ACPVI MULTIBLOC. APPLICATION EN ÉPIDÉMIOLOGIE ANIMALE

Stéphanie BOUGEARD [1],
Mohamed HANAFI [2], El Mostapha QANNARI [2]

RÉSUMÉ

Des analyses factorielles permettant l'étude conjointe de $(K+1)$ tableaux de données sont proposées pour le cas où un tableau Y est prédit à l'aide de K tableaux X_k $(k = 1, \ldots, K)$. Ces méthodes factorielles sont basées sur une extension de l'Analyse en Composantes Principales sur Variables Instrumentales $(ACPVI)$, appelée aussi analyse des redondances. Une discussion sur le positionnement des méthodes développées par rapport aux méthodes existantes est présentée. La démarche est illustrée sur la base d'une étude de cas en épidémiologie animale.

Mots-clés : Analyse en composantes principales sur variables instrumentales, analyse des redondances, analyse multibloc.

ABSTRACT

We discuss factor analytic methods to study a set of $(K + 1)$ data tables where we wish to predict a data set Y from K other data sets X_k $(k = 1, \ldots, K)$. The methods of analysis are based on an extension of Principal Component Analysis on Instrumental Variables, also called Redundancy Analysis. We outline the general approach and show its relationships to existing methods. The method of analysis is illustrated on the basis of a case study in animal epidemiology.

Keywords : Principal component analysis on instrumental variables, redundancy analysis, multiblock analysis.

1. Introduction

1.1. Introduction à l'épidémiologie analytique

L'épidémiologie est l'étude des maladies et des facteurs de santé dans une population. Les caractéristiques épidémiologiques sont souvent multi-factorielles et résultent d'interactions entre un ou plusieurs agents pathogènes, une population et le milieu dans lequel vit celle-ci. Les propriétés de chacun de ces trois

1. AFSSA, Département d'épidémiologie animale – Beaucemaine, BP53, 22440 Ploufragan. E-mail : s.bougeard@ploufragan.afssa.fr

2. ENITIAA-INRA, Unité de Sensométrie et Chimiométrie – Rue de la Géraudière BP 82225, 44322 Nantes Cedex.

éléments conditionnent l'épidémiologie de la maladie. L'épidémiologie analytique est un outil permettant l'étude des causes apparentes et des évènements directement ou indirectement associés au phénomène de santé étudié. Ce type d'étude vise à mesurer l'intensité de la liaison entre les facteurs étudiés et la maladie, et ainsi déterminer les facteurs de risque associés au développement de la maladie.

1.2. Description des données

Les études d'épidémiologie analytique dans le domaine animal sont réalisées au travers d'enquêtes menées en élevage et à l'abattoir. Elles sont structurées en thèmes : les caractéristiques de la ferme (taille de l'élevage, performances zootechniques, autres productions animales, ...), la conduite d'élevage (taux de renouvellement, technique de reproduction, nombre d'animaux par portée, nombre de bandes d'animaux, ...), l'habitat des animaux (enregistrements bio-climatiques, ventilation, isolation, chauffage, ...), l'alimentation et l'abreuvement des animaux (enregistrements alimentaires, mode de distribution, nombre de mangeoires, origine des aliments, ...), l'état sanitaire du troupeau (dosages sérologiques, pesées, maladies chroniques, taux de réformes, vaccinations, traitements antibiotiques, ...), les pratiques d'hygiène (protocoles de nettoyage et de désinfection) et les mesures de bio-sécurité (mesures sanitaires de l'éleveur et des visiteurs, équarrissage, ...). Un exemple de structure simplifiée des données est résumé dans la figure 1.

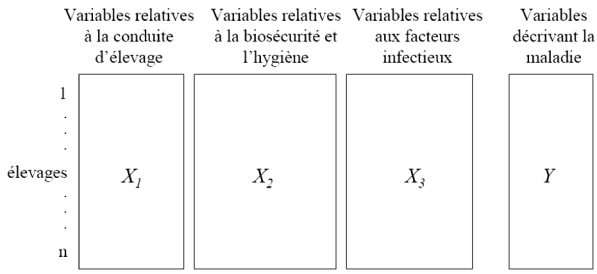

FIG 1. — Illustration de la structure classique des données d'épidémiologie animale.

1.3. Objectifs du traitement statistique

Les objectifs du traitement statistique des données d'épidémiologie animale sont à la fois descriptifs et prédictifs. La description des données passe par la compréhension du mode de fonctionnement de l'ensemble des élevages étudiés (études des blocs X_k et Y ainsi que des variables à l'intérieur des blocs), mais aussi par la visualisation des ressemblances et différences entre les élevages (étude des individus). Pour cela, des représentations factorielles synthétiques, orientées vers l'explication de la maladie, permettent de comprendre les liens complexes entre variables et blocs de variables. La modélisation des données est réalisée en parallèle à l'étude descriptive et permet de déterminer les variables qui peuvent réduire l'apparition ou la prévalence de la maladie. Parmi les variables X, sont sélectionnés les facteurs de risque de la maladie

Y. Du fait du grand nombre de variables X et du sens biologique des blocs dans lesquels celles-ci sont organisées (groupe de variables liées à l'hygiène par exemple), il apparaît utile de mesurer à la fois l'importance des variables mais aussi des blocs de variables dans l'explication de la maladie. Le développement de méthodes statistiques en vue du traitement des données d'épidémiologie animale doit prendre en compte à la fois la particularité des données et des objectifs de traitement associés.

1.4. Position du problème

La problématique statistique relève à la fois de la description de tableaux et de la modélisation à partir de tableaux multiples. Les données sont organisées en $(K+1)$ tableaux : K tableaux X_k orientés vers l'explication d'un tableau Y. Les variables de tous ces tableaux sont mesurées sur les mêmes individus. L'Analyse en Composantes Principales sur Variables Instrumentales ($ACPVI$) (Rao, 1964 ; Van Den Wollenberg, 1977 ; Sabatier, 1987a est une méthode adaptée à l'explication d'un tableau Y par un tableau X. Des variables latentes, combinaisons linéaires des variables X orientées vers l'explication du tableau Y, sont utilisées pour décrire les liaisons entre les variables X et Y au travers de représentations factorielles. De plus, ces variables latentes peuvent être utilisées pour modéliser le tableau Y. Nous proposons des méthodes, dérivant d'extensions de l'Analyse en Composantes Principales sur Variables Instrumentales, dans le cas où le tableau X est structuré en K sous-tableaux. Parmi celles-ci, nous distinguons une méthode que nous désignons par $ACPVI$ multibloc, basée sur l'extraction de composantes globales, synthèses de composantes partielles liées à chacun des tableaux. La démarche adoptée privilégie le cadre de l'$ACPVI$ pour étudier la prédiction d'un tableau à l'aide d'un ensemble d'autres tableaux. Il faut souligner que d'autres approches ont été proposées pour l'analyse de $(K+1)$ tableaux, dans des contextes privilégiant d'autres méthodes. Nous pouvons citer les travaux de Kissita (2003) privilégiant le cadre de l'analyse canonique, ceux de Vivien (2002) privilégiant le cadre de la régression PLS, ainsi que les méthodes PLS multibloc (Wold, 1984) et PLS multibloc hiérarchique (Wold *et al.*, 1996). L'approche PLS offre un cadre général à un grand nombre de ces méthodes (Wold, 1982 ; Tenenhaus, 1998, 1999 ; Tenenhaus *et al.*, 2005a,b).

