REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES

Adaptation de la régression PLS au cas de la régression après analyse des correspondances multiples

Revue de statistique appliquée, tome 45, n° 2 (1997), p. 89-99

<http://www.numdam.org/item?id=RSA_1997__45_2_89_0>

Rev. Statistique Appliquée, 1997, **XLV** (2), 89-99

ADAPTATION DE LA RÉGRESSION PLS AU CAS DE LA RÉGRESSION APRÈS ANALYSE DES CORRESPONDANCES MULTIPLES

P. Cazes

LISE-CEREMADE, Université Paris 9 Dauphine,
Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16

RÉSUMÉ

Après avoir effectué des rappels et donné des compléments méthodologiques sur la Régression PLS et sur la Régression après Analyse des Correspondances Multiples, on adapte la première technique à la seconde, ce qui permet d'obtenir des composantes principales non corrélées sur lesquelles on peut effectuer la régression.

Mots-clés : *Analyse en Composantes Principales, Analyse des Correspondances Multiples, Régression PLS, Régression après Analyse des Correspondances Multiples.*

ABSTRACT

After giving principles and properties of PLS Regression and Regression after Multiple Correspondence Analysis, the first technics is matched to the second. Thus, we obtain uncorrelated principal components, on which regression can be performed.

Keywords : *Principal Components Analysis, Multiple Correspondence Analysis, PLS Regression, Regression after Multiple Correspondence Analysis.*

1. Introduction

Le but de ce papier est d'adapter la régression PLS au cas de la régression après analyse des correspondances multiples (*cf.* [1]). En effet le premier pas de cette méthode qui fournit des composantes principales corrélées sur lesquelles on effectue la régression peut être considéré comme le premier pas d'une régression PLS (où l'Analyse en Composantes Principales est remplacée par l'Analyse des Correspondances Multiples (A.C.M.)). L'intérêt de cette adaptation est de fournir des composantes principales non corrélées sur lesquelles on pourra effectuer la régression.

L'article est divisé en trois parties. Dans la première, on fait des rappels, dans un cadre plus général (*i.e.* avec des métriques quelconques) sur la régression PLS, et on donne des compléments méthodologiques. Dans la deuxième partie, on rappelle les principes de la régression après A.C.M., on donne quelques propriétés

de cette technique, et on fournit également des compléments méthodologiques. Enfin la dernière partie concerne l'adaptation de la régression PLS au cas de la régression après A.C.M.

2. Rappels et compléments sur la régression P.L.S. (*cf.* [3], [5], [6])

2.1. Rappels

On suppose qu'on a $p+q$ variables quantitatives $x_1, x_2, \ldots, x_p, y_1, y_2, \ldots, y_q$ mesurées sur un échantillon de taille n et l'on désigne par X et Y les tableaux $n \times p$ et $n \times q$ associés. On désire expliquer les variables $y_1, \ldots, y_q$ en fonction des variables $x_1, x_2, \ldots, x_p$. Le but de la régression PLS est de fournir de nouvelles variables $\psi_1, \psi_2, \ldots, \psi_r$ $(r \leqslant p)$ combinaisons linéaires des variables explicatives x_i, non corrélées, fonction de la liaison entre les y_j et les x_i, variables sur lesquelles on effectuera la régression de chaque y_j.

On suppose l'espace R^n muni de la métrique des poids $N = D_p$ (en général $\frac{1}{n} Id_n$, Id_n étant la métrique identité, ce qui revient à donner même poids à chaque observation) et les espaces R^p et R^q des métriques définies respectivement par les matrices M_{11} et M_{22}. On supposera les variables centrées, et l'on pose[1] :

$V_{11} = X'NX$: matrice variance des variables explicatives
$V_{12} = X'NY = (V_{21})'$: matrice des covariances entre les variables x_i et les variables y_j

On posera encore :

$$E_0 = X,\ F_0 = Y,\ V_{11}^{(0)} = V_{11},\ V_{12}^{(0)} = \left(V_{21}^{(0)}\right)' = V_{12} \tag{1}$$

notations qui seront utiles dans le processus itératif de la régresion PLS.

