Université de Montréal

COMPARAISON ENTRE LA RÉGRESSION SUR LES COMPOSANTES PRINCIPALES (PCR), LA RÉGRESSION PAR ANALYSE DES VALEURS LATENTES (LRR) ET LA RÉGRESSION PAR MOINDRES CARRÉS PARTIELS (PLS)

par

Abdelhai JAHOURI

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en Statistique

septembre 2001

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

COMPARAISON ENTRE LA RÉGRESSION SUR LES COMPOSANTES PRINCIPALES (PCR), LA RÉGRESSION PAR ANALYSE DES VALEURS LATENTES (LRR) ET LA RÉGRESSION PAR MOINDRES CARRÉS PARTIELS (PLS)

présenté par

Abdelhai JAHOURI

a été évalué par un jury composé des personnes suivantes :

Yves LEPAGE
(président-rapporteur)

Robert CLÉROUX
(directeur de recherche)

Aziz LAZRAQ
(co-directeur)

Narayan Chandra GIRI
(membre du jury)

Mémoire accepté le :
La date d'acceptation

SOMMAIRE

L'objectif principal du présent mémoire est de comparer trois méthodes alternatives de la régression linéaire, il s'agit de la régression sur les composantes principales (PCR)[1], la régression par analyse des valeurs latentes (LRR)[2] et la régression par moindres carrés partiels (PLS)[3]. Ces trois méthodes peuvent être utilisées même lorsque la taille de l´échantillon est inférieure au nombre de prédicteurs.

Quatre exemples concrets ont été analysés dans le cas multiple et multivarié, et ce, selon le nombre de variables explicatives par rapport au nombre d'observations. En effet, les deux premiers exemples envisagent le cas où le nombre d'observations est supérieur au nombre de variables explicatives et les deux autres le cas contraire.

Le but de ces analyses est d'évaluer à travers certains nombre de critères, les performances et les portées de chaque méthode.

Les critères pris en considération sont les moyennes et les écarts-type résiduels ainsi que la corrélation entre les valeurs observées et les valeurs estimées de la variable ou des variables dépendantes du modèle.

Nous avons alors comparé ces différentes méthodes en utilisant des algorithmes appropriés afin de déterminer un modèle concret et adéquat.

---

1. En anglais, PCR: principal component regression
2. LRR: latent root regression analysis
3. PLS: partial least squares regression

SUMMARY

The main objective of the present thesis is to compare three alternative methods of the linear regression: principal component regression, latent root regression analysis and partial least squares regression.

These three methods can also be used in the case on here the sample size is smaller than the number of predictors.

Four concrete examples have been analysed in the case of multiple and multivariate regression according to the number of explanatory variables in comparison with the number of observations.

In fact, the two first examples consider the case where the number of observations is greater to the number of independent variables and the two others the opposite case.

The goal of these analyses is to evaluate through a set of criteria, the performance and the range of each method.

The criteria taken in consideration are the residual mean and the residual standard deviation as well as the correlation between the observed values and the estimated values of the variables depending on model.

We have then compared these different methods using fitting algorithms in order to determine a concrete and adequate model.

REMERCIEMENTS

Je tiens à témoigner toute ma gratitude à mon directeur de recherche, Robert CLÉROUX, dont les directives, les conseils et la pédagogie m'ont été précieux pour la réalisation de ce mémoire.

Je le remercie encore pour sa patience, sa grande disponibilité et son support financier.

Je remercie mon codirecteur Aziz LAZRAQ, qui a fourni des corrections et des suggestions pour l'amélioration de ce mémoire.

Spécialement, un grand remerciement à ma mère, mes frères et soeurs pour leur soutien moral et leurs encouragements durant toute ma vie estudiantine.

Je remercie aussi, mon épouse, pour sa présence à mes côtés et son soutien moral.

Pour terminer, un remerciement particulier à tous mes amis.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

En science expérimentale, les phénomènes étudiés peuvent souvent être décrits sous la forme de modèle et ce, afin de comprendre les mécanismes d'un système : ces modèles comprenant aussi bien des variables explicatives que des variables à expliquer. L'objectif est de décrire et de trouver une relation adéquate entre ces deux groupes de variables.

Certes, la régression linéaire est une méthode statistique très utilisée dans la modélisation et ce, dans divers domaines. Dans cette méthode, on estime souvent les paramètres par la méthode des moindres carrés qui, en présence du phénomène de la multicolinéarité a notamment l'inconvénient de conduire à des estimateurs dont les variances sont très grandes et par conséquent aboutissent à un modèle instable. Le remède généralement utilisé dans ce cas est l'élimination d'une ou plusieurs variables explicatives. Mais dans certains domaines comme la chimie et plus particulièrement dans les applications concernant les données de chromatographie ou de spectrographie, nous devons préserver nos variables explicatives.

En outre, il se peut que le nombre d'observations soit inférieur au nombre de variables explicatives.

A cet égard, plusieurs auteurs ont recours à des alternatives fixant comme objectif: l'amélioration de la qualité du modèle de régression linéaire multiple. Nous rappelons les principes de quelques techniques de régression, spécialement les techniques employées pour atténuer les effets de la multicolinéarité. Il s'agit de la régression sur les composantes principales (PCR): (Drapper et Smith, 1998);

(Rawlings, Pantula et Dickey, 1998), la régression par analyse des valeurs latentes (LRR): (Webster, Gunst et Mason, 1974); (Vigneau, Bertrand et Qannari, 1996) et la régression par moindres carrés partiels (PLS): (Tenenhaus, Gauchi et Menardo, 1995); (Gauchi, 1995); (Tenenhaus, 1998).