2. Méthode

2.1. $ACPVI$ d'un tableau Y expliqué par un tableau X

L'objectif est de réaliser une représentation factorielle ainsi qu'une prédiction du tableau Y contenant Q variables $(y_1, \ldots, y_Q)$ à partir de P variables $X = (x_1, \ldots, x_P)$. Toutes ces variables sont mesurées sur les mêmes n individus et supposées centrées. On adopte, ici et dans la suite de l'article, la métrique unité usuelle associée à l'espace vectoriel considéré. La détermination de la première composante de l'$ACPVI$ telle qu'elle est proposée par Rao (1964) est basée sur la minimisation du critère (1) :

$$\text{Minimiser} \quad \left\| YY' - \lambda tt' \right\|^2 \quad \text{avec} \quad t = Xw \tag{1}$$

Concrètement, ceci revient à minimiser la distance entre la matrice de produits scalaires entre individus dans l'espace des variables de Y et la représentation des individus sur la première composante t, contrainte d'être dans l'espace engendré par les variables de X. Du fait de son intérêt aussi bien théorique que pratique, plusieurs auteurs ont donné différentes interprétations de l'$ACPVI$ en montrant qu'elle peut être définie comme une ACP des projections des variables de Y dans l'espace engendré par les variables de X (voir par exemple Saporta (2006)) ou qu'elle peut naturellement s'intégrer dans le cadre de l'approche PLS appliquée à deux blocs (Tenenhaus, 1998). Parmi ces auteurs, nous distinguons Van Den Wollenberg (1977) et Sabatier (1987a,b) qui ont démontré plusieurs propriétés de cette méthode et bien défini son positionnement par rapport à l'analyse en composantes principales (ACP) et l'analyse canonique. Nous retiendrons ici la formulation qui fait apparaître l'$ACPVI$ de Y par rapport à X comme étant une ACP de Y sous la contrainte que les composantes principales t sont des combinaisons linéaires des variables constituant X (variables instrumentales). En effet, la première composante t de l'$ACPVI$ de Y par rapport à X peut être déterminée de manière à maximiser le critère (2).

$$\text{Maximiser} \quad \sum_{q=1}^{Q} cov^2(y_q, t) \quad \text{avec} \quad t = Xw \quad \text{et} \quad ||t|| = 1 \tag{2}$$

Ce problème a pour solution $w^{(1)}$, vecteur propre de la matrice $[(X'X)^{-1}X'YY'X]$ associé à la plus grande valeur propre $\lambda^{(1)}$. Les composantes de l'$ACPVI$ d'ordre supérieur à 1 sont obtenues en considérant une démarche itérative consistant, à chaque pas, à déterminer une composante t standardisée, maximisant le même critère, en imposant la contrainte supplémentaire que cette composante soit orthogonale aux composantes déterminées lors des étapes précédentes. Par souci de cohérence avec la démarche qui sera poursuivie dans le cadre de l'$ACPVI$ multibloc, nous proposons de déterminer les composantes selon la procédure de déflation préconisée dans le cadre de la régression PLS. Celle-ci consiste à considérer, à chaque pas, les résidus de la régression des variables de X sur les composantes déterminées aux étapes précédentes. S'agissant de composantes orthogonales, il apparaît intuitivement que ces deux procédures (contrainte d'orthogonalité et déflation) sont équivalentes. Cette équivalence à été montrée de manière formelle par Tenenhaus (1998, p. 131) dans le cadre de la régression PLS et par Nocairi *et al.* (2005) dans le cadre de l'analyse discriminante. Cette démonstration peut être facilement adaptée au présent contexte.

Nous pouvons également montrer que la composante t, solution du problème d'optimisation (2), est également solution du problème (3), consistant à déterminer deux variables latentes, $t = Xw$ et $u = Yv$, de manière à maximiser le critère :

$$\text{Maximiser} \quad cov^2(u, t) \quad \text{sous les contraintes} \quad ||t|| = ||v|| = 1 \tag{3}$$

Afin de démontrer cette propriété, nous pouvons remarquer que $cov^2(u,t) = (1/n^2)(u't)^2 = (1/n^2)(v'Y't)^2$. Le maximum, pour v, de cette quantité est atteint pour $v = Y't/||Y't||$, ce qui permet d'obtenir en définitive $cov^2(u,t) = (1/n^2)t'YY't = \sum_q cov^2(y_q,t)$. Nous reconnaissons dans cette dernière expression le critère (2). Ceci corrobore les liens déjà démontrés entre l'*ACPVI*, la régression *PLS* et l'analyse inter-batterie de Tucker (Van der Geer, 1984 ; Chessel et Mercier, 1993 ; Burnham *et al.*, 1996).

L'*ACPVI* peut être vue comme étant une extension de la régression linéaire multiple à l'explication simultanée de plusieurs variables Y. La méthode dite *reduced-rank regression* (Muller, 1981 ; Davies et Tso, 1982) préconise que chaque variable du tableau Y soit modélisée par régression sur les composantes $(t^{(1)}, \ldots, t^{(h)})$ déterminées à l'aide de l'*ACPVI*.

Il faut noter que comme la solution de l'*ACPVI* est basée sur l'inversion de la matrice $(X'X)$, des problèmes d'instabilité peuvent surgir en présence de quasi-colinéarité entre les variables du tableau X, comme cela est d'ailleurs le cas pour la régression linéaire multiple (Cazes, 1975). Parmi les méthodes préconisées pour contrer cette difficulté, nous pouvons citer la régression *ridge* (Hoerl et Kennard, 1970) et la régression *PLS* (Wold *et al.*, 1983 ; Tenehaus, 1998).