Au premier pas de cette régression, on recherche les combinaisons linéaires $\underline{\psi} = X\,\underline{a}$, $\underline{\eta} = Y\,\underline{c}$ de covariance maximale, les formes linéaires $\underline{a} \in (R^p)^*$ et $\underline{c} \in (R^q)^*$ étant normées pour les métriques M_{11}^{-1} et M_{22}^{-1} respectivement induites par M_{11} et M_{22} dans $(R^p)^*$ et $(R^q)^*$:

$$\underline{a}'M_{11}^{-1}\underline{a} = \underline{c}'M_{22}^{-1}\,\underline{c} = 1\ . \tag{2}$$

La covariance à maximiser étant égale à $\underline{\psi}'N\,\underline{\eta} = \underline{a}'\,V_{12}\,\underline{c}$, il est facile de voir (*cf.* [3]) que $\underline{a}$ et $\underline{c}$ doivent vérifier les équations

$$\left.\begin{array}{rcl} M_{11}V_{12}\underline{c} & = & \lambda\,\underline{a} \\ M_{22}V_{21}\underline{a} & = & \lambda\,\underline{c} \end{array}\right\} \tag{3}$$

[1] On désigne de façon classique par A' la transposée d'une matrice A

avec

$$\text{Cov}\,(\psi, \eta) = \underline{a}'\, V_{12}\, \underline{c} = \lambda\,. \tag{4}$$

On déduit alors de (3) et (4) que $\underline{a}$ (resp. $\underline{c}$) est le vecteur propre $\underline{a}_1$ (resp. $\underline{c}_1$) normé (pour M_{11}^{-1} (resp. M_{22}^{-1})) de $M_{11}\ V_{12}\ M_{22}\ V_{21}$ (resp. $M_{22}\ V_{21}\ M_{11}\ V_{12}$) associé à la plus grande valeur propre de cette application.

On pose alors :

$$\underline{\psi}_1 = E_0\, \underline{a}_1 \ ;\ \underline{\eta}_1 = F_0\, \underline{c}_1 \tag{5}$$

et on désigne par A_1 l'opérateur de projection dans R^n (toujours muni de la métrique N) sur l'hyperplan orthogonal à $\underline{\psi}_1$ et par $E_1 = A_1 E_0$ (resp. $F_1 = A_1 F_0$) les matrices résiduelles de E_0 (resp. F_0) après régression sur $\underline{\psi}_1$.

On pose encore :

$$V_{12}^{(1)} = E_1' N F_1 \ ;\ V_{11}^{(1)} = E_1' N E_1 \tag{6}$$

et il est facile de voir que :

$$V_{12}^{(1)} = E_1' N F_0 \ ;\ V_{11}^{(1)} = E_1' N E_0 \tag{7}$$

Au second pas de l'algorithme, on recherche les combinaisons linéaires $\underline{\psi} = E_1\, \underline{a}$ et $\underline{\eta} = F_1\, \underline{c}$ de covariance maximale sous les contraintes de normalisation (2).

Les solutions $\underline{a}_2$ et $\underline{c}_2$ sont donc respectivement les vecteurs propres associés à la plus grande valeur propre de $M_{11} V_{12}^{(1)} M_{22} V_{21}^{(1)}$ et $M_{22} V_{21}^{(1)} M_{11} V_{12}^{(1)}$.

On pose alors $\underline{\psi}_2 = E_1\, \underline{a}_2$, $\underline{\eta}_2 = F_1\, \underline{c}_2$, et on considère les matrices résiduelles E_2 et F_2 après régression de E_1 et F_1 sur $\underline{\psi}_2$. On peut alors continuer le processus en recherchant $\underline{a}$ et $\underline{c}$ tels que Cov $(E_2\, \underline{a},\ F_2\, \underline{c})$ soit maximale sous les contraintes (2).

De façon générale, au pas k, on recherchera le couple de vecteurs normés $(\underline{a}, \underline{c})$ tel que Cov $(E_{k-1}\, \underline{a},\ F_{k-1}\, \underline{c})$ soit maximum.

2.2. Remarques et compléments

1) La régression PLS classique correspond au cas où les métriques M_{11} et M_{22} sont les métriques usuelles de R^p et R^q respectivement.