Dans le premier chapitre, nous allons présenter l'aspect théorique des différentes méthodes suggérées. Nous commençons par un aperçu sur la régression multivariée considérée comme base pour les régressions étudiées et particulièrement la régression en composantes principales et la régression par l'analyse des valeurs latentes. Ensuite, on distingue ces deux régressions par leurs caractéristiques et leurs spécificités. Un grand intérêt dans la littérature est donné à la régression par moindres carrés partiels.

Dans le deuxième chapitre, nous illustrons les différentes méthodes par deux exemples où le nombre de variables explicatives est inférieur au nombre d'observations et dont on peut parler du problème de la colinéarité. En effet, nous présentons et nous commentons les différents résultats en se basant sur les sorties de plusieurs programmes considérés.

Dans le troisième chapitre, la compétition des trois méthodes sera illustrée par le traitement et l'analyse de deux exemples où le nombre de variables explicatives est largement supérieur au nombre d'observations et sera basée sur deux sortes de données : les données de calibration pour former nos différents modèles et les données de validation pour les tester en calculant l'erreur de prédiction ainsi que les corrélations entre les valeurs observées et les valeurs estimées des variables dépendantes.

Notons que la comparaison sera basée sur l'écart type résiduel en fonction du nombre de composantes introduites dans le modèle. Cette fonction est obtenue par:

$$\sigma_e = \sqrt{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij} - \hat{y}_{ij})^2}$$

où y_{ij}: $j$ème observation sur la $i$ème composante de $y : p \times 1$.

En introduisant étape par étape une composante validée par la procédure FORWARD pour le cas de la régression sur les composantes principales et la régression par moindres carrés partiels et par la sélection proposée par la règle de Webster, Gunst et Mason qui consiste à éliminer tous les valeurs latentes correspondants à une valeur propre inférieure à 0,05 et un coefficient de la variable dépendante inférieur à 0,20 dans le cas de la régression par analyse des valeurs latentes.

Les méthodes sont comparées pour deux critères: d'une part la valeur de σ_e à minimiser et d'autre part son comportement autour de ce minimum. Cette approche est considérée comme robuste (Vigneau, Bertrand et Qannari, 1996). En effet, le nombre de composantes reste optimum et valable pour des données de calibration comme pour des données de validation. En revanche, elle sera basée aussi sur les corrélations entre les valeurs observées et les valeurs estimées des variables dépendantes et la stabilité des modèles.

Chapitre 1

ASPECT THÉORIQUE DE LA PROBLÉMATIQUE

1.1. Régression linéaire multivariée

1.1.1. Présentation

Soit $\underline{Y}$: p $\times$ 1 un vecteur aléatoire que l'on cherche à expliquer linéairement par le vecteur $\underline{X}$: $q \times 1$. Le modèle empirique associé est:

$$Y = XB + U$$

où

p: le nombre de variables à expliquer;

q: le nombre de variables explicatives;

n: le nombre d'observations;

Y : $n \times p$ matrice à expliquer à partir de X;

X : $n \times q$ matrice explicative:variables contrôlables;

B : $q \times p$ matrice des coefficients de régression;

U : $n \times p$ matrice des erreurs (des résidus).

1.1.2. Les hypothèses

On distingue deux sortes d'hypothèses:

1.1.2.1. *Les hypothèses stochastiques*

$H1$: les valeurs de la matrice X sont observées sans erreur;
$H2$:$E(U) = 0$, l'espérance mathématique des erreurs est nulle;
$H3$:$E(U'U) = \Sigma$, homoscédasticité (la matrice variance-covariance des erreurs est constante inconnue);
$H4$: les lignes de U sont non corrélées entre elles (ou encore sont indépendantes).

1.1.2.2. *Les hypothèses structurelles*

$H5$: l'absence de colinéarité entre les variables explicatives, cela implique que la matrice $(X'X)$ est non singulière c'est à dire que la matrice inverse $(X'X)^{-1}$ existe;
$H6$: le vecteur $\underline{Y}$ possède une distribution multinormale.

1.1.3. L'estimation de la matrice des coefficients de régression B et de la matrice de variance-covariance Σ

Par la méthode du maximum de vraisemblance, on peut déterminer les estimateurs de B et Σ, on suppose que le rang $X = q$.

Le logarithme de la fonction de vraisemblance est donné par:

$$\begin{aligned} l(B,\Sigma) &= -\frac{n}{2}log|2\pi\Sigma| - \frac{1}{2}tr(Y - XB)\Sigma^{-1}(Y - XB)' \\ &= -\frac{n}{2}log|2\pi\Sigma| - \frac{1}{2}tr\Sigma^{-1}(Y - XB)'(Y - XB) \end{aligned}$$

Théorème 1.1.1. *Le maximum de la fonction* $l(B,\Sigma)$ *est atteint lorsque:*

$B = \hat{B} = (XX')^{-1}X'Y$

et

$\Sigma = \hat{\Sigma} = n^{-1}Y'PY$

avec $P = [I - X(XX')^{-1}X']$

Preuve:

Soient $\hat{Y} = X\hat{B} = X(XX')^{-1}X'Y$, et $\hat{U} = Y - X\hat{B} = PY$

$Y - XB = (Y - X\hat{B}) + (X\hat{B} - XB) = \hat{U} + X(\hat{B} - B)$

et donc, $(Y - XB)'(Y - XB) = [\hat{U} + X(\hat{B} - B)]'[\hat{U} + X(\hat{B} - B)]$