2.2. Extensions de l'*ACPVI* à l'explication d'un tableau Y par un tableau X partitionné en K blocs

2.2.1. Notations

Nous disposons à présent d'un tableau de variables quantitatives X partitionné en K blocs, $X = [X_1|\ldots|X_K]$, et d'un tableau Y contenant Q variables quantitatives à expliquer. Chaque tableau X_k est un tableau $(n \times p_k)$ dont les lignes correspondent aux mêmes individus. Soit $P = \sum_k p_k$ le nombre total de variables du tableau concaténé X. Le tableau Y est mesuré sur les mêmes n individus. Par la suite, toutes ces variables sont supposées être centrées. Les objectifs sont de décrire les liens entre les variables X et Y et d'expliquer le tableau Y à partir des K tableaux X_k $(k = 1, \ldots, K)$. Pour cela, des composantes $t = Xw$, qui constituent des résumés du tableau concaténé $X = [X_1|\ldots|X_K]$ sont déterminées. Chaque composante globale t est supposée être également la synthèse des K composantes partielles $t_k = X_k w_k$ qui constituent elles-mêmes des résumés des tableaux X_k $(k = 1, \ldots, K)$. Les composantes partielles peuvent servir à des fins d'interprétation des résultats ou à des fins de prédiction des variables Y.

2.2.2. ACPVI des tableaux Y et X_k

Afin de déterminer les composantes partielles t_k, synthèses des liens entre chaque tableau X_k et le tableau Y, une première solution consiste à maximiser le critère (4) :

$$\sum_{k=1}^{K}\sum_{q=1}^{Q} cov^2(y_q, t_k) \quad \text{avec} \quad t_k = X_k w_k \quad \text{et} \quad ||t_k|| = 1 \tag{4}$$

La solution de ce problème revient en définitive à effectuer de manière indépendante K $ACPVI$ des couples de tableaux (Y, X_k). Cette méthode ne répond qu'à une partie des objectifs formulés précédemment. Les composantes partielles obtenues sont intéressantes du point de vue de l'explication des liens entre chaque tableau X_k et le tableau Y. Cependant, ces $ACPVI$ restituent l'inertie de Y dans des directions propres à chaque tableau X_k. Les différentes composantes, déterminées de manière indépendante les unes des autres, ne véhiculent pas nécessairement des informations concordantes. De ce fait, l'interprétation des résultats peut s'avérer fastidieuse par manque d'une vision synthétique.

2.2.3. ACPVI des tableaux Y et X

Une seconde solution consiste à chercher à la fois les composantes globales $t = Xw$ associées au tableau concaténé $X = [X_1|\dots|X_K]$ et les composantes partielles $t_k = X_k w_k$ associées aux tableaux X_k, de manière à maximiser le critère (5) :

$$\sum_{q=1}^{Q} cov^2(y_q, t) \text{ avec } t = \sum_{k=1}^{K} a_k t_k, \quad t_k = X_k w_k, \quad ||t|| = 1 \quad \text{et} \quad ||t_k|| = 1 \quad (5)$$

La solution de ce problème est donnée par l'$ACPVI$ de Y par rapport au tableau concaténé X. En effet, soit $t = Xw$ avec w, vecteur propre associé à la plus grande valeur propre de la matrice $[(X'X)^{-1}X'YY'X]$. Ce vecteur w est partitionné en K sous-vecteurs w_k relatifs aux différents tableaux X_k de la façon suivante : $w = [w_1|\dots|w_K]'$. Si l'on pose $t_k^* = X_k w_k$, la solution du problème (5) consiste à prendre $t_k = t_k^*/||t_k^*||$ et $a_k = ||t_k^*||$.

Cette solution, à première vue satisfaisante, présente toutefois un inconvénient. En effet, la méthode étant basée sur l'$ACPVI$ des tableaux Y et X ne prend pas en compte la structure en blocs du tableau X pour la détermination des composantes globales t. De plus, elle implique l'inversion de la matrice de variance-covariance $(X'X)$. Ceci pourrait constituer un problème lié à l'instabilité des résultats lorsque les variables du tableau concaténé X présentent des quasi-colinéarités. Le problème se pose de manière plus aiguë que précédemment, car même s'il n'y a pas a priori de redondance à l'intérieur des blocs X_k pris séparément, il se peut que la concaténation des tableaux X_k engendre des problèmes de quasi-colinéarité. Afin de contourner cette difficulté souvent rencontrée en pratique, nous proposons une démarche basée sur un critère original.

2.2.4. ACPVI multibloc

Comme dans le paragraphe 2.2.3, la méthode proposée permet d'extraire directement une composante globale $t = Xw$ orientée vers l'explication du tableau Y, cette composante étant contrainte à être la synthèse des composantes partielles t_k. Le critère (5) ci-après constitue une variante du critère (6), car il est basé sur la même fonction à maximiser, mais d'autres

contraintes sont imposées. Le problème (6) consiste à maximiser :

$$\sum_{q=1}^{Q} cov^2(y_q, t) \text{ avec } t = \sum_{k=1}^{K} a_k t_k,\ t_k = X_k w_k, \quad \sum_{k=1}^{K} a_k^2 = 1 \text{ et } ||t_k|| = 1 \quad (6)$$

Afin de déterminer les solutions de ce problème, nous proposons de développer des propriétés liées au critère (6). Dans un premier temps, nous montrons que la composante t, solution du critère (6), est également solution du critère (7) :

$$cov^2(u,t) \quad \text{avec} \quad t = \sum_{k=1}^{K} a_k t_k, \quad t_k = X_k w_k, \quad u = Yv, \qquad (7)$$
$$\sum_{k=1}^{K} a_k^2 = 1 \quad \text{et} \quad ||t_k|| = ||v|| = 1$$

L'avantage de cette formulation est d'exhiber une composante $u = Yv$ dans l'espace engendré par les variables de Y, liée de manière optimale à la composante globale t. Pour la démonstration de cette propriété, nous pouvons remarquer que la maximisation de $cov^2(u,t) = cov^2(Yv,t)$ par rapport à v, conduit à la solution $v = Y't/||Y't||$. En remplaçant cette solution dans l'expression $cov^2(Yv,t)$, nous trouvons $cov^2(u,t) = (1/n^2)t'YY't = \sum_q cov^2(y_q,t)$. Le critère (8) ci-après, permet de montrer que la composante u est liée aux composantes partielles t_k pour $k = (1,\ldots,K)$. Ce critère stipule que les composantes optimales t_k des problèmes (6) et (7), sont également solutions du problème (8).

$$\sum_{k=1}^{K} cov^2(u, t_k) \quad \text{avec} \quad t_k = X_k w_k, \quad u = Yv, \quad ||t_k|| = ||v|| = 1 \qquad (8)$$