2) Par construction les composantes PLS $\underline{\psi}_k$ sont non corrélées, et l'on peut montrer (*cf.* [3]) que les $\underline{a}_k$ sont orthonormés (pour M_{11}^{-1}).

3) On déduit de (7) que les résultats de la régression PLS sont inaltérés si à chaque pas de l'algorithme, on ne corrige que le tableau $X = E_0$ et pas $Y = F_0$. Au pas k, on recherche alors les combinaisons linéaires $\underline{\psi} = E_{k-1}\, \underline{a}$ et $\underline{\theta} = F_0\, \underline{c}$ de covariance maximale sous les contraintes (2).

4) Après avoir extrait les composantes PLS $\underline{\psi}_k$, composantes qui par construction dépendent de la liaison entre les variables à expliquer y_j et les variables

explicatives x_i, on peut effectuer la régression de chaque y_j sur les ψ_k. Au lieu de considérer les ψ_k, on aurait pu aussi faire la régression sur les composantes principales $\underline{G}_k$ (qui comme les $\underline{\psi}_k$ sont non corrélées) issues de l'ACP du triplet (X, M_{11}, D_p).

L'avantage de raisonner sur les composantes PLS et non sur les composantes principales pour effectuer la régression est d'avoir, comme on vient de le dire, des composantes explicatives dépendant de la liaison entre les y_j et les x_i, ce qui permet d'optimiser dans un certain sens chaque régression, si on ne garde que les premières composantes PLS, contrairement au cas de la régression sur les composantes principales, où certaines composantes G_k de faible variance peuvent être éliminées alors qu'elles peuvent être corrélées de façon non négligeable à certaines des variables à expliquer y_j. Par contre les résultats obtenus avec la régression PLS sont biaisés, puisqu'on utilise pour expliquer les y_j des composantes qui sont déjà fonction des liaisons entre les y_j et les variables à expliquer x_i. On aura donc intérêt à utiliser un échantillon d'épreuve pour juger de la validité de la régression obtenue avec les $\underline{\psi}_k$.

5) A chaque pas k de la régression PLS, on extrait le premier facteur issu de l'ACP (non centrée) du triplet $\left(V_{12}^{(k-1)}, M_{11}, M_{22}\right)$ tandis que les composantes PLS $\underline{\psi}_k$ et $\underline{\eta}_k$ peuvent être considérées comme calculées à partir des tableaux E'_{k-1} et F_{k-1} rajoutés respectivement en colonnes et en lignes supplémentaires à $V_{12}^{(k-1)}$ (on projette chaque ligne de E_{k-1} (resp. F_{k-1}) sur le premier axe factoriel dans R^p (resp. R^q) issu de l'ACP de $V_{12}^{(k-1)}$).

Au lieu de corriger à chaque pas les tableaux $E_0 = X$ et $F_0 = Y$, on aurait pu faire l'ACP (non centrée) du triplet (V_{12}, M_{11}, M_{22}) les tableaux X et Y étant rajoutés en supplémentaires. Les composantes $\underline{\psi}_k = X\,\underline{a}_k$ et $\underline{\eta}_k = Y\,\underline{c}_k$ ainsi obtenues maximisent successivement Cov $(\underline{\psi}_k, \underline{\eta}_k)$ sous la contrainte que les $\underline{a}_k$ (resp. $\underline{c}_k$) sont normés et orthogonaux (pour la métrique M_{11}^{-1} (resp. M_{22}^{-1})) aux $\underline{a}_m$ (resp. $\underline{c}_m$) pour $m \leqslant k-1$.

On obtient alors l'analyse de co-inertie de Chessel et Mercier (*cf.* [4]) entre les triplets (X, M_{11}, N) et (Y, M_{22}, N), analyse qui généralise au cas de métriques quelconques l'analyse interbatteries de Tucker (*cf.* [7]). Le nombre maximum de composantes que l'on peut extraire est égal au rang de V_{12} et donc inférieur ou égal à $\mathrm{Min}(\mathrm{rang}X,\ \mathrm{rang}Y)$ alors qu'en régression PLS, il est égal au rang de X.