$= \hat{U}'\hat{U} + (\hat{B} - B)'X'X(\hat{B} - B) + \hat{U}'X(\hat{B} - B) + (\hat{B} - B)'X'\hat{U}$

or $\hat{U}'X(\hat{B} - B) = Y'P'X(\hat{B} - B)$

$= Y'(X - X(XX')^{-1}X'X)(\hat{B} - B)$

$= Y'.0.(\hat{B} - B) = 0$

et $[(\hat{B} - B)'X'\hat{U}]' = \hat{U}'X(\hat{B} - B) = 0 \Leftrightarrow (\hat{B} - B)'X'\hat{U} = 0$

D'où:

$tr\Sigma^{-1}(Y - XB)'(Y - XB) = tr\Sigma^{-1}[\hat{U}'U + (\hat{B} - B)'X'X(\hat{B} - B)]$

$= tr\Sigma^{-1}\hat{U}'U + tr\Sigma^{-1}(\hat{B} - B)'X'X(\hat{B} - B)$

on a:

$\hat{U} = PY$

$\hat{U}'\hat{U} = Y'P'PY = Y'P^2Y = Y'PY = n\hat{\Sigma}$

car $P' = P$ et $P^2 = P$

P est une matrice symétrique et idempotente.

D'où $l(B,\Sigma)$ s'écrit:

$l(B,\Sigma) = -\frac{n}{2}log|2\pi\Sigma| - \frac{n}{2}tr\Sigma^{-1}\hat{\Sigma} - \frac{1}{2}tr\Sigma^{-1}(\hat{B} - B)'X'X(\hat{B} - B).$

Maximiser $l(B,\Sigma)$ par rapport B revient à minimiser

$tr\Sigma^{-1}(\hat{B}-B)'X'X(\hat{B}-B)$

Cette expression, positive, est minimale lorsque $B=\hat{B}$, elle vaut alors 0.

Il reste à maximiser $l(\hat{B},\Sigma)$ par rapport à Σ

$l(\hat{B},\Sigma) = -\frac{np}{2}log2\pi - \frac{n}{2}log|\Sigma| - \frac{n}{2}tr\Sigma^{-1}\hat{\Sigma} = -\frac{np}{2}log2\pi - \frac{n}{2}[log|\Sigma| + tr\Sigma^{-1}\hat{\Sigma}]$.

On utilisera le théorème suivant:

Théorème 1.1.2. *Soient A et Σ deux matrices, $A \geq 0, \Sigma > 0$*

et $f(\Sigma) = |\Sigma|^{-\frac{n}{2}}exp(-\frac{1}{2}tr\Sigma^{-1}A)$.

Alors, f est maximale pour $\Sigma = n^{-1}A$

et $f(n^{-1}A) = |n^{-1}A|^{-\frac{n}{2}}exp-\frac{np}{2}$

Preuve: (Mardia, 1979) pages (104-105).

$l(\hat{B},\Sigma)$ vérifie toutes les conditions du théorème. Par conséquent le Σ qui maximise $l(\hat{B},\Sigma)$ est $\hat{\Sigma}$.

$l(\hat{B},\Sigma) = -\frac{np}{2}log2\pi - \frac{n}{2}log|\Sigma| - \frac{1}{2}tr\Sigma^{-1}(n\hat{\Sigma})$

et donc on est dans les conditions du théorème.

(En effet, $logf(\Sigma) = -\frac{n}{2}log|\Sigma| - \frac{1}{2}tr\Sigma^{-1}(n\hat{\Sigma})$)

et donc le Σ qui maximise $l(\hat{B},\Sigma)$ est donné par:

$n^{-1}A = n^{-1}n\hat{\Sigma} = \hat{\Sigma} = n^{-1}Y'PY$.

1.1.4. La mesure de la qualité de l'ajustement dans le cas de la régression multivariée

On considère le vecteur $\underline{Y}' = (Y_1,Y_2,....,Y_p)$ des variables à expliquer et le vecteur $\underline{X}' = (X_1,X_2,...,X_q)$ des variables explicatives.

On suppose que le vecteur $\begin{pmatrix} Y \\ \hline X \end{pmatrix}$ a pour matrice de variance-covariance

$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}$ où

$\Sigma_{yy} : p \times p, \quad \Sigma_{xx} : q \times q \quad$ et $\quad \Sigma_{xy} = \Sigma'_{yx} : p \times q.$

Définition 1.1.1. *On appelle indice de redondance $\rho I(Y,X)$ comme il a été défini par Stewart et Love (1968) et généralisé par Gleason (1976):*

$$\rho I(Y,X) = \frac{tr(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})}{tr\Sigma_{yy}},$$

ou encore,

$$\rho I(Y,X) = \frac{\sum_{j=1}^{p} \sigma_{yj}^2 \bar{R}^2(Y_j; X_1,X_2,...,X_q)}{\sum_{j=1}^{p} \sigma_{yj}^2},$$

où $\bar{R}^2(Y_j; X_1,X_2,...,X_q)$ est le carré du coefficient de corrélation multiple entre Y_j : la jème composante du vecteur Y et les variables explicatives $X_1,X_2,...,X_q$.
σ_{yj}^2 : étant la variance de la jème composante du vecteur Y.

L'indice de redondance est donc une moyenne pondérée (par les variances) des carrés des coefficients de corrélation multiple entre les composantes du vecteur à prédire et le vecteur de prédiction. C'est aussi la fraction de la variance totale de Y expliquée par la régression multivariée de Y sur X.

1.2. La régression sur les composantes principales

1.2.1. Présentation et estimation des paramètres

Cette approche repose sur l'utilisation comme variables explicatives, des valeurs des composantes principales calculées à partir de la matrice X, les données sont supposées centrées-réduites.

