des données écologiques recueillies à 20 sites de l'étang de Thau (côte méditerranéenne de la France) dans le cadre du programme ÉCOTHAU (Amanieu *et al.*, [1]). Les données, présentées au tableau 1 ci-dessous, sont tirées de Legendre et Legendre ([6], tableau 13.2).

Dans ce tableau, les deux variables-réponse *Bna* et *Ma* représentent des quantités de deux types de bactéries hétérotrophes aérobies dans les sites examinés, après transformation logarithmique des abondances ; elles forment la matrice des variables dépendantes $\mathbf{Y}$. La matrice des variables explicatives $\mathbf{X}$ contient deux variables chimiques, NH_4 et la *production bactérienne*, ainsi qu'une *variable spatiale* qui représente les coordonnées des sites le long du grand axe géographique de l'étang de Thau.

L'ACR linéaire et polynomiale ont été calculées pour les données du tableau 1. Les résultats ci-dessous ont été obtenus : deux axes canoniques ont été trouvés par l'ACR linéaire et polynomiale. La méthode polynomiale explique 67,62 % de la variance totale de la matrice des variables-réponse **Y** alors que la méthode linéaire n'explique que 18,70 % de la variance de **Y**. Les résultats des tests de signification des deux modèles sont présentés plus bas ; nous verrons que seul le modèle polynomial est adéquat pour expliquer les variables-réponse *Bna* et *Ma*.

Sites	Bna	Ma	NH_4	Prod. bact.	Var. spatiale
N°	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4,615	10,003	0,307	0,274	–9,4173
2	5,226	9,999	0,207	0,213	–7,1865
3	5,081	9,636	0,140	0,134	–5,8174
4	5,278	8,331	1,371	0,177	–6,8322
5	5,756	8,929	1,447	0,091	–4,6014
6	5,328	8,839	0,668	0,272	–4,2471
7	4,263	7,784	0,300	0,460	–1,8632
8	5,442	8,023	0,329	0,253	–0,4940
9	5,328	8,294	0,207	0,235	0,8751
10	4,663	7,883	0,223	0,362	–0,1398
11	6,775	9,741	0,788	0,824	–1,1546
12	5,442	8,657	1,112	0,419	0,2145
13	5,421	8,117	1,273	0,398	4,9824
14	5,602	8,117	0,956	0,172	3,9676
15	5,442	8,487	0,708	0,141	3,4602
16	5,303	7,955	0,637	0,360	6,3515
17	5,602	10,545	0,519	0,261	5,8441
18	5,505	9,687	0,247	0,450	4,8293
19	6,019	8,700	1,664	0,287	4,6762
20	5,464	10,240	0,182	0,510	6,5527

Tableau 1. Données recueillies à 20 sites de l'étang de Thau le 25 octobre 1988.

Les équations quadratiques générées par le programme pour les deux variables-réponse du tableau 1 sont décrites ci-dessous. Les équations obtenues par l'algorithme empirique de régression polynomiale du paragraphe 2 pour ajuster la première variable bactérienne ($\mathbf{y}_1$) par une combinaison des variables explicatives $\mathbf{x}_1$, $\mathbf{x}_2$, et $\mathbf{x}_3$ sont :

$$\mathbf{x}_{23} = -5{,}62\,\mathbf{x}_2 + 0{,}02\,\mathbf{x}_3 + 0{,}09\,\mathbf{x}_2\mathbf{x}_3 + 7{,}64\,\mathbf{x}_2^2 + 0{,}0008\,\mathbf{x}_3^2 + 1{,}08 \qquad (a)$$

$$\mathbf{x}_{123} = 0{,}27\,\mathbf{x}_1 + 1{,}23\,\mathbf{x}_{23} - 0{,}27\,\mathbf{x}_1\mathbf{x}_{23} + 0{,}10\,\mathbf{x}_1^2 - 0{,}003 \qquad (b)$$

$$\hat{\mathbf{y}}_1 = \mathbf{x}_{123} + 4{,}82 \qquad (c)$$

En ce qui concerne la deuxième variable bactérienne ($\mathbf{y}_2$), les équations sont les suivantes :

$$\mathbf{x}_{23} = -5{,}87\,\mathbf{x}_2 - 0{,}34\,\mathbf{x}_3 + 0{,}23\,\mathbf{x}_2\mathbf{x}_3 + 8{,}42\,\mathbf{x}_2^2 + 0{,}0235\,\mathbf{x}_3^2 + 0{,}44 \qquad (a)$$