Les composantes $\underline{\psi}_k$ obtenues en analyse de co-inertie et qui peuvent servir de variables explicatives (au lieu des $\underline{x}_i$) pour expliquer les y_j sont, comme dans le cas de la régression PLS, centrées, et tiennent compte de la liaison entre les y_j et les x_i. Par contre, elles ont le désavantage de ne plus être non corrélées.

C'est cette façon d'opérer (*i.e.* sans corriger les tableaux X et Y) qui est utilisée dans la régression après analyse des correspondances multiples, l'ACP étant remplacée par l'Analyse Factorielle des Correspondances (AFC), le tableau Y (resp. X) par le tableau disjonctif complet associé aux indicatrices de la ou des variables à expliquer (resp. explicatives). Si on veut obtenir, pour faire la régression, des composantes explicatives non corrélées, il faut adapter la

procédure PLS au cas de l'analyse des correspondances multiples. C'est l'objet après des rappels sur la régression après analyse des correspondances multiples (§ 3) du paragraphe 4.

3. La Régression après Analyse des Correspondances Multiples (*cf.* [1], [2] et [3])

3.1. Rappels et Compléments

On suppose ici que toutes les variables $x_1, \ldots, x_i, \ldots, x_p, y_1, \ldots, y_j, \ldots, y_q$ ont été découpées en tranches, et l'on désigne par K_{x_i} (resp. K_{y_j}) l'ensemble des modalités de la variable x_i $(1 \leqslant i \leqslant p)$ (resp. y_j $(1 \leqslant j \leqslant q)$) et par K_X (resp. K_Y) l'union disjointe des K_{x_i} (resp. K_{y_j}), *i.e.* l'ensemble de toutes les modalités explicatives (resp. à expliquer) :

$$K_X = U\{K_{x_i} \mid i = 1, p\} \quad ; \quad K_Y = U\left\{K_{y_j} \mid j = 1, q\right\}$$

Si E désigne l'ensemble des n observations, on considère alors le tableau disjonctif complet S_{EK_X} qu'on notera également S, associé aux variables x_i et dont le terme général $S(e,k)$ est défini par : $\forall e \in E, \ \forall k \in K_{x_i} \subset K_X$:

$$\begin{aligned} S(e,k) &= 1 \text{ si } e \text{ a adopté la modalité } k \text{ de } x_i \\ &= 0 \text{ sinon} \end{aligned}$$

On note de même T_{EK_Y}, ou plus simplement T, le tableau disjonctif complet associé aux variables y_j, et l'on désignera par $T(e,k)(e \in E, \ k \in K_Y)$, le terme général de T.

La régression après analyse des correspondances multiples revient à effectuer les étapes suivantes :

1) Après la division en tranches des variables x_i et y_j, construire le tableau

$$C = T'S$$

qui rassemble l'ensemble des pq tableaux de contingence croisant toute variable x_i $(1 \leqslant i \leqslant p)$ avec toute variable y_j $(1 \leqslant j \leqslant q)$.

2) Effectuer l'analyse des correspondances du tableau C. On désignera par $\left(\varphi_\alpha^{K_X}, \varphi_\alpha^{K_Y}\right)$ le $\alpha^{ème}$ couple de facteurs associés de variance 1 issu de cette analyse et par λ_α la valeur propre correspondante.

3) Rajouter le tableau S en supplémentaire de C, *i.e.* projeter sur les r premiers axes factoriels trouvés en 2) les profils des lignes e du tableau S. Soit $F_\alpha(e)$[2] la coordonnée du profil de la ligne e de E sur l'axe factoriel α. Compte tenu de

[2] On emploie ici la notation F_α au lieu de ψ_α, cette notation étant plus classique en analyse des correspondances.

ce que $\sum\{S(e,k) \mid k \in K_X\} = p$, on a :

$$F_\alpha(e) = \frac{1}{p}\sum\{\varphi_\alpha^k S(e,k) \mid k \in K_X\} \tag{8}$$

4) Effectuer la régression usuelle de chaque y_j (avant découpage en classes) sur les F_α.