En effet, en remplaçant dans le critère (7) la composante t par $\sum_k a_k t_k$, nous obtenons $cov^2(t,u) = [\sum_k a_k cov(u,t_k)]^2$. La maximisation de ce critère par rapport à a_k pour $k = (1,\ldots,K)$, conduit à $a_k = cov(u,t_k)/\sqrt{\sum_l cov^2(u,t_l)}$. En remplaçant cette valeur dans l'expression ci-dessus, nous obtenons : $cov^2(t,u) = \sum_k cov^2(u,t_k)$. C'est à partir du critère (8) que la solution du problème est développée. Nous avons :

$$\sum_k cov^2(u,t_k) = (1/n^2)\sum_k [w_k' X_k' u]^2$$
$$= (1/n^2)\sum_k [b_k'(X_k'X_k)^{-1/2} X_k' u]^2 \qquad (9)$$

en posant $b_k = (X_k'X_k)^{1/2} w_k$. La contrainte de norme $||t_k|| = 1$ se traduit par $||b_k|| = 1$. Pour u fixé, la valeur optimale de b_k est $b_k = (X_k'X_k)^{-1/2} X_k' u / ||(X_k'X_k)^{-1/2} X_k' u||$. En reportant cette valeur dans l'expres-

sion (9) et en remplaçant u par Yv, il s'ensuit que :

$$\sum_k cov^2(u, t_k) = (1/n^2) \sum_k v'Y'X_k(X_k'X_k)^{-1}X_k'Yv \qquad (10)$$

Ce qui permet de conclure que le vecteur $v^{(1)}$ qui maximise ce critère, est le premier vecteur propre normé, associé à la plus grande valeur propre $\lambda^{(1)}$, de la matrice $H = (1/n^2)\sum_k Y'X_k(X_k'X_k)^{-1}X_k'Y$. La composante associée à ce vecteur dans l'espace engendré par les variables de Y, découle de $u^{(1)} = Yv^{(1)}$. Les composantes partielles sont données par $t_k^{(1)} = X_k w_k^{(1)} = X_k(X_k'X_k)^{-1/2}b_k^{(1)} = P_k u^{(1)}/||P_k u^{(1)}||$, où $P_k = X_k(X_k'X_k)^{-1}X_k'$ est le projecteur sur l'espace engendré par les variables de X_k. Les coefficients sont donnés par $a_k^{(1)} = cov(u^{(1)}, t_k^{(1)})/\sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})} = ||P_k u^{(1)}||/\sqrt{\sum_l ||P_l u^{(1)}||^2}$. La composante globale $t^{(1)}$ est donnée par $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)} = \sum_k P_k u^{(1)}/\sqrt{\sum_k ||P_k u^{(1)}||^2}$.

Il apparaît que la solution du problème considéré est basée sur le premier vecteur propre de $H = (1/n^2)\sum_k Y'X_k(X_k'X_k)^{-1}X_k'Y$. Etant donné que les projecteurs P_k sont symétriques et idempotents, il s'ensuit que $H = (1/n^2)\sum_k (P_kY)'(P_kY)$. La méthode consiste donc à réaliser une ACP du tableau obtenu par concaténation verticale des projections de Y sur les espaces engendrés par les blocs X_k. Cela revient à chercher une direction commune v dans l'espace $\Re^Q$ telle que les nuages associés à P_kY $(k = 1, \ldots, K)$ aient une inertie moyenne restituée la plus élevée possible. La première valeur propre de la matrice H est égale à $\lambda^{(1)} = (1/n^2)\sum_k ||P_k u^{(1)}||^2 = (1/n)\sum_k var(P_k u^{(1)})$. Ainsi, cette valeur est d'autant plus grande qu'il existe une direction dans l'espace Y expliquée par les différents tableaux X_k. Naturellement, la contribution relative du tableau X_k dans cette explication peut être évaluée par $\left((1/n^2)||P_k u^{(1)}||^2\right)/\lambda^{(1)}$, et il est aisé de vérifier que cette quantité est égale à $(a_k^{(1)})^2$.

Afin de déterminer les composantes d'ordre 2, la même démarche est effectuée en remplaçant les tableaux X_k et Y par leurs résidus respectifs de la régression sur la première composante globale $t^{(1)}$. Cette procédure peut être répétée plusieurs fois pour obtenir des composantes globales $(t^{(1)}, \ldots, t^{(h)})$, ainsi que les composantes partielles associées. Les composantes $(t^{(1)}, \ldots, t^{(h)})$ ainsi obtenues peuvent servir à des fins de prédiction, en régressant les variables Y sur celles-ci. Ainsi, la procédure de déflation conduit à l'obtention de composantes globales orthogonales, qui, de proche en proche, restituent la variabilité de Y. Le nombre de composantes à introduire dans le modèle peut être déterminé par une procédure de validation croisée (Stone, 1974), basée sur la minimisation de l'erreur moyenne de prédiction.

La méthode décrite dans ce paragraphe est appelée par la suite $ACPVI$ multibloc.

2.2.5 Cas particuliers

Lorsque le tableau X est composé d'un seul bloc, le critère (6) montre que l'application de l'$ACPVI$ multibloc conduit à l'$ACPVI$. Supposons maintenant que le tableau X soit disposé en autant de blocs qu'il a de variables : $X_1 = [x_1], \ldots, X_P = [x_P]$. Il est clair, à partir du critère (7), que l'application de l'$ACPVI$ multibloc conduit à la régression $PLS2$ de Y sur X^* ; X^* étant obtenu à partir du tableau X par standardisation des variables.

Considérons maintenant le cas de K tableaux $X_1, \ldots, X_K$, et désignons par Y le tableau concaténé $[X_1|\ldots|X_K]$. L'application de l'$ACPVI$ multibloc à ce cas de figure vise en définitive à explorer les relations entre les tableaux X_k. A partir du critère (10), nous pouvons conclure que dans ce cas, nous sommes conduits à chercher une variable $u = Yv$ telle que $||v|| = 1$ qui maximise la quantité $\sum_k u'X_k(X_k'X_k)^{-1}X_k'u = u'(\sum_k P_k)u$. Nous en déduisons que u est un vecteur propre de $\sum_k P_k$ associé à la plus grande valeur propre. Ceci est la solution de la première étape de l'analyse canonique de Carroll (1968).

2.2.6. Comparaison de méthodes

Différentes méthodes sont basées sur la maximisation du critère $\sum_k cov^2(u, t_k)$, critère (8) de l'$ACPVI$ multibloc. Nous pouvons citer l'analyse de concordance généralisée ($CONCORg$) (Lafosse et Hanafi, 1997 ; Kissita *et al.*, 2004), l'analyse de co-inertie multiple orthogonale ($ACIMO$) (Vivien, 2002) et l'analyse canonique généralisée avec tableau de référence ($ACGTR$) (Kissita, 2003). Dans le cas d'un seul tableau Y, la méthode PLS multibloc ($MPLS$) (Wold, 1984) est aussi basée sur la maximisation du même critère, comme cela est montré par Vivien (2002). Les déflations associées à la méthode $MPLS$ sont réalisées sur les composantes t comme recommandé par Westerhuis et Coenegracht (1997). Toutes ces méthodes fournissent des résultats différents car les contraintes portant sur les composantes ou la stratégie de déflation diffèrent (tableau 1).