$$\mathbf{x}_{123} = -0{,}82\,\mathbf{x}_1 + 1{,}60\,\mathbf{x}_{23} - 1{,}28\,\mathbf{x}_1\mathbf{x}_{23} + 0{,}40\,\mathbf{x}_1^2 - 0{,}01 \qquad (b)$$

$$\hat{\mathbf{y}}_2 = \mathbf{x}_{123} + 8{,}99 \qquad (c)$$

Les équations ci-dessus illustrent le processus d'approximation des variables-réponse : lors de la première itération de l'algorithme de régression polynomiale, les variables explicatives $\mathbf{x}_2$ et $\mathbf{x}_3$ ont été fusionnées (équations *a*) pour former une nouvelle variable combinée $\mathbf{x}_{23}$. Cette nouvelle variable a été fusionnée à la variable $\mathbf{x}_1$ (équations *b*) lors de la deuxième itération, formant la variable combinée $\mathbf{x}_{123}$. La variable-réponse **y** a enfin été régressée sur cette dernière variable combinée pour fournir l'approximation finale.

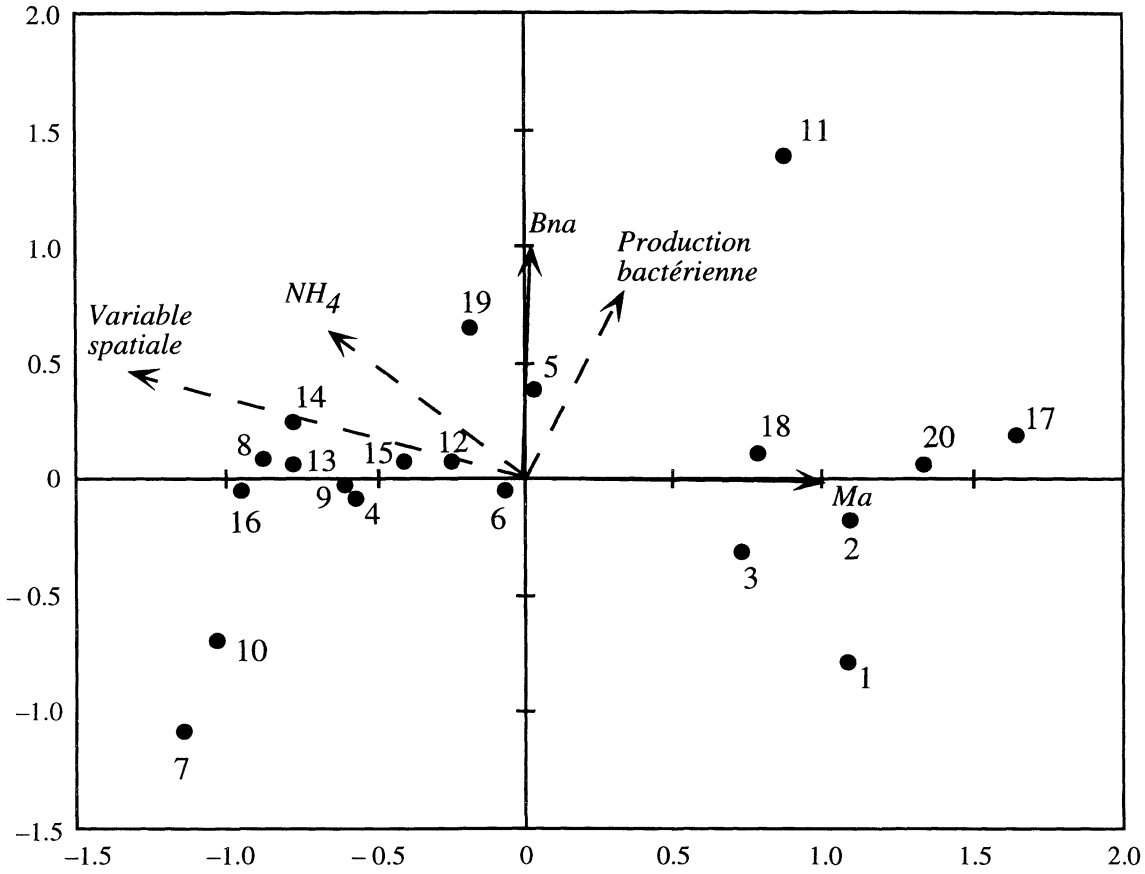

Figure 1. Diagramme de double projection de l'ACR pour les données du tableau 1. Le diagramme représente les espèces (flèches pleines), les sites (points avec les numéros) et les variables explicatives (flèches en tirets). Les coordonnées des variables explicatives ont été multipliées par le facteur 2.5 pour assurer une meilleure représentation.