Soit

$$y_j^*(e) = \sum\{g_j^\alpha F_\alpha(e) \mid \alpha = 1, r\} \tag{9}$$

la valeur approchée de y_j pour l'individu e.

Compte tenu de (8), on a :

$$y_j^*(e) = \sum\{b_j^k S(e,k) \mid k \in K_X\} \tag{10}$$

avec

$$b_j^k = \frac{1}{p}\sum\{g_j^\alpha \varphi_\alpha^k \mid \alpha = 1, r\} \tag{11}$$

soit si $i(e)$ désigne la tranche de x_i dans laquelle tombe e :

$$y_j^*(e) = \sum\left\{b_j^{i(e)} \mid i = 1, p\right\} \tag{12}$$

qui n'est rien d'autre qu'une formule d'analyse de variance.

3.2. Quelques propriétés (cf [2], [3])

Les composantes principales F_α sont centrées; par contre elles ne sont pas non corrélées.

Par ailleurs, soit[3]

$$G_\alpha(e) = \frac{1}{q}\sum\{\varphi_\alpha^k T(e,k) \mid k \in K_Y\} \tag{13}$$

la coordonnée sur l'axe factoriel α du profil de la colonne e du tableau T' rajouté en supplémentaire à C. Alors on a

$$\begin{array}{lll} \text{Cov}\,(F_\alpha, G_\beta) & = \sqrt{\lambda_a} & \text{si} \quad \alpha = \beta \\ & = 0 & \text{si} \quad \alpha \neq \beta \end{array} \tag{14}$$

[3] On emploie ici la notation G_α (au lieu de η_α), cette notation étant plus classique en analyse des correspondances.

De plus, si φ^{K_X} et φ^{K_Y} sont des fonctions centrées de variance 1 (avec les métriques des poids déduites de l'analyse des correspondances de C), et si on pose

$$F(e) = \frac{1}{p}\sum\{\varphi^k S(e,k) \mid k \in K_X\} \ ; \ G(e) = \frac{1}{q}\sum\{\varphi^k T(e,k) \mid k \in K_Y\} \tag{15}$$

Alors F et G sont de covariance maximale si $F = F_1, G = G_1$, cette covariance étant d'après (14) égale à $\sqrt{\lambda_1}$.

De plus si on impose à φ^{K_X} (resp. φ^{K_Y}) d'être non corrélé aux $\varphi_\beta^{K_X}$ $(1 \leqslant \beta \leqslant \alpha - 1)$ (resp. $\varphi_\beta^{K_Y}$ $(1 \leqslant \beta \leqslant \alpha - 1)$), le couple (F, G) de covariance maximale est le couple (F_α, G_α).

3.3. Remarques

1) Comme l'ont remarqué Chessel et Mercier (*cf.* [4]), l'analyse des correspondances du tableau C revient à l'analyse de co-inertie des deux triplets suivants $\left(SD_{K_X}^{-1} - 1_{ns}, \frac{1}{p}D_{K_X}, \frac{1}{n}Id_n\right)$ et $\left(TD_{K_Y}^{-1} - 1_{nt}, \frac{1}{q}D_{K_Y}, \frac{1}{n}Id_n\right)$, D_{K_X} (resp. D_{K_Y}) étant la matrice diagonale des proportions des différentes modalités de K_X (resp. K_Y), 1_{ns} (resp. 1_{nt}) la matrice $n \times s$ (resp. $n \times t$) ne comportant que des 1, s (resp. t) le nombre de colonnes de S (resp. T) (*i.e.* le cardinal de K_X (resp. K_Y)), et Id_n la matrice identité d'ordre n.

   L'analyse des correspondances du tableau C est aussi équivalente à l'analyse de co-inertie des deux triplets suivants : $\left(\frac{1}{p}S, \left(\frac{1}{p}D_{K_X}\right)^{-1}, \frac{1}{n}Id_n\right)$ et $\left(\frac{1}{q}T, \left(\frac{1}{q}D_{K_Y}\right)^{-1}, \frac{1}{n}Id_n\right)$, cette façon de voir étant plus en accord avec l'esprit de cet article (*cf.* fin de la remarque 5) du § 2.2, *cf.* § 3.2, ainsi que les calculs qui seront développés au § 4.