Pour chaque variable expliquée, on peut considérer le modèle linéaire suivant:

$$y = X\beta + \epsilon$$

où $\epsilon \sim N(0,\sigma^2 H)$ et $H = I - n^{-1}11'$

n : le nombre d'observations;

q : le nombre de variables explicatives;

y : $n \times 1$, vecteur de n observations de la variable à expliquer;

X : $n \times q$, matrice de n observations prises sur les variables explicatives;

ϵ : le vecteur de n résidus.

Soient $\lambda_1,\lambda_2,...,\lambda_q$: les valeurs propres non nulles de $(X'X)$.

Soit V la matrice des vecteurs propres orthonormés associés aux valeurs propres $\lambda_1,\lambda_2,...,\lambda_q$,

telle que : $V'V = I$.

On pose $W = XV$,

W étant la transformation des composantes principales. Alors, l'équation du modèle devient:

$$y = W\gamma + \epsilon$$

où $\gamma = V'\beta$

L'estimation des paramètres est donnée par :

$\hat{\gamma} = (W'W)^{-1}W'y$ et $\hat{\epsilon} = y - W\hat{\gamma}$

où

$$\hat{\gamma}_i = \lambda_i^{-1} w_i' y$$

γ_i : la $i$ème composante du vecteur γ: q $\times$ 1,

w_i : la $i$ème colonne de la matrice W: n $\times$ q.

$$\hat{\gamma}_i = \frac{r_{yw_i}}{\sqrt{\lambda_i}}$$

avec $(i = 1,2...,q)$.

r_{yw_i} : le coefficient de corrélation entre y et w_i.

La relation entre les coefficients ordinaires et les coefficients des composantes principales est donnée par :

$$\hat{\beta} = V\hat{\gamma}$$

car:

$$\hat{y} = XV\hat{\gamma}.$$

Dans le cas où la matrice $(X'X)$ est non singulière, le vecteur $\hat{\beta}$ obtenu par l'intermédiaire des composantes principales est théoriquement équivalent au vecteur $\hat{\beta}$ obtenu par les moindres carrés. La variance des coefficients de régression

est explicitée par les valeurs propres de $(X'X)$:

$$Var(\hat{\gamma}_i) = \frac{\sigma^2}{\lambda_i}$$

avec $(i = 1,....,q)$

Il en résulte que:

$$Var(\hat{\beta}_j) = Var(\sum_{i=1}^{q} v_{ji}\hat{\gamma}_i) = \sigma^2 \sum_{i=1}^{q} \frac{v_{ji}^2}{\lambda_i}$$

avec $(j = 1,...,q)$

v_{ji}: le $j$ème élément du $i$ème vecteur propre.

Ces variances sont donc d'autant plus grandes que les éléments des vecteurs v_i sont importants alors que la valeur propre correspondante λ_i est faible.

Si la matrice $(X'X)$ est singulière, on ne dispose que de $r < q$ valeurs propres non nulles et par conséquent r composantes principales. Les relations entre $\hat{\gamma}$,$\hat{\beta}$ et $Var(\hat{\beta}_j)$ restent valables si on considère que i varie de 1 à r au lieu de 1 à q. Ainsi, s'il existe une colinéarité entre par x_1,x_2,x_3 et x_4, par exemple une relation de type $x_4 = x_1 + x_2 - x_3$, alors on obtient la relation entre $\hat{\beta}_1$,$\hat{\beta}_2$,$\hat{\beta}_3$ et $\hat{\beta}_4$: $\hat{\beta}_4 = \hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3$.

Dans le cas où on choisit h composantes principales parmi les r composantes disponibles, soient les composantes associées aux plus grandes valeurs propres de $(X'X)$, il suffit de considérer que V est constitué des h vecteurs propres correspondant aux composantes principales retenues.

Le vecteur $\hat{\beta}$ qu'on obtient est différent du vecteur $\hat{\beta}$ obtenu par les moindres carrés. On le note $\hat{\beta}_{cp}$, c'est un estimateur biaisé mais le biais est compensé par la réduction de la variance des $\hat{\beta}_j$.

L'idée principale de la régression sur les composantes principales est d'extraire de la matrice X l'information utile, en négligeant les fluctuations aléatoires contenues dans les dernières composantes (Palm, 1994).

1.3. La régression par analyse des valeurs latentes

Cette approche est une variante de la régression sur les composantes principales proposée par Webster, Gunst et Mason (1974). Ces auteurs suggèrent de calculer les composantes non à partir de $(X^{'}X)$ mais à partir de $(A^{'}A)$ avec :

$$A = [y|X].$$

Dans le cas de la non colinéarité, on a q+1 vecteurs propres associés aux q+1 valeurs propres formant la matrice V.

$$V = [v_1 v_2 v_3 .....v_{q+1}]$$

avec

$$v_i^{'} = [v_{0i} v_{1i} .....v_{qi}]$$

$(i = 1,2,...,q+1)$.

Les q+1 vecteurs des valeurs des composantes principales sont donnés par:

$$W = [w_1 w_2 .....w_{q+1}] = AV.$$

La $i$ème composante principale de A est donnée par :

$$w_i = Av_i$$

ou aussi,

$$w_i = v_{0i} y_i + X v_i^0$$

avec $v_i^0 = [v_{1i}.....v_{qi}]$.

En présence de colinéarité, on élimine les composantes qui sont sans grand intérêt, c'est à dire on élimine les composantes correspondant à une faible valeur de λ_i et pour lesquelles v_{0i} est également faible.

(Rappelons que les λ_i sont les valeurs propres de AA' mesure la variabilité de la $i$ème composante).