Nous ne présentons pas les autres résultats numériques fournis par notre programme. Nous nous limitons à la présentation du diagramme de double projection correspondant à l'ACR polynomiale. Ce diagramme, présenté à la figure 1, contient trois types de données : les sites, les variables-réponse et les variables explicatives. Nous avons choisi la normalisation («scaling») de type 1 qui préserve les distances euclidiennes entre les sites. La disposition des sites est approximativement la même dans le diagramme de l'ACR linéaire (non présenté) et polynomiale (Figure 1). Les variables environnementales *production bactérienne* et *variable spatiale* sont plus fortement corrélées (corrélations multiples) aux axes de l'ACR polynomiale qu'aux axes de l'ACR linéaire ; ces variables sont donc mieux représentées dans le diagramme de l'ACR polynomiale, ce qui est indiqué par des flèches plus longues. Les relations entre les

variables bactériennes (*Bna, Ma*) et les variables environnementales, représentées par les angles, ne diffèrent que légèrement entre les deux diagrammes de double projection.

Des tests par permutation ont été réalisés pour les deux modèles de régression afin de déterminer leur niveau de signification. Pendant ces tests, les lignes de la matrice des variables-réponse **Y** ont été permutées par rapport aux lignes de la matrice des variables explicatives **X**, produisant une matrice $\mathbf{Y}_{\mathbf{perm}}$. À chaque permutation, une matrice de variables ajustées $\hat{\mathbf{Y}}_{\mathbf{perm}}$ a été obtenue de $\mathbf{Y}_{\mathbf{perm}}$ par application de la procédure de régression en question. Une statistique pseudo-*F* a été calculée pour chacune des matrices permutées $\mathbf{Y}_{\mathbf{perm}}$ à l'aide de la formule suivante:

$$pseudo\text{-}F = \frac{variance\ de\ \hat{\mathbf{Y}}_{\mathbf{perm}}}{variance\ totale\ de\ \mathbf{Y} - variance\ de\ \hat{\mathbf{Y}}_{\mathbf{perm}}} \qquad (10)$$

On compte combien de fois la valeur de la statistique pseudo-*F* associée à une matrice permutée $\mathbf{Y}_{\mathbf{perm}}$ est supérieure ou égale à la valeur de la statistique associée à la matrice initiale **Y**, ce qui permet de déterminer la probabilité associée au modèle pour les données en question. Dans cette étude, nous avons généré 999 matrices randomisées $\mathbf{Y}_{\mathbf{perm}}$ en calculant les valeurs correspondantes de la statistique pseudo-*F* pour les deux méthodes de régression. Cette dernière statistique a servis à calculer la probabilité (*P*) que le modèle de la régression choisi n'explique pas les données observées. Généralement, les valeurs de la statistique *P* inférieures à 0,05 sont considérées significatives.

Les tests par permutation, réalisés sur les données du tableau 1 ont abouti aux résultats suivants : la valeur de *P* pour le modèle linéaire est égale à 0,312, tandis que la valeur de *P* pour le modèle polynomial est 0,011. Le modèle polynomial est donc le seul modèle significatif pour les données du tableau 1. Nous avons également testé la signification de la *quantité additionnelle de variance* expliquée par le modèle polynomial par rapport au modèle linéaire, en utilisant la procédure de test par permutation. Une procédure supplémentaire, dont les détails ne sont pas décrits ici, a été mise en œuvre pour tester la signification de la différence de variance expliquée par les deux méthodes. La valeur de *P* pour la différence des variances expliquées, obtenue par cette procédure après 999 permutations, est égale à 0,031, ce qui confirme l'intérêt du modèle polynomial par rapport au modèle linéaire pour expliquer les données du tableau 1.

Nous ne pouvons garantir que des résultats aussi intéressants seront produits dans tous les cas par la méthode polynomiale. Plusieurs essais comparatifs des méthodes linéaire et polynomiale ont montré que les résultats dépendent vraiment des données examinées, si bien qu'on ne peut prédire *a priori* laquelle des deux méthodes fournira les résultats les plus significatifs. Cependant, ces essais ont montré que l'algorithme empirique de régression polynomiale décrit au paragraphe 2 parvient souvent à expliquer une partie de la variance de la matrice des variables-réponse qui reste inexpliquée par l'ACR basée sur la régression linéaire multiple.

5. PROPRIÉTÉS PRINCIPALES DU NOUVEL ALGORITHME DE RÉGRESSION POLYNOMIALE ET DE L'ACR POLYNOMIALE

Nous avons décrit un algorithme empirique de régression polynomiale qui peut être incorporé dans l'analyse canonique de redondance pour modéliser les relations polynomiales entre les matrices de variables-réponse et de variables explicatives. L'utilisation de cet algorithme dans le cadre de l'ACR polynomiale permet généralement d'augmenter le pourcentage de variance expliqué, par rapport à l'ACR classique basée sur la régression linéaire.