2) Les composantes principales F_α étant corrélées, si on élimine une composante principale, parce qu'elle est peu liée aux y_j, ou qu'elle est associée à une valeur propre faible, il faut recalculer les coefficients de régression g_j^α dans la formule (9). La formule (11) n'est donc pas additive dans la mesure où la suppression d'une composante principale F_β ne revient pas simplement à supprimer le terme $g_j^\beta \varphi_\beta^k$ associé, mais oblige à recalculer les autres termes de cette formule.

3) Au lieu de faire la régression sur les F_α, on aurait pu faire la régression sur les composantes principales issues de l'analyse des correspondances du tableau disjonctif complet S (en mettant le tableau C en supplémentaire pour visualiser les liaisons entre les x_i et les y_j) qui ont l'avantage d'être non corrélées.

   Les formules (8) à (12) restent encore valables ($\varphi_\alpha^{K_X}$ étant maintenant le $\alpha^{ème}$ facteur de variance 1, et F_α la composante principale associée, *i.e.* le facteur sur E de variance la valeur propre λ_α, dans l'analyse des correspondances de S), les formules (9) et (11) étant additives dans la mesure où la suppression (ou l'ajoût) d'une composante principale revient simplement à supprimer (ou rajouter) le terme correspondant dans ces formules.

L'avantage de raisonner sur les composantes principales issues de l'analyse des correspondances du tableau C est d'avoir des composantes dépendant de la liaison entre les y_j et les x_i, ce qui permet d'optimiser dans un certain sens les régressions. Par contre les résultats de ces régressions sont biaisées puisqu'on prend pour variables explicatives des variables dépendant des liaisons entre les y_j et les x_i. On aura donc intérêt à utiliser un échantillon test pour valider les résultats obtenus. On retrouve ainsi des commentaires analogues à ceux déjà effectués dans le cas de la régression PLS (*cf.* remarque 4) du § 2.2).

4) L'intérêt de découper les variables en tranches puis de faire la régression sur les composantes principales issues de l'analyse des correspondances du tableau C avec S en supplémentaire (ou de S avec C en supplémentaire) réside dans les points suivants

   a) du fait du découpage en classes, on prend en compte les liaisons non linéaires entre les y_j et les x_i.

   b) On visualise ces liaisons par la représentation simultanée de K_X et K_Y fournie par l'analyse des correspondances de C (ou de S avec C en supplémentaire).

   c) On peut traiter sans problème le cas où les variables explicatives sont qualitatives (elles sont naturellement définies par des classes) ou le cas d'un mélange de variables qualitatives et quantitatives (par découpage en tranches de ces dernières).

   d) La méthodologie précédente se transpose aisément au cas où certaines variables y_j sont qualitatives. Il suffit, pour ces variables, d'effectuer l'analyse factorielle discriminante et la discrimination sur les F_α, au lieu de faire la régression.

   e) Comme dans toute méthode de régression après analyse factorielle, la méthodologie précédente permet de s'affranchir des problèmes de liaisons entre les variables explicatives, et d'éliminer le bruit dû aux fluctuations d'échantillonnage.

   f) Dans l'analyse des correspondances de C avec S en supplémentaire (ou l'inverse) le profil de chaque ligne k ($k \in K_{y_j} \subset K_Y$) est le centre de gravité des profils des lignes e de S associées aux individus e tombant dans la tranche k de y_j. L'analyse des correspondances de C correspond donc à une analyse interclasses. Il s'agit en fait d'une analyse factorielle discriminante où l'on a remplacé la métrique de Mahalanobis par la métrique du chi-deux.

   g) Un exemple de régression après analyse des correspondances multiples est donné dans CAZES (1976) (*cf.* [1]).

   h) En général la régression après analyse des correspondances multiples est utilisée quand on a une seule variable à expliquer (*i.e.* $q = 1$).