Si $\lambda_i = 0$ avec $v_{0i} \neq 0$ alors, on estime y_i par:

$$\hat{y}_i = -(v_{0i}^{-1}) X v_i^0 = -(v_{0i}^{-1}) \sum_{r=1}^{q} v_{ri} x_r,$$

$\hat{y}_i$ est le $i$ème prédicteur associé à la $i$ème valeur latente w_i.

En pratique, Webster, Gunst et Mason en 1974, proposent d'éliminer la composante i lorsque simultanément $\lambda_i < 0{,}05$ et $v_{0i} < 0{,}20$.

Après avoir éliminé les composantes de faible intérêt, on a l'expression de $\hat{y}$ donnée par:

$$\hat{y} = \sum_{i=1}^{l} q_i \hat{y}_i$$

avec

$q_i = \frac{v_{0i}^2}{\lambda_i} \left(\frac{1}{\sum_{j=1}^{l} \frac{v_{0j}^2}{\lambda_j}} \right)$ et $q_i \geq 0$;

l: le nombre de valeurs propres retenues.

Les éléments du vecteur des coefficients de régression s'obtiennent par les relations suivantes:

$$\hat{\beta}_{w_j} = c \sum_{i=1}^{l} \frac{v_{ji} v_{0i}}{\lambda_i}$$

$(j = 1,....,q)$

avec

$c = \frac{-1}{\sum_{i=1}^{l} \frac{v_{0i}^2}{\lambda_i}}$ d'où le vecteur des coefficients de régression est donné par:

$$\hat{\beta}_w = c \sum_{i=1}^{l} \frac{v_{ji} v_i^0}{\lambda_i}$$

Démonstration voir: (Webster, Gunst et Mason, 1974).

1.4. La régression par moindres carrés partiels(PLS)

1.4.1. Introduction

L'utilisation de la régression par moindres carrés partiels s'est développée dans le domaine de la chimie.

Cette méthode présente à la fois des analogies avec la régression sur les composantes principales et avec la régression par analyse des valeurs latentes. Elle consiste à remplacer les q variables explicatives initiales par $h \leq r$ combinaisons linéaires t_k avec (k=1,...,h) de ces variables:

$$T = [t_1,....,t_h] = XV$$

et à utiliser ces combinaisons linéaires comme des variables explicatives.

La détermination des vecteurs t_k, se fait en tenant compte des variables dépendantes.

1.4.2. La procédure PLS multivariée (PLS2)

Soit $Y = (Y_1,Y_2,...,Y_p)$: variables à expliquer à partir de X,

$X = (X_1,X_2,...,X_q)$: variables explicatives.

On suppose$\begin{pmatrix} Y \\ X \end{pmatrix}$: $(p+q) \times 1$ un vecteur aléatoire avec la matrice variance-covariance $\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$ où

$\Sigma_{yy} : p \times p, \quad \Sigma_{xx} : q \times q \quad$ et $\quad \Sigma_{xy} = \Sigma'_{yx} : p \times q.$

La procédure PLS est une méthode pas à pas permettant de déterminer les composantes adéquates.

On pose

$$X_{(0)} = X, \quad Y_{(0)} = Y \quad et \quad \Sigma_{(0)} = \begin{pmatrix} \Sigma_{yy_{(0)}} & \Sigma_{yx_{(0)}} \\ \Sigma_{xy_{(0)}} & \Sigma_{xx_{(0)}} \end{pmatrix}.$$

On cherche des fonctions linéaires $t_1 = \alpha'_1 X_{(0)}$ et $u_1 = \beta'_1 Y_{(0)}$.

L' objectif est de maximiser $cov(t_1,u_1)$

$$cov(t_1,u_1) = cov(\alpha'_1 X_{(0)},\beta'_1 Y_{(0)}) = \alpha'_1 \Sigma_{xy_{(0)}} \beta_1$$

sous les contraintes:

$||\alpha_1|| = 1$ et $||\beta_1|| = 1$.

En utilisant la méthode de Lagrange on a:

$$L(\alpha_1,\beta_1,\theta_1,\theta_2) = \alpha'_1 \Sigma_{xy_{(0)}} \beta_1 - \theta_1(\alpha'_1 \alpha_1 - 1) - \theta_2(\beta'_1 \beta_1 - 1)$$

θ_1, θ_2 sont les multiplicateurs de Lagrange.

En dérivant par rapport aux paramètres on a:

$\frac{\partial L}{\partial \theta_1} = -(\alpha'_1 \alpha_1 - 1) = 0$

$\frac{\partial L}{\partial \theta_2} = -(\beta_1^{'}\beta_1 - 1) = 0$

$\frac{\partial L}{\partial \alpha_1} = \Sigma_{xy_{(0)}}\beta_1 - 2\theta_1\alpha_1 = 0$

$\frac{\partial L}{\partial \beta_1} = \Sigma_{yx_{(0)}}\alpha_1 - 2\theta_2\beta_1 = 0.$

On obtient :

$\alpha_1^{'}\Sigma_{xy_{(0)}}\beta_1 = 2\theta_1 = 2\theta_2 = \varsigma_1$

$\Sigma_{xy_{(0)}}\beta_1 = \varsigma_1\alpha_1$

$\Sigma_{yx_{(0)}}\alpha_1 = \varsigma_1\beta_1$

$\Sigma_{xy_{(0)}}\Sigma_{yx_{(0)}}\alpha_1 = \varsigma_1^2\alpha_1 = \lambda_1\alpha_1$

$\Sigma_{yx_{(0)}}\Sigma_{xy_{(0)}}\beta_1 = \varsigma_1^2\beta_1$

d'où:

α_1: est le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice $\Sigma_{xy_{(0)}}\Sigma_{yx_{(0)}}$.