TABLEAU 1. — Contraintes portant sur les composantes et déflations de différentes méthodes multiblocs permettant le traitement de K tableaux X_k $(k = 1, \ldots, K)$ orienté vers l'explication d'un tableau Y. $u = Yv$ et $t_k = X_k w_k$ sont les composantes dans l'espace des variables Y et X_k respectivement.

Méthode	**Contrainte de norme**	**Déflations**
$ACGTR$	$\|\|u\|\| = \|\|t_k\|\| = 1$	X_k sur t_k
$ACIMO$	$\|\|v\|\| = \|\|w_k\|\| = 1$	X_k sur t_k, Y sur u
$CONCORg$	$\|\|v\|\| = \|\|w_k\|\| = 1$	X_k sur w_k
$MPLS$	$\|\|v\|\| = \|\|w_k\|\| = 1$	X et Y sur t
$ACPVI$ multibloc	$\|\|v\|\| = \|\|t_k\|\| = 1$	X et Y sur t

Du fait des contraintes de norme, il ressort que l'$ACPVI$ multibloc est davantage orientée vers l'explication de Y que les autres méthodes citées. En effet, en imposant aux variables latentes partielles d'être normées, la méthode est focalisée sur la restitution de la variabilité de Y. D'autre part,

la procédure de déflation de l'$ACPVI$ multibloc conduit à l'obtention de variables latentes globales orthogonales, optimisant l'explication de Y par l'ensemble des variables X (Westerhuis et Coenegracht, 1997).

Parmi les méthodes conçues pour la prédiction de Y à partir de plusieurs blocs X_k $(k = 1, \ldots, K)$, nous pouvons également citer la méthode PLS multibloc hiérarchique ($HPLS$) (Wold *et al.*, 1996) dont les objectifs sont similaires à ceux de l'$ACPVI$ multibloc : calculer, dimension par dimension, des composantes partielles (t_k pour X_k, u pour Y) et en déduire une composante globale t, combinaison linéaire des composantes partielles t_k. Dans un premier temps, des composantes partielles t_k et u, associées respectivement aux tableaux X_k et Y, sont déterminées. Par la suite, la composante globale est définie en effectuant une régression $PLS1$ de u sur le tableau concaténé des composantes partielles $[t_1|\ldots|t_K]$. L'algorithme est itératif et converge, permettant ainsi la détermination simultanée de ces composantes partielles et globales (voir Wold *et al.* (1996) pour le détail de l'algorithme). A chaque étape, la méthode considère les résidus de la régression de X et Y sur les composantes globales déterminées précédemment. Le principal inconvénient de la méthode PLS multibloc hiérarchique est qu'elle ne dispose pas d'un critère clair à optimiser et que l'algorithme de résolution est complexe et long en temps de calcul. Westerhuis *et al.* (1998) et Vivien (2002) indiquent que cette méthode détermine des variables latentes globales davantage liées aux blocs X_k $(k = 1, \ldots, K)$ qu'au tableau Y.

3. Application

Nous présentons une application de l'$ACPVI$ multibloc en épidémiologie animale et comparons les résultats de cette méthode avec ceux de la régression PLS multibloc.

3.1. Données d'épidémiologie animale

Les données d'épidémiologie animale sont issues d'une enquête analytique sur la maladie de l'amaigrissement du porcelet (Rose *et al.*, 2003). L'un des principaux facteurs infectieux de cette maladie est le circovirus $PCV2$. Le tableau de données comporte 158 élevages sur lesquels sont mesurées 36 variables organisées en cinq tableaux décrits en annexe A. Le tableau Y, composé de trois variables, mesure le taux d'animaux positifs au $PCV2$ (truies, porcs charcutiers et porcelets). Les variables X sont organisées en quatre tableaux : X_1 relatif aux mesures de bio-sécurité et d'hygiène, X_2 reflétant la conduite d'élevage, X_3 lié à la structure de l'élevage et X_4 relatif aux co-facteurs infectieux et vaccins. Les variables qualitatives ont été codées selon un codage disjonctif complet. Le premier objectif de l'étude est de décrire les liens entre les variables et entre les blocs de variables, et de déterminer les variables permettant de différencier les élevages. Le second objectif est à la fois de déterminer, parmi les variables de X, celles qui sont facteurs de risque ou, au contraire, facteurs protecteurs de la contamination des élevages au $PCV2$ et de mesurer l'influence des blocs de variables X_k $(k = 1, \ldots, 4)$ dans

l'explication de la positivité des élevages au circovirus $PCV2$. Les variables, ayant des unités de mesure différentes, sont centrées et réduites.

3.2. Description de tableaux structurés en blocs

3.2.1. Interprétation des composantes

Le tableau 2 permet de montrer que les trois premières composantes globales de l'$ACPVI$ multibloc sont suffisantes pour décrire le tableau Y (87.2% de l'inertie de Y expliquée).

TABLEAU 2. — Inertie expliquée du tableau Y, ainsi que de chaque variable y_q, par les composantes globales t pour les cinq premières dimensions.

Dimension h	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	...	Somme
$CIRCOPS$	**52,8**	8,8	29,0	3,0	2,6	...	100
$CIRCOPC$	19,3	**65,6**	0,2	8,5	2,4	...	100
$CIRCOTR$	**56,8**	15,5	14,2	1,8	7,2	...	100
Total	**40,6**	**33,4**	13,3	4,8	3,8	...	100

Le tableau 3 donne l'importance des composantes partielles t_k dans la construction de la composante globale t. Celle-ci est reflétée par les coefficients a_k^2 qui, rappelons-le, sont déterminés de manière à avoir $t = \sum_k a_k t_k$ et $\sum_k a_k^2 = 1$. La première composante globale $t^{(1)}$ de l'$ACPVI$ multibloc est la synthèse des composantes partielles $t_1^{(1)}$, $t_2^{(1)}$ et $t_3^{(1)}$. La seconde composante globale $t^{(2)}$ est surtout déterminée par les composantes partielles $t_1^{(2)}$ et $t_2^{(2)}$, et la troisième composante globale $t^{(3)}$ par la composante partielle $t_2^{(3)}$.

TABLEAU 3. — Contribution relative des composantes partielles t_k dans la construction de la composante globale t pour les cinq premières dimensions.

