# La régression multivariée contrainte PLS

Philippe Casin, Francois Marque

# La régression multivariée contrainte PLS

**Philippe CASIN et François MARQUE**

Laboratoire ID2

UFR Mathématique, Informatique, Mécanique

Université de METZ

Ile du Saulcy, 57005 METZ Cedex

**Résumé :** L'objet de cette communication est de montrer les problèmes de manque de robustesse des solutions de la régression multivariée contrainte, obtenues grâce à une analyse canonique entre deux ensembles de résidus de régressions. Une autre méthode, proche des techniques PLS, est alors proposée.

**Mots-clés :** Régression multivariée contrainte, analyse canonique, méthodes PLS

**Abstract :** This paper aims to show lack of robustness that can occurs for solutions of Multivariate Reduced Rank Regression because these solutions are computed by a canonical analysis between two sets of residuals of regression ; another method, close to PLS methods, is proposed afterwards

**Key-words :** Multivariate reduced rank regression, canonical analysis, PLS methods

## ECONOMETRIE – ANALYSE DES DONNEES

# 1. Introduction

Présentée pour la première fois par Anderson (1951), la régression multivariée contrainte (RMC) est une technique qui présente aujourd'hui de nombreuses applications en économie et en finance (par exemple, Perron et Campbell (1992), Gouriéroux, Monfort, Renault (1993), notamment pour les applications en finance et en économétrie des données de panel ; d'autre part, Anderson (2002), Johansen (1988, 1995, et 2000) pour les applications dans le cas d'un processus non stationnaire avec relations de cointégration entre les variables).

Anderson (1951) a montré algébriquement que la RMC se ramène à une analyse canonique entre deux ensembles de résidus de régression ; or, cette analyse canonique peut être très sensible à de faibles variations des données de départ et, par conséquent, les solutions de la RMC peuvent être instables, au sens où une faible perturbation des données d'origine peut modifier substantiellement les résultats obtenus.

Nous reformulons alors le problème posé, en explicitant les propriétés que doit posséder une méthode robuste ; la recherche de solutions possédant ces propriétés nous conduit à proposer une autre méthode, proche des techniques PLS (Partial Least Squares, Tenenhaus (1998), Casin (2000)).

# 2. Le problème

## 2.1 Les données

On considère $P$ variables $Y_p$, $Q$ variables $X_q$ et $R$ variables $Z_r$ . Toutes ces variables sont quantitatives et observées pour les mêmes $n$ individus. Sans perdre en généralité, on suppose que les variables $Y_p$, $p=1,\ldots,P$, $X_q$, $q=1,\ldots Q$ et $Z_r$, $r=1,\ldots,R$ sont centrées et réduites. Enfin, on note $Y=\left(Y_1,\cdots,Y_P\right)$ la matrice $n\times P$, $X=\left(X_1,\cdots,X_Q\right)$ la matrice $n\times Q$ et $Z=\left(Z_1,\cdots,Z_R\right)$ la matrice $n\times R$, $Y_j$ (resp. $X_j$, $Z_j$) étant le vecteur-colonne des n observations pour la variable considérée.

## 2.2 Le modèle multivarié contraint

Les variables $Y$ étant les variables à expliquer et les variables $X$ et $Z$ les variables explicatives, il s'agit de déterminer une matrice $B$ de dimensions $Q\times P$ et une matrice $C$ de dimension R x P telles que :

$$Y=XB+ZC+U$$

$U$ désignant la matrice $n\times P$ des résidus.
Ce modèle multivarié est contraint lorsque $B=\alpha\beta$ où $\alpha$ est une matrice de dimension $Q\times S$, et $\beta$ une matrice de dimension $S\times P$ avec $S<Q$.