4. Régression PLS après Analyse des Correspondances Multiples

4.1. Le principe

On a vu (*cf.* § 3.2; *cf.* également la 1ère remarque du § 3.3) que quand on effectue l'analyse des correspondances du tableau C, le couple de premiers facteurs normés associés $\left(\varphi_1^{K_X}, \varphi_1^{K_Y}\right)$ est le couple de fonctions $\left(\varphi^{K_X}, \varphi^{K_Y}\right)$ normés qui maximise la covariance entre les fonctions F et G définies par (15), la valeur maximale de la covariance étant égale à $\sqrt{\lambda_1}$.

On est donc dans le cadre du premier pas de la régression PLS, à ceci près qu'il faut remplacer les tableaux X et Y par les tableaux S/p et T/q respectivement (l'introduction des facteurs $1/p$ et $1/q$ dans les tableaux disjonctifs complets vient du fait qu'on raisonne sur des profils), les métriques M_{11} et M_{22} étant les métriques du chi-2 induites par l'analyse des correspondances du tableau C. L'espace R^n étant muni de la métrique des poids $D_p = \frac{1}{n} Id_n$ induite par l'analyse des correspondances de S (ou de T) le tableau V_{12} considéré au début du § 2 n'est autre que le tableau $\frac{1}{npq} C'$ dont l'analyse des correspondances est identique à celle de C.

On peut donc, au lieu d'extraire les facteurs successifs issus de l'analyse des correspondances de C, qui produisent des facteurs F_α centrés, mais corrélés, opérer comme dans la régression PLS en corrigeant à chaque pas le tableau disjonctif complet des variables explicatives S.

Les résultats de la régression PLS étant inchangés si on ne corrige pas la matrice des variables à expliquer, ici le tableau disjonctif complet T (*cf.* 3ème remarque du § 2.2), nous ne modifierons pas ici par souci de simplification le tableau T.

Soit $S^*(e,k)/p \quad (k \in K_X)$ la composante k de la projection de la ligne e du tableau S/p sur la droite engendrée par $\underline{F_1}$ dans R^n. On a :

$$S^*(e,k) = u(k) F_1(e) \tag{16}$$

avec

$$u(k) = \sum \{S(e,k) F_1(e) \mid e \in E\} / \underline{F}_1' \underline{F}_1 \ . \tag{17}$$

On peut noter puisque F_1 est centré que :

$$\sum \{u(k) \mid k \in K_X\} = p \sum \{F_1(e) \mid e \in E\} / \underline{F}_1' \underline{F}_1 = 0 \ . \tag{18}$$

Les marges du tableau S^* sont donc nulles.

On pose alors :

$$\forall e \in E, \ \forall k \in K_X, \quad S^{(1)}(e,k) = S(e,k) - S^*(e,k) = S(e,k) - u(k) F_1(e) \tag{19}$$

et on peut noter que les tableaux S et $S^{(1)}$ ont mêmes marges.

On pose de même :

$$C^{(1)} = T' S^{(1)} \tag{20}$$

dont le terme général s'écrit : $\forall k \in K_Y,\ \forall k' \in K_X$:

$$\begin{aligned} C^{(1)}(k,k') &= \sum \{T(e,k) S^{(1)}(e,k') \mid e \in E\} \\ &= \sum \{T(e,k)(S(e,k') - u(k')F_1(e)) \mid e \in E\} \\ &= C(k,k') - u(k') \sum \{T(e,k) F_1(e) \mid e \in E\} \end{aligned}$$

Il est facile de voir compte tenu de (18) et de ce que $\sum \{T(e,k) \mid k \in K_Y\} = q$, que les tableaux C et $C^{(1)}$ ont même marge. Il en résulte que les métriques induites par l'analyse des correspondances du tableau $C^{(1)}$ sont identiques à celles induites par l'analyse des correspondances de C.

Remplaçant le tableau S par $S^{(1)}$, on peut alors rechercher le couple de fonctions $\left(\varphi^{K_X}, \varphi^{K_Y}\right)$ normées (avec les mêmes métriques que précédemment) qui maximise la covariance entre les fonctions F et G définies par (15) (où S est remplacé par $S^{(1)}$). La solution $\left(\varphi_2^{K_X}, \varphi_2^{K_Y}\right)$ est fournie par le premier couple de facteurs normés associés issu de l'analyse des correspondances du tableau $C^{(1)}$. Si λ_2 est la valeur propre la plus grande issue de cette analyse, la covariance maximale est égale à $\sqrt{\lambda_2}$. Si F_2 et G_2 sont les fonctions sur E associées, il est immédiat de vérifier que F_2 est non corrélée à F_1.