β_1: est le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice$\Sigma_{yx_{(0)}}\Sigma_{xy_{(0)}}$.

Après avoir déterminé α_1, β_1 et t_1, on considère la régression de $X_{(0)}$ et aussi $Y_{(0)}$ sur t_1:

$X_{(0)} = P_1t_1 + X_{(1)}$

où:

$P_{(1)} = \Sigma_{xt_1(0)}\sigma^{-1}_{t_1t_1(0)}$

et

$Y_{(0)} = R_1t_1 + Y_{(1)}$

où:

$R_1 = \Sigma_{yt_1(0)}\sigma^{-1}_{t_1t_1(0)}$,

$X_{(1)}$ et $Y_{(1)}$ sont les vecteurs résiduels.

On a:

$\Sigma_{Xt_1(0)} = cov(X_{(0)},t_1) = cov(X_{(0)},\alpha_1^{'}X_{(0)}) = \Sigma_{XX_(0)}\alpha_1$

$\sigma_{t_1t_1(0)} = cov(\alpha_1^{'}X_{(0)},\alpha_1^{'}X_{(0)}) = \alpha_1^{'}\Sigma_{XX_{(0)}}\alpha_1 = var(t_1)$

$\Sigma_{Yt_1(0)} = cov(Y_{(0)},\alpha_1' X_{(0)}) = \Sigma_{YX_{(0)}}\alpha_1$,

t_1 est la première composante de la régression PLS.

Détermination de la deuxième composante de la régression PLS.

La matrice de covariance du vecteur résiduel $\begin{pmatrix} Y_{(1)} \\ X_{(1)} \end{pmatrix}$ est donnée par :

$$\Sigma_1 = \begin{pmatrix} \Sigma_{YY_{(1)}} & \Sigma_{YX_{(1)}} \\ \Sigma_{XY_{(1)}} & \Sigma_{XX_{(1)}} \end{pmatrix}$$

$$= \Sigma_{(0)} - \begin{pmatrix} \Sigma_{Yt_1(0)} \\ \Sigma_{Xt_1(0)} \end{pmatrix} \sigma^{-1}_{t_1t_1(0)} (\Sigma_{t_1Y_{(0)}} \Sigma_{t_1X_{(0)}}).$$

Si le vecteur $\begin{pmatrix} Y \\ X \end{pmatrix}$ est multinormal alors $\begin{pmatrix} Y_{(1)} \\ X_{(1)} \end{pmatrix}$ l'est aussi.

A partir de $\begin{pmatrix} Y_{(1)} \\ X_{(1)} \end{pmatrix}$ on obtient α_2,β_2 et les deux fonctions linéaire $t_2 = \alpha_2' X_{(1)}$ et $u_2 = \beta_2' Y_{(1)}$.

En maximisant $cov(t_2,u_2)$ sous les contraintes :

$||\alpha_2|| = ||\beta_2|| = 1$, on a le système d'équations suivant:

$\Sigma_{XY_{(1)}} \Sigma_{YX_{(1)}} \alpha_2 = \lambda_2 \alpha_2$

$\Sigma_{YX_{(1)}} \Sigma_{XY_{(1)}} \beta_2 = \lambda_2 \beta_2$

d'où on considère la régression de $X_{(2)}$ sur t_2 aussi la régression de $Y_{(2)}$ sur t_2.

On a:

$$X_{(1)} = P_2 t_2 + X_{(2)}$$

où $P_2 = \Sigma_{XX_{(1)}} \alpha_2 var(t_2)^{-1}$

$$Y_{(1)} = R_2 t_2 + Y_{(2)}$$

où $R_2 = \Sigma_{YX_{(1)}} \alpha_2 var(t_2)^{-1}$

avec

$var(t_2) = \alpha_2^{'} \Sigma_{XX_{(1)}} \alpha_2$.

D'où on a:

$$Y = Y_{(0)} = R_1 t_1 + R_2 t_2 + Y_{(2)}.$$

La détermination de la $h$ème composante de régression PLS (t_h)

A l´étape (h-1) on considère le vecteur résiduel $\begin{pmatrix} Y_{(h-1)} \\ X_{(h-1)} \end{pmatrix}$ dont la matrice de variance-covariance est donnée par:

$$\Sigma_{(h-1)} = \begin{pmatrix} \Sigma_{YY_{(h-1)}} & \Sigma_{YX_{(h-1)}} \\ \Sigma_{XY_{(h-1)}} & \Sigma_{XX_{(h-1)}} \end{pmatrix}$$

$$= \Sigma_{(h-2)} - \begin{pmatrix} \Sigma_{Yt_{h-1}(h-2)} \\ \Sigma_{Xt_{h-1}(h-2)} \end{pmatrix} \sigma^{-1}_{t_{h-1}t_{h-1}(h-2)} (\Sigma_{t_{h-1}Y_{(h-2)}} \Sigma_{t_{h-1}X_{(h-2)}}).$$

Si le vecteur $\begin{pmatrix} Y \\ X \end{pmatrix}$ est multinormal alors $\begin{pmatrix} Y_{(h-1)} \\ X_{(h-1)} \end{pmatrix}$ l'est aussi.

A partir de $\begin{pmatrix} Y_{(h-1)} \\ X_{(h-1)} \end{pmatrix}$ on obtient α_h,β_h et les deux fonctions linéaires $t_h = \alpha_h^{'} X_{(h-1)}$ et $u_h = \beta_h^{'} Y_{(h-1)}$.