Dimension	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
$\%a_1^{(h)2}$	**32,7**	**37.1**	17,5	**33,7**	29,5
$\%a_2^{(h)2}$	27,6	**43,3**	**42,2**	20,0	**49,4**
$\%a_3^{(h)2}$	**30,0**	13,6	21,6	14,1	8,6
$\%a_4^{(h)2}$	9,7	6,1	18,7	**32,2**	12,4
Total	100	100	100	100	100

3.2.2. Représentation factorielle de l'ensemble des variables

Les variables de X et de Y peuvent être représentées sur la base des composantes globales t, orthogonales par construction. La figure 2 illustre cette représentation pour les deux premières dimensions $t^{(1)}$ et $t^{(2)}$. Les variables Y relatives aux taux de truies ($CIRCOTR$) et de porcelets ($CIRCOPS$) séropositifs au $PCV2$ sont liées et non corrélées au taux de porcs séropositifs ($CIRCOPC$). En effet, les truies et les porcelets sont élevés ensemble, ce qui explique que leurs taux de séroconversion soient comparables. Les porcs à l'engrais sont élevés dans d'autres bâtiments et ont donc des modes de contamination différents. De plus, pour déterminer les facteurs de risque de la maladie, il est essentiel de raisonner sur l'ensemble des variables Y. En effet, un élevage ayant un profil à risque est un élevage où les truies et les porcelets séroconvertissent peu (faible immunité passive transmise de la truie aux porcelets) et où les porcs à l'engrais séroconvertissent fortement (forte immunité active liée à l'infection virale) (Rose *et al.*, 2003). Les variables ayant des coordonnées négativement corrélées à la composante $t^{(1)}$ sont donc associées au profil à risque pour la séroconversion au virus. La variable $FOSMAT$, relative à la profondeur des préfosses en maternité, par exemple, est interprétée comme un facteur de risque pour l'élevage. A l'inverse, les variables ayant des coordonnées positivement corrélées à la composante $t^{(1)}$ sont associées à un profil protecteur pour la séroconversion au virus. La variable $PEDILUV$, définissant l'utilisation d'un pédiluve dans chaque salle de l'élevage, est interprétée comme un facteur protecteur pour l'élevage.

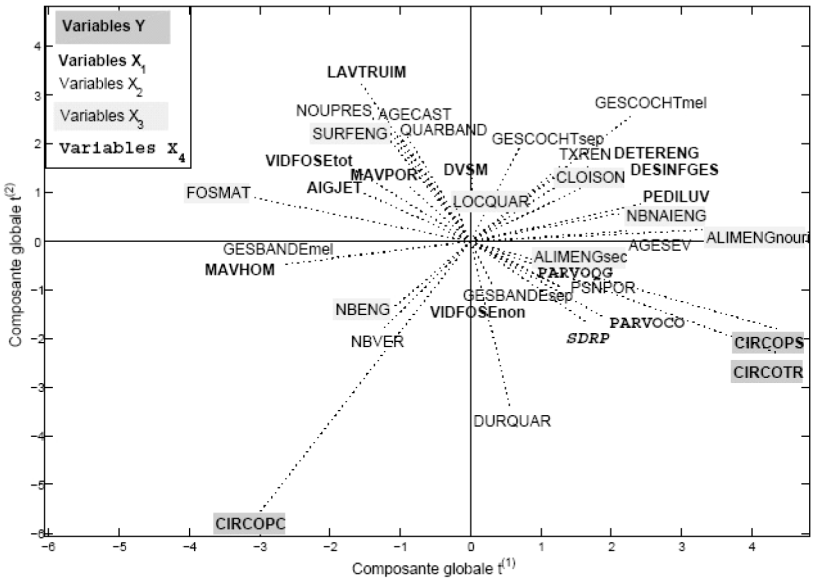

FIG 2. — Représentation factorielle des variables X et Y à partir de leurs coefficients pour les composantes globales $t^{(1)}$ et $t^{(2)}$.

3.3. Prédiction à partir de tableaux structurés en blocs

3.3.1. Nombre optimal de dimensions

Sur la base d'une étude de validation croisée, l'erreur moyenne de prédiction ($RMSE_V$) est calculée en fonction du nombre de composantes globales de l'$ACPVI$ multibloc, introduites dans le modèle de prédiction. A titre de comparaison, nous avons également calculé les valeurs de $RMSE_V$ correspondant à des modèles de régression sur la base de composantes obtenues par la méthode PLS multibloc. Ces résultats sont reportés dans la figure 3. Il ressort qu'une seule variable latente de l'$ACPVI$ est utile pour la prédiction des variables de Y. La performance de la méthode PLS multibloc, avec deux variables latentes, est légèrement supérieure à celle de l'$ACPVI$ multibloc.

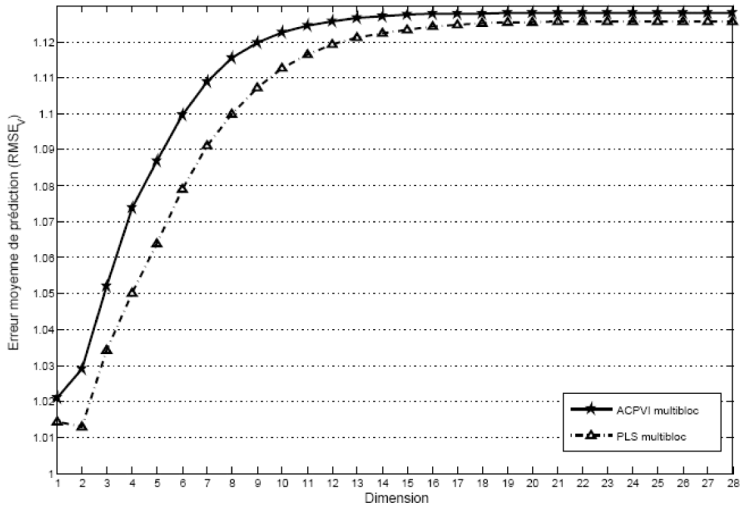

FIG 3. — Comparaison des erreurs moyennes de prédiction ($RMSE_V$) de l'$ACPVI$ multibloc et de la méthode PLS multibloc.

Par ailleurs, nous avons appliqué la méthode PLS hiérarchique multibloc au jeu de données considéré mais les résultats, non présentés ici, se sont avérés très decevants en comparaison de ceux obtenus par l'$ACPVI$ multibloc et la méthode PLS multibloc.

3.3.2. Poids des variables X dans l'explication des variables Y

Les coefficients de régression associés aux trois variables à prédire, donnés en annexe B, sont calculés à partir du nombre optimal de dimensions : une seule pour l'$ACPVI$ multibloc et deux pour la méthode PLS multibloc. Les résultats donnés par l'$ACPVI$ multibloc et la méthode PLS multibloc diffèrent légèrement. L'$ACPVI$ multibloc donne des coefficients de régression concordants (dans le sens où les facteurs de risque et les facteurs protecteurs sont les mêmes) pour $CIRCOPS$ et $CIRCOTR$ et opposés à ceux de $CIRCOPC$. L'objectif à atteindre serait un élevage où les porcelets ont

une bonne immunité passive (issue de leur mère) et une faible immunité active (pas de contact avec le virus $PCV2$), où les porcs charcutiers sont faiblement séropositifs au $PCV2$ (pas de contact avec le virus) et où les truies sont fortement séropositives afin de transmettre le plus d'anticorps à leurs porcelets. Ce qui revient à dire qu'un facteur de risque est ici un facteur diminuant l'immunité passive des porcelets et des truies ($\beta_{CIRCOPS} < 0$ et $\beta_{CIRCOTR} < 0$) et augmentant l'immunité active des porcs ($\beta_{CIRCOPC} > 0$).