On peut ensuite continuer le processus qui revient donc à corriger les tableaux $S^{(1)}$ puis $C^{(1)}$ pour obtenir des tableaux $S^{(2)}$ et $C^{(2)}$, extraire le premier couple de facteurs normés associés issu de l'analyse des correspondances de $C^{(2)}$, en déduire la combinaison linéaire associée F_3, corriger en conséquence $S^{(2)}$ et $C^{(2)}$ etc.

On voit donc que cette façon d'opérer revient à faire une suite d'analyse des correspondances. Si on effectue r analyses, on obtient pour expliquer les y_j (avant découpage en classes) r composantes F_α centrées, non corrélées.

4.2. Remarques

1) On peut noter que les tableaux $C^{(1)}, C^{(2)}, \ldots, C^{(r)}$ peuvent comporter des valeurs négatives, mais comme ils ont des marges positives, on peut sans difficulté en faire l'analyse des correspondances. De plus, toutes les valeurs propres issues de l'analyse des correspondances de $C^{(\alpha)}$ sont inférieures ou égales à 1 (donc comprises entre 0 et 1). En effet, la plus grande valeur propre $\lambda_{\alpha+1}$ issue de l'analyse de $C^{(\alpha)}$ est égale au carré de la covariance optimisée au pas $\alpha + 1$, qui est donc inférieure à λ_1 carré de la covariance optimisée sans contraintes (à part la contrainte de norme) au premier pas, λ_1 étant inférieure ou égale à 1, comme plus grande valeur propre issue de l'analyse des correspondances du tableau $C^{(0)} = C$ qui ne comporte pas d'éléments négatifs.

2) Pour les mêmes raisons que pour la régression PLS (*cf.* fin de la remarque 4) du § 2.2) et que pour la régression après ACM (*cf.* fin de la remarque 3) du § 3.3) les résultats fournis par la régression PLS après ACM sont biaisés.

3) La méthodologie précédente s'applique bien sûr sans problème si on a une seule variable à expliquer soit si $q = 1$.

4) La méthode de régression PLS développée ci-dessus diffère de la régression PLS qualitative préconisée par Tenenhaus *et al.* (1995) (*cf.* [6]) : dans cette dernière méthode, on effectue la régression PLS classique sur les composantes principales respectivement issues de l'analyse des correspondances des tableaux disjonctifs complets S et T, puis on se sert des formules de reconstitution pour reconstituer les y_j (*i.e.* les $T(e,k),\ k \in K_Y$) en fonction des x_i (*i.e.* des $S(e,k),\ k \in K_X$)).

Remerciements

L'auteur remercie vivement M. Tenenhaus pour avoir relu l'article et effectué des suggestions pour l'améliorer.

Bibliographie

[1] CAZES P. (1976). Régression par boule et par l'analyse des correspondances, R.S.A., Vol. 24 n° 4, pp. 5-22.

[2] CAZES P. (1980). L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives, Les Cahiers de l'Analyse des Données, Vol. V n° 2, pp. 145-161.

[3] CAZES P. (1996). Méthodes de Régression, Polycopié de 3ème cycle, Université Paris Dauphine.

[4] CHESSEL D., MERCIER P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement, In : *Biométrie et Environnement*, LEBRETON J.D., ASSELAIN B. (Eds), Masson, Paris, pp. 15-44.

[5] TENENHAUS M., GAUCHI J.P., MENARDO C. (1995). Régression PLS et applications, RSA, Vol. 43, n° 1, pp. 7-63.

[6] TENENHAUS M. (1995). A partial least squares approach to multiple regression, redundancy analysis, and canonical analysis, Les Cahiers de Recherche de HEC, CR 550/1995.

[7] TUCKER L.R. (1958). An inter-battery method of factor analysis, *Psychometrica*, Vol. 23 n° 2, pp. 111-136.