## 2.3 La solution

La solution (Anderson (1951), Johansen (1995) par exemple) est obtenue en maximisant le logarithme de la fonction de vraisemblance, après avoir supposé que chaque ligne $i$ de $U$ est

un vecteur aléatoire à *P* dimensions suivant une loi de Gauss d'espérance le vecteur nul et de matrice de variances-covariances $\Omega$ indépendante de *i*
$R^Y$ (resp. $R^X$) désignant le tableau dont les colonnes sont les résidus des régressions des variables X (resp Y) par rapport aux variables Z, les S variables $R^X\alpha$ engendrent l'espace des S premières variables canoniques issues de l'analyse des tableaux $R^Y$ et $R^{X.}$ . La valeur de *S* est obtenue en testant l'hypothèse de nullité des corrélations canoniques de rang supérieur à *S*.

Une fois les *S* colonnes de $X\alpha$ déterminées, l'équation s'écrit :

$$Y = X\alpha\beta + ZC + U$$

$\beta$ et C sont obtenus en régressant Y par rapport à $R^X\alpha$ et C.

## 3. La stabilité des solutions

Les solutions du modèle de régression multivariée contrainte sont donc obtenues en effectuant une analyse canonique entre deux espaces engendrés par des résidus de régressions.
Deux problèmes se posent alors, le premier concernant la stabilité des résidus, le second la stabilité des résultats de l'analyse canonique. Dans ce paragraphe, nous mettons en évidence les critères que doit vérifier N, une combinaison linéaire des colonnes de $R^X$, soit N = $R^X\alpha$, pour être robuste.

### 3.1 La stabilité des résidus des régressions

Le modèle s'écrit $Y = XB + ZC + U$ , où U est une variable aléatoire à P dimensions.

Soit $t_p$ , p=1,…P une petite perturbation de la variable $Y_p$ non corrélée avec les variables Y et les variables Z, et soit $R_p^{Y+t}$ le résidu de la régression de $Y_p + t_p$ par rapport aux variables Z.
Puisque $t_p$ est non corrélé avec les variables *Z* :

$$R_p^{Y+t} = R_p^Y + t_p \text{ et } R^2\left(R_p^Y, R_p^{Y+t}\right) = \frac{\mathrm{Cov}^2\left(R_p^Y, R_p^{Y+t}\right)}{Var(R_p^Y)Var(R_p^{Y+p})} = \frac{1}{1 + \frac{Var(t_p)}{Var(R_p^Y)}}$$

Lorsque la variable $Y_p$ est très bien expliquée par les variables Z, alors la variance du résidu de la régression de Y par rapport aux variables Z est faible. $Var(R_p^Y)$ est alors petit par rapport à la variance de la perturbation, c'est à dire par rapport à $Var(t_p)$ et donc le coefficient de détermination entre $R_p^Y$ *et* $R_p^{Y+t}$ est faible. Une faible variation des données d'origine $Y_p$ peut donc avoir pour conséquence une importante modification d'une des variables d'un des deux ensembles et donc une modification importante des résultats de l'analyse canonique, tant en ce qui concerne les valeurs propres que les vecteurs propres issus de cette analyse.

**Critère 1** :

Pour que $N=R^X\alpha$ soit stable, il est nécessaire que N soit fortement corrélée aux résidus dont la variance est élevée, c'est à dire qu'il est nécessaire que $\sum_p R^2(N,R_p^Y)Var(R_p^Y)$ ait une valeur élevée.

### 3.2 La stabilité des résultats de l'analyse canonique

Soit $d_q$ une "petite" pertubation des colonnes de $R_q^X$ pour p=1,...,P non corrélée avec les variables Y et avec les variables Z.