4. Conclusion

L'$ACPVI$ multibloc permet à la fois d'explorer les relations entre K tableaux X_k ($k = 1, \ldots, K$) et un tableau Y, et de prédire Y à partir de ces blocs. Cette méthode est basée sur la détermination simultanée de composantes globales et partielles. Les composantes partielles sont les synthèses des tableaux X_k. Les composantes globales sont construites de façon à résumer les composantes partielles. De ce fait, elles apparaissent comme étant la synthèse du tableau concaténé X orienté vers l'explication du tableau Y. Elles permettent soit de décrire le lien (X, Y) soit de prédire Y à partir de X. Les composantes globales jouent un rôle central dans l'analyse car elles permettent la représentation de toutes les variables dans un espace commun. L'avantage de l'$ACPVI$ multibloc par rapport aux autres méthodes est d'avoir un critère à optimiser clair et résumant les objectifs de la méthode. L'$ACPVI$ multibloc permet ainsi de gérer la complexité des données d'épidémiologie animale en prenant en compte l'organisation de celles-ci en blocs et en fournissant des informations au niveau des variables et des blocs de variables.

Sur la base de l'étude de cas considérée, il ressort que l'$ACPVI$ multibloc et la méthode PLS multibloc donnent des résultats légèrement différents, avec des capacités prédictives équivalentes. Cependant, l'$ACPVI$ multibloc nécessite l'inversion de la matrice de variance-covariance associée à chaque bloc, ce qui peut poser des problèmes d'instabilité en présence de quasi-colinéarité au sein de chaque bloc de variables. Dans ce cas, la PLS multibloc peut constituer une alternative. Des développements concernant la comparaison de ces deux méthodes seront étudiés ultérieurement, en se basant notamment sur les travaux de Vivien (2002).

Il ressort de la comparaison de l'$ACPVI$ multibloc aux méthodes existantes que le choix des contraintes de normes sur les composantes ou sur les axes est important. Il conviendrait d'explorer dans un cadre plus général offert par l'approche PLS (Wold, 1982 ; Tenenhaus *et al.*, 2005b), le lien entre ces contraintes et les schémas d'estimation (mode A ou B, ainsi que les modèles centroïde, factoriel ou structurel).

A. Annexe : Description des variables et des blocs de variables

Bloc	Variable	Description
	CIRCOPS	Taux de porcelets positifs au circovirus $PCV2$ en post-sevrage
Y	CIRCOPC	Taux de porcs positifs au circovirus $PCV2$ en engraissement
	CIRCOTR	Taux de truies positives au circovirus $PCV2$
	MAVPOR	Sens de circulation des animaux (marche en avant)
	MAVHOM	Sens de circulation des hommes (marche en avant)
	PEDILUV	Utilisation d'un pédiluve dans chaque salle de l'élevage
	AIGJET	Utilisation d'une aiguille jetable par truie pour les vaccins
X_1	DETERENG	Détersion de la salle d'engraissement après lavage
	VIDFOSEnon	Vidange de la fosse (partielle *vs* pas de vidange)
	VIDFOSEtot	Vidange de la fosse (partielle *vs* totale)
	DESINFGES	Désinfection des stalles de gestation
	LAVTRUIM	Lavage des truies à l'entrée en maternité
	DVSM	Durée du vide sanitaire en maternité
	GESBANDEmel	Séparation physique des bandes de truies gestantes (externe *vs* mélange)
	GESBANDEsep	Séparation physique des bandes de truies gestantes (externe *vs* séparées)
	GESCOCHTmel	Position des cochettes au sein des travées (externe *vs* mélange)
	GESCOCHTsep	Position des cochettes au sein des travées (externe *vs* séparées)
	NOUPRES	Présence d'une nursery pour une partie de la bande
X_2	AGECAST	Age des verrats à la castration
	TXREN	Taux de renouvellement du troupeau de truies
	AGESEV	Age des porcelets au sevrage
	PSNPOR	Nombre de portées par case en post-sevrage
	QUARBAND	Nombre de lots par salle en quarantaine
	DURQUAR	Durée moyenne de la quarantaine (en semaines)
	NBVER	Nombre moyen de verrats introduits par an
	NBENG	Nombre d'élevages engraisseurs dans un rayon de 2 km autour de l'élevage
	NBNAIENG	Nombre d'élevages naisseurs-engraisseurs dans un rayon de 2 km autour de l'élevage
	CLOISON	Cloisons entre les préfosses en engraissement
X_3	ALIMENGnourri	Type d'alimentation en engraissement (soupe *vs* nourrisoupe)
	ALIMENGsec	Type d'alimentation en engraissement (soupe *vs* sec)
	SURFENG	Surface des cases en engraissement
	LOCQUAR	Type de locaux en quarantaine (semi-claustration ou claustration)
	FOSMAT	Profondeur des préfosses en maternité
	SDRP	Vaccination des truies contre le virus $SDRP$
X_4	PARVOQG	Utilisation du même antigène contre le parvovirus en quarantaine et lors de la gestation
	PARVOCO	Taux de cochettes positives au parvovirus