En maximisant $cov(t_h,u_h)$ sous les contraintes:

$||\alpha_h|| = ||\beta_h|| = 1$, on a le système d'équations suivant:

$\Sigma_{XY_{(h-1)}} \Sigma_{YX_{(h-1)}} \alpha_h = \lambda_h \alpha_h$

$\Sigma_{YX_{(h-1)}} \Sigma_{XY_{(h-1)}} \beta_h = \lambda_h \beta_h$

d'où on considère la régression de $X_{(h)}$ sur t_h et aussi la régression de $Y_{(h)}$ sur t_h.

On a:

$$X_{(h-1)} = P_h t_h + X_{(h)}$$

où $P_h = \Sigma_{XX_{(h-1)}} \alpha_h var(t_h)^{-1}$

$$Y_{(h-1)} = R_h t_h + Y_h$$

où $R_h = \Sigma_{YX_{(h-1)}} \alpha_h var(t_h)^{-1}$

avec

$var(t_h) = \alpha_h^{'} \Sigma_{XX_{(h-1)}} \alpha_h$.

D'où on a:

$$Y = Y_{(0)} = R_1 t_1 + R_2 t_2 + ... + R_h t_h + Y_{(h)}$$

$$= \Sigma_{i=1}^{h} R_i t_i + Y_{(h)}.$$

La matrice de variance-covariance $\Sigma_{(h)}$ du vecteur $\begin{pmatrix} Y_{(h)} \\ X_{(h)} \end{pmatrix}$ est donnée par:

$$\Sigma_h = \Sigma_{(h-1)} - \begin{pmatrix} \Sigma_{Yt_h(h-1)} \\ \Sigma_{Xt_h(h-1)} \end{pmatrix} \sigma^{-1}_{t_h t_h(h-1)} (\Sigma_{t_h Y_{(h-1)}} \Sigma_{t_h X_{(h-1)}}).$$

1.4.3. Les propriétés des composantes PLS

Propriété 1.4.1. *Si* $\begin{pmatrix} Y \\ X \end{pmatrix}$ *suit une loi multinormale alors* (t_1,t_2,....,t_h) *avec* $h \leq q$ *suit une distribution multinormale.*

Preuve: (Lazraq et Cléroux, 2001).

Propriété 1.4.2. *Les composantes principales sont indépendantes deux à deux.*

Preuve:

Par hypothèse on a t_1 indépendant de $X_{(1)}$

d'où $cov(t_1,X_{(1)}) = 0$.

Montrons que t_1 est indépendante de t_2.

$cov(t_1,t_2) = cov(t_1,\alpha_2^{'}X_{(1)}) = cov(t_1,X_{(1)})\alpha_2 = 0.$

On suppose que $t_1,t_2,.....,t_{h-1}$ sont indépendantes deux à deux et on montre que $t_1,t_2,.....,t_h$ sont aussi indépendantes.

$cov(t_{h-1},t_h) = cov(t_{h-1},\alpha_h^{'}X_{(h-1)}) = cov(t_{h-1},X_{(h-1)})\alpha_h = 0$

$$\begin{aligned} cov(t_{h-2},t_h) &= cov(t_{h-2},X_{(h-1)})\alpha_h \\ &= cov(t_{h-2},X_{(h-2)} - P_{h-1}t_{h-1})\alpha_h \\ &= (cov(t_{h-2},X_{(h-2)}) - P_{h-1}cov(t_{h-2},t_{h-1}))\alpha_h \\ &= 0 \end{aligned}$$

Puisque :

$cov(t_{h-2},X_{(h-2)}) = 0 \quad$ et $\quad cov(t_{h-2},t_{h-1}) = 0$

$$\begin{aligned} cov(t_{h-3},t_h) &= cov(t_{h-3},X_{(h-1)})\alpha_h \\ &= cov(t_{h-3},X_{(h-3)}) - P_{h-2}t_{h-2} - P_{h-1}t_{h-1})\alpha_h \\ &= 0 \end{aligned}$$

Ainsi de suite, d'où :

$cov(t_{h-4},t_h) = cov(t_{h-5},t_h) = \cdots = cov(t_1,t_h) = 0.$

1.4.4. La signification des composantes t_h dans le modèle de régression PLS

Par définition, l'indice de redondance entre $Y_{(h-1)}$ et t_h est donné par:

$$\begin{aligned}\rho I(Y_{(h-1)},t_h) &= \frac{tr(\Sigma_{YX_{(h-1)}}\alpha_h\sigma^{-1}_{t_ht_h}\alpha_h^{'}\Sigma_{XY_{(h-1)}})}{tr(\Sigma_{YY_{(h-1)}})} \\ &= \frac{\alpha_h^{'}\Sigma_{XY_{(h-1)}}\Sigma_{YX_{(h-1)}}\alpha_h}{var(t_h)tr(\Sigma_{YY_{(h-1)}})} \\ &= \frac{\lambda_h}{var(t_h)tr(\Sigma_{YY_{(h-1)}})}\end{aligned}$$

où ρI : représente la fraction de la variance de $Y_{(h-1)}$ expliquée par t_h par rapport à la variance totale de $Y_{(h-1)}$,

t_h n'explique pas la matrice résiduelle $Y_{(h-1)}$ si et seulement si $\rho I = 0 \Leftrightarrow \lambda_h = 0$ ou encore,

on a :

$$Y_{(h-1)} = R_h t_h + Y_{(h)}$$

et puisque $cov(t_h,Y_h) = 0$ on peut écrire :

$$var(Y_{(h-1)}) = var(R_h t_h) + var(Y_h)$$

$var(Y_{(h-1)}) = \Sigma_{YY_{(h-1)}}$,

$var(R_h t_h) = R_h var(t_h) R_h^{'}$,

avec, $R_h = \frac{\Sigma_{YX_{(h-1)}}\alpha_h}{var(t_h)}$.