Ecrivons $N=\sum_{q=1}^{Q}\lambda_q R_q^X$ avec $\sum_{q=1}^{Q}(\lambda_q)^2=1$ et notons $M=\sum_{q=1}^{Q}\lambda_q(R_q^X+d_q)=N+D$ alors, en procédant comme dans le paragraphe précédent, on obtient :

$$R^2(N,M)=\frac{1}{1+\frac{\text{var}(D)}{\text{var}(N)}}$$

$R^2$(M,N) est d'autant plus proche de 1 que $\frac{\text{Var(D)}}{\text{Var(N)}}$ est petit, donc que $\frac{\lambda' V\lambda}{\text{Var(N)}}$ est petit, V étant la matrice de variances-covariances entre les perturbations ; si on suppose que les perturbations ont même variance et sont deux à deux indépendantes, alors la valeur de $\lambda' V\lambda$ est constante, quel que soit $\lambda$, et donc $R^2$(M,N) est d'autant plus proche de 1 que Var(N) sera élevé, c'est à dire que $N$ sera proche des premières composantes principales issues de l'analyse du tableau $R^Y$. A l'inverse, si $\text{var}(N)$ a une faible valeur, alors $R^2(N,M)$ sera proche de 0, ce qui signifie que les solutions de la régression multivariée contrainte sont très instables.

**Critère 2 :**

Pour qu'une faible perturbation des variables de départ n'entraîne pas une modification importante des résultats de l'analyse canonique, il est nécessaire que Var(N) ait une valeur élevée.

## 4. La régression multivariée contrainte PLS

L'objet de ce paragraphe est de construire une nouvelle technique, la régression multivariée contrainte PLS (RMC PLS), dont les solutions au problème posé par la RMC sont peu sensibles à de faibles perturbations des données.

### 4. 1 Le critère à maximiser

A l'étape 1, considérons $N^1$, combinaison linéaire des résidus $R_q^X$, c'est à dire telle que $N^1=R^X\alpha^1$, avec $\left(\alpha^1\right)'\alpha^1=1$.

Les deux conditions suivantes doivent être remplies par $N^1$ :

- il est nécessaire que $B^1=\sum_p R^2\left(N^1,R_p^Y\right)\mathrm{var}\left(R_p^Y\right)$ ait une valeur élevée ( critère 1)
- il est nécessaire que $A^1=\mathrm{var}\left(N^1\right)$ soit élevée ( critère 2)

Comme il n'est pas possible de maximiser simultanément $A^1$ et $B^1$, on choisit ici de maximiser un compromis, le produit $A^1B^1$ :

$$A^1B^1=\mathrm{Var}(N^1)\sum_p R^2(N^1,R_p^Y)\mathrm{Var}(R_p^Y)$$

$$=\sum_p \mathrm{Cov}^2(N^1,R_p^Y)$$

$$=\frac{(\alpha^1)'(R^X)'R^Y(R^Y)'R^X\alpha^1}{n^2}$$

A l'étape 2, le problème se pose dans les mêmes termes, sauf qu'il s'agit de déterminer une variable $N^2$, dont le pouvoir explicatif complète le pouvoir explicatif de la variable $N^1$ ; autrement dit, on impose que la corrélation entre $N^1$ et $N^2$ est nulle, ou, ce qui est équivalent, que $N^2$ est une combinaison linéaire des résidus des régressions des variables $R_q^X$ par la variable $N^1$. On note $R^{X1}$ la matrice dont les q colonnes sont ces résidus.
$N^2$ étant donc orthogonal à $N^1$, si $R^{Y_1}$ est la matrice dont les colonnes $R_p^{Y1}$ sont les résidus des régressions des variables $R_p^Y$ par la variable $N^1$, alors $(R_p^Y)'N^2=(R_p^{Y1})'N^2$ ; il s'agit de déterminer $N^2$ telle que $N^2=R_q^X\alpha^2$, avec $\left(\alpha^2\right)'\alpha^2=1$, de telle que manière que $A^2=\mathrm{var}\left(N^2\right)$ et $B^2=\sum_p R^2\left(N^2,R_p^{Y1}\right)\mathrm{var}\left(R_p^{Y1}\right)$ aient des valeurs élevées. Comme à l'étape 1, on maximisera donc le produit