B. Annexe : Coefficients de régression de Y sur l'ensemble des variables X pour l'$ACPVI$ multibloc et la PLS multibloc

		$ACPVI$ multibloc (nbdim=1)			PLS multibloc (nbdim=2)		
Bloc	Variable	CIRCOPS	CIRCOPC	CIRCOTR	CIRCOPS	CIRCOPC	CIRCOTR
	MAVPOR	-0,03	0,02	-0,03	-0,07	-0,04	-0,08
	MAVHOM	**-0,11**	0,07	**-0,11**	-0,09	0,07	-0,09
	PEDILUV	**0,10**	-0,07	**0,10**	**0,11**	**-0,12**	0,09
	AIGJET	-0,06	0,04	-0,06	-0,06	-0,04	-0,08
X_1	DETERENG	0,07	-0,05	0,07	0,04	**-0,12**	0,02
	VIDFOSEnon	-0,03	0,02	-0,03	0,02	0,05	0,03
	VIDFOSEtot	-0,07	0,05	-0,07	-0,06	-0,05	-0,08
	DESINFGES	0,08	-0,06	0,08	0,04	**-0,17**	0,00
	LAVTRUIM	-0,06	0,04	-0,06	-0,08	**-0,12**	**-0,11**
	DVSM	0,00	0,00	0,00	-0,03	-0,04	-0,04
	GESBANDEmel	-0,08	0,05	-0,08	-0,09	0,01	**-0,10**
	GESBANDEsep	0,01	-0,01	0,01	0,01	0,01	0,02
	GESCOCHTmel	0,09	-0,06	0,09	0,08	**-0,11**	0,06
	GESCOCHTsep	0,03	-0,02	0,03	-0,03	0,00	-0,03
	NOUPRES	-0,06	0,04	-0,06	-0,06	-0,06	-0,09
X_2	AGECAST	-0,05	0,03	-0,05	-0,05	-0,07	-0,08
	TXREN	0,05	-0,04	0,05	0,04	-0,09	0,02
	AGESEV	0,09	-0,06	0,09	0,08	**-0,12**	0,06
	PSNPOR	0,05	-0,04	0,05	0,03	-0,03	0,03
	QUARBAND	-0,04	0,02	-0,04	-0,02	-0,08	-0,05
	DURQUAR	0,02	-0,02	0,02	0,05	0,08	0,08
	NBVER	-0,05	0,03	-0,05	-0,03	**0,14**	0,00
	NBENG	-0,04	0,03	-0,04	-0,05	0,07	-0,03
	NBNAIENG	0,08	-0,06	0,08	0,04	-0,05	0,03
	CLOISON	0,05	-0,04	0,05	0,04	**-0,10**	0,02
X_3	ALIMENGnourri	**0,13**	-0,09	**0,13**	0,09	**-0,12**	0,07
	ALIMENGsec	0,03	-0,02	0,03	0,00	0,00	0,00
	SURFENG	-0,04	0,03	-0,04	-0,01	-0,09	-0,04
	LOCQUAR	-0,01	0,01	-0,01	-0,03	-0,05	-0,05
	FOSMAT	**-0,13**	0,09	**-0,12**	**-0,14**	0,02	**-0,14**
	SDRP	0,07	-0,05	0,07	0,09	0,02	**0,11**
X_4	PARVOQG	0,04	-0,02	0,03	0,04	0,00	0,04
	PARVOCO	0,08	-0,05	0,07	0,09	0,00	**0,10**

Références

A.J. Burnham, R. Viveros et J.F. Macgregor (1996), Framework for latent variable multivariate regression. *Journal of chemometrics*, 10 :31-45.

J.D. Carroll (1968), A generalization of canonical correlation analysis to three or more sets of variables. *In 76th annual convention of the American psychological association*, pages 227-228.

P. Cazes (1975), Protection de la régression par utilisation de contraintes linéaires et non linéaires. *Revue de statistique appliquée*, XXIII(3) :37-57.

D. Chessel et P. Mercier (1993), Couplage de triplets statistiques et liaisons espèces environnement. *In* LEBRETON J.D. et ASSELAIN B., éditeur : *Biométrie et environnement*, pages 15-44. MASSON, Paris.

P.T. Davies et M. K.S. Tso (1982), Procedures for reduced-rank regression. *Appl. Statist.*, 31 :244-255.

A.E. Hoerl et R.W. Kennard (1970), Ridge regression : biaised estimation for nonorthogonal problems. *Technometrics*, 12 :55-67.

G. Kissita (2003), *Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués*. Thèse de doctorat, Université Paris Dauphine

G. Kissita, P. Cazes, M. Hanafi et R. Lafosse (2004), Deux méthodes danalyse factorielle du lien entre deux tableaux de variables partitionnés. *Revue de statistique appliquée*, LII(3) :73-92.

R. Lafosse et M. Hanafi (1997), Concordance dun tableau avec K tableaux : définition de K+1uples synthétiques. *Revue de statistique appliquée*, XLV (4) :111-126.

K.E. Muller (1981), Relationships between redundancy analysis, canonical correlation and multivariate regression. *Psychometrika*, 46(2) :139-142.

H. Nocairi, M. Hanafi et E.M. Qannari (2005), Approche continuum de la discrimination de type ridge. *Revue de statistique appliquée*, LIII(2) :29-41.

C.R. Rao (1964), The use and interpretation of principal component analysis in applied research. *Sankhya, A.*, 26 :329-358.

N. Rose, G. Larour, G. Le Digherher, E. Eveno, J.P. Jolly, P. Blanchard, A. Oger, M. Le Dinma, A. Jestin et F. Madec (2003), Risk factors for porcine post-weaning multisystemic wasting syndrome (PMWS) in 149 french farrow-to-finish herds. *Preventive Veterinary Medicine*, 61 :209-225.

R. Sabatier (1987a), Analyse factorielle de données structurées et métriques. *Statistique et analyse des données*, 12(3) :75-96.

R. Sabatier (1987b), *Méthodes factorielles en analyse de données : approximations et prise en compte de variables concomitantes*. Thèse de doctorat, Université des sciences et techniques du Languedoc.

G. Saporta (2006), *Probabilités, analyse des données et statistique (2nd édition)*. Technip, Paris.

M. Stone (1974), Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(1) :111-147.

M. Tenenhaus (1998), *La régression PLS. Théorie et pratique*. Technip, Paris.

M. Tenenhaus (1999), *L'approche PLS. Revue de statistique appliquée*, 47(2) :5-40.

M. Tenenhaus, J. Pages, L. Ambroisine et C. Guinot (2005a), PLS methodology to study relationships between hedonic judgements and product characteristics. *Food quality and preference*, 16 :315-325.

M. Tenenhaus, V.E. Vinzi, Y.M. Chatelin et C. Lauro (2005b), PLS path modeling. *Computational statistics and data analysis*, 48 :159-205.

J.P. Van De Geer (1984), Linear relations among K sets of variables. *Psychometrika*, 49 (1) :79-94.

A. Van Den Wollenberg (1977), Redundancy analysis : an alternative for canonical correlation analysis. *Psychometrika*, 42(2) :207-219.

M. Vivien (2002), *Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications*. Thèse de doctorat, Université de Montpellier 1.

J.A. Westerhuis et P.M.J. Coenegracht (1997), Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of chemometrics*, 11(5) :379-392.

J.A. Westerhuis, T. Kourti et J.F. Macgregor (1998), Analysis of multiblock and hierarchical PCA and PLS model. *Journal of chemometrics*, 12 :301-321.

H. Wold (1982), Soft modelling : the basic design and some extensions. In K.G JÖRESKOG et H. WOLD, éditeurs : *System under indirect observation*. Part 2, pages 1-54. North-Holland, Amsterdam.

S. Wold (1984), Three PLS algorithms according to SW. In S. WOLD, éditeur : *Symposium MULDAST (multivariate analysis in science and technology)*, pages 26-30, Umea University, Sweden.

S. Wold, N. Kettaneh et K. Tjessem (1996), Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics*, 10 :463-482.

S. Wold, H. Martens et H. Wold (1983), The multivariate calibration problem in chemistry solved by the PLS method. In Ruhe A. B. et KASTROM, éditeurs : *Proceedings of the Conference on Matrix Pencils*, pages 286-293. Springer Verlag, Heidelberg.