Notre algorithme de régression polynomiale, calculé indépendamment pour chacune des variables-réponse **y** (i.e. pour chacune des colonnes) de la matrice **Y**, assure l'amélioration successive de l'approximation obtenue par la régression linéaire multiple de **y** sur la matrice des variables explicatives **X**, contenant m colonnes. Ainsi, pour chacune des variables-réponse **y** de **Y**, la matrice **X** est réduite $m - 1$ fois ; chaque fois, la matrice des résidus du vecteur **y** est calculée et régressée sur les deux variables (colonnes) de **X** les plus appropriées. Ces deux variables, qui sont sélectionnées lors d'une procédure locale d'approximation quadratique, sont ensuite combinées pour former une nouvelle variable (colonne) qui les remplace dans la matrice **X** pour la suite de la procédure de régression.

Les polynômes générés pour les variables-réponse de **Y** ne sont pas toujours les mêmes ; ils peuvent varier d'une variable **y** à l'autre. Cette différence peut être utilisée pour regrouper les variables-réponse en classes, par exemple.

Une nouvelle façon de calculer les coordonnés des variables explicatives est proposée pour l'ACR polynomiale. Il s'agit d'employer la corrélation linéaire multiple au lieu de la corrélation linéaire simple utilisée en ACR classique et de lui adjoindre le signe de la corrélation linéaire simple.

L'ACR polynomiale ne garantit pas que le modèle polynomial sera toujours supérieur au modèle linéaire pour expliquer la variance de la matrice des variables-réponse. Si les deux modèles sont significatifs pour les données observées, la différence entre la variance de **Y** expliquée par le modèle polynomial et par le modèle linéaire peut être très importante. Au cours de nombreux essais réalisés sur des données artificielles et réelles de types variés, l'ACR polynomiale a permis, dans plusieurs cas, d'expliquer une partie de la variance de **Y** qui restait inexpliquée par l'ACR linéaire classique.

BIBLIOGRAPHIE

[1] AMANIEU M., LEGENDRE P., TROUSSELLIER M., FRISONI G.-F., «Le programme Écothau : théorie écologique et base de la modélisation», *Oceanol. Acta*, 12, (1989), 189-199.

[2] BREIMAN, L., FRIEDMAN, J. H., «Estimating optimal transformations for multiple régression and correlation», *J. Amer. Statist. Assoc.*, 80, (1985), 580-598.

[3] DE VEAUX, R., «Finding transformations for regression using ACE algorithm», *Sociol. Methods & Res.*, 18, (1989), 327-359.

[4] DURAND, J. F., «Generalized principal component analysis with respect to instrumental variables via univariate spline transformations», *Computational Statistics & Data Analysis*, 16, (1993), 423-440.

[5] HILL, M. O., «Reciprocal averaging : an eigenvector method of ordination», *J. Ecol.*, 61, (1973), 237-249.

[6] LEGENDRE, P. et LEGENDRE, L., *Numerical ecology*, 2nd English edition, Amsterdam, Elsevier Science BV, 1998.

[7] RAMSAY, J. O., «Monotone regression splines in action», *Stat. Sci.*, 3, (1988), 426-461.

[8] RAO, C. R., «The use and interpretation of principal component analysis in applied research», *Sankhyaá, Ser. A*, 26, (1964), 329-358.

[9] TER BRAAK, C. J. F., «Ordination», *Data analysis in community and landscape ecology*, R. H. G. Jongman, C. J. F. ter Braak and O. F. R. Van Tongeren (eds.), Wageningen, Pudoc, (1987), 91-173.

[10] TER BRAAK, C. J. F., *CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis* (version 2.1), Wageningen, Agricultural Mathematics Group, Ministry of Agriculture and Fisheries, 1988.

[11] TER BRAAK, C. J. F., *Update notes : CANOCO* (version 3.10), Agricultural Mathematics Group, Wageningen, 1990.

[12] TER BRAAK, C. J. F., «Canonical community ordination. Part I : Basic theory and linear methods», *Écoscience*, 1, (1994), 127-140.

[13] VAN DER BURG, E. et DE LEEUW, J., «Non-linear canonical correlation», *British J. Math. Statist. Psych.*, 36, (1983), 54-80.

[14] WINSBERG, S. et RAMSAY, J. O., «Monotonic transformations to additivity using splines», *Biometrica*, 67, (1980), 669-674.