$$A^2B^2=\sum_p \mathrm{Cov}^2(N^2,R_p^{Y1})$$

$$=\frac{(\alpha^2)'(R^{X1})'R^{Y1}(R^{Y1})'R^{X1}\alpha^2}{n^2}$$

Plus généralement, à l'étape k+1, on détermine la valeur maximale de

$$A^{k+1}B^{k+1}=\sum_p \mathrm{Cov}^2(N^{k+1},R_p^{Yk})$$

$$=\frac{(\alpha^{k+1})'(R^{Xk})'R^{Yk}(R^{Yk})'R^{Xk}\alpha^{k+1}}{n^2}$$

avec $\left(\alpha^{k+1}\right)'\alpha^{k+1}=1$, $R^{Yk}$ (resp. $R^{Xk}$) désignant la matrice dont les colonnes $R_p^{Yk}$ (resp. $R_q^{Xk}$) sont les résidus des régressions des variables $R_p^{Y}$ (resp. $R_q^{X}$) par les variables $N^1,\ldots,N^k$.

### 4.2 La solution

Le vecteur $\alpha^{k+1}$ maximisant $\left(\alpha^{k+1}\right)'\left(R^{Xk}\right)'R^{Yk}\left(R^{Yk}\right)'R^{Xk}\alpha^{k+1}$ sous la contrainte $\left(\alpha^{k+1}\right)'\alpha^{k+1}=1$ est le premier vecteur propre de $\left(R^{Xk}\right)'R^{Yk}\left(R^{Yk}\right)'R^{Xk}$.

A chaque étape k, il convient d'examiner la valeur de $A^k$ et la valeur de $B^k$ ; la variable $N^k$ obtenue ne sera retenue que si ces valeurs sont suffisamment élevés. Ainsi, il est possible de ne pas retenir les résultats d'une étape et de retenir ceux d'une étape ultérieure.
Notons aussi que si, à l'issue de l'étape k, $\sum_p \mathrm{Var}(R_p^{Yk})$ ou $\sum_q \mathrm{Var}(R_q^{Xk})$ ont de petites valeurs, il est inutile de poursuivre l'analyse.

## 6 Conclusion

L'analyse canonique est peu utilisée en pratique à cause des problèmes d'interprétation de ses résultats. En effet, l'analyse canonique s'intéresse aux relations linéaires existant entre des espaces engendrés par les variables et non directement aux relations linéaires existant entre les variables des deux ensembles, ce qui se traduit ici par la non robustesse des solutions obtenues par la RMC. Il est préférable d'utiliser d'autres techniques, comme celle proposée ici, pour obtenir des solutions robustes.

## Références


[1] ANDERSON, T.W. [1951], «*The statistical analysis of time series* » John Wiley, New York
[2] [1] ANDERSON, T.W.[2002], «Reduced rank regression in cointegrated models", *Journal of Econometrics*» 106, 203-216
[3] CASIN Ph. [2000], «Une nouvelle solution au problème de Kloek et Mennes », *Annales d'économie et de statistique*» 193-209
[4] GOURIEROUX Ch., MONTFORT A., RENAULT E.[1993], « Test sur le Noyau, l'Image et le Rang de la Matrice des Coefficients d'un Modèle Linéaire Multivarié, *Annales d'Economie et de Statistique*» 32, 81-112.
[5] JOHANSEN S. [1988], «Statistical analysis of cointegration vectors", *Journal of economic dynamics and Control* »12, 231-254
[6] JOHANSEN S. [1995], «*Likehood-based inference in cointegrated vector auto-regressive models*», Oxford University Press
[7] JOHANSEN S. [2000], «Modelling of cointegration in the vector auto-regressive models", *Economic Modelling*» 17, 359-373
[8] PERRON P., CAMPBELL J.Y. [1992], «Racines unitaires en macroéconomie : le cas multidimensionnel», *Annales d'économie et de statistique*» 27 , 1-50
[9] TENENHAUS M. [1998], « La régression PLS », Technip, Paris