Finalement, on a:

$\rho I(Y_{(h-1),t_h}) = \frac{tr(var(R_h t_h))}{tr(var(Y_{(h-1)}))}$

D'où t_h n'explique pas $Y_{(h-1)}$ si et seulement si $tr(var(R_h t_h)) = 0$.

Concernant, la matrice à expliquer Y on a:

$$\begin{aligned} \rho I(Y,t_h) &= \frac{tr(var(R_h t_h))}{tr(var(Y))} \\ &= \frac{\alpha_h' \Sigma_{XY_{(h-1)}} \sigma_{t_h t_h}^{-1} \Sigma_{YX_{(h-1)}} \alpha_h}{tr(\Sigma_{YY})} \\ &= \frac{\lambda_h}{var(t_h) tr(\Sigma_{YY})}. \end{aligned}$$

L'indice de redondance entre Y et t_h est donné par:
si $\rho I \Longleftrightarrow \lambda_h = 0$ alors la composante principale t_h n'explique pas le vecteur, Y.

1.4.5. L'inférence statistique sur ρI

On considère que le vecteur $\begin{pmatrix} Y \\ X \end{pmatrix}$ suit une loi multinormale avec matrice de variance-covariance $\Sigma = \Sigma_{(0)}$.

Après avoir observé un échantillon de n observations on estime la matrice variance-covariance par:

$$\begin{aligned} S = S_{(0)} &= \begin{pmatrix} S_{YY_{(0)}} & S_{YX_{(0)}} \\ S_{XY_{(0)}} & S_{XX_{(0)}} \end{pmatrix} \\ &= \frac{1}{n-1} \begin{pmatrix} A_{YY_{(0)}} & A_{YX_{(0)}} \\ A_{XY_{(0)}} & A_{XX_{(0)}} \end{pmatrix} \end{aligned}$$

avec,

$A_{YY_{(0)}} = \Sigma_{\alpha=1}^{n} (Y_\alpha - \bar{Y})(Y_\alpha - \bar{Y})'$,
$A_{XX_{(0)}} = \Sigma_{\alpha=1}^{n} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$,
$A_{XY_{(0)}} = \Sigma_{\alpha=1}^{n} (X_\alpha - \bar{X})(Y_\alpha - \bar{Y})'$,

$A_{XY_{(0)}} = A'_{YX_{(0)}}$.

On estime $\rho I(Y,X)$ par:

$$RI(Y,X) = \frac{tr(S_{YX}S_{XX}^{-1}S_{XY})}{tr(S_{YY})}$$

et $\rho I(Y,t_h)$ par:

$$RI(Y,t_h) = \frac{tr(S_{Yt_h}s_{t_ht_h}^{-1}S_{t_hY})}{tr(S_{YY})} = \frac{tr(A_{Yt_h}a_{t_ht_h}^{-1}A_{t_hY})}{tr(A_{YY})}.$$

D'après Mardia, Kent et Bibby (1979), on a:

$A_{YY.t_h} = A_{YY} - A_{Yt_h}a_{t_ht_h}^{-1}A_{t_hY} \sim Wishart \quad W_p(\Sigma_{YY},n-2)$

et

$a_{t_ht_h}$ et A_{Yt_h} sont indépendants.

On a donc:

sous l'hypothèse nulle: $\rho I(Y,t_h) = 0$

$\Sigma_{Yt_h} = cov(Y,\alpha'_hX) = \Sigma YX\alpha_h = 0$

$A_{Yt_h}a_{t_ht_h}^{-1}A_{t_hY} \sim Wishart \quad W_p(\Sigma_{YY},1)$

avec,

$a_{t_ht_h}$ et A_{Yt_h} indépendants.

On a donc le rapport:

$$\begin{aligned}\frac{RI(Y,t_h)}{1-RI(Y,t_h)} &= \frac{A_{Yt_h}a_{t_ht_h}^{-1}A_{t_hY}}{tr(A_{YY.t_h})} \\ &= \frac{tr[W_p(\Sigma_{YY},1)]}{tr[W_p(\Sigma_{YY},n-2)]} \\ &= \frac{V_1}{V_2}\end{aligned}$$

et

$$P(\frac{RI(Y,t_h)}{1-RI(Y,t_h)} \leq v) \quad = \quad P(V_1 - vV_2 \leq 0).$$

En utilisant le théorème 3.1 de l'article Lazraq et Cléroux (1988a).

On a:

Théorème 1.4.1. *Si le vecteur* $\begin{pmatrix} Y \\ X \end{pmatrix}$ *est multinormal et, sous* H_0,

$$P(\frac{RI(Y,t_h)}{1-RI(Y,t_h)} \leq v) = P(W \leq 0)$$

où

$W = V_1 - vV_2$ *est distribué comme* $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$ *où les* W_i *sont iid et N(0,1) et où* $\lambda_1, \lambda_2, ..., \lambda_{p(n-1)}$ *sont les p valeurs propres de* Σ_{YY} *ayant chacune une multiplicité q et les p valeurs propres de* $-v\Sigma_{YY}$ *ayant une multiplicité n-1-q.*

D'où on rejette $H_0 : \rho I(Y,t_h) = 0$ au niveau de signification α, si $\frac{RI(Y,t_h)}{1-RI(Y,t_h)} > c_\alpha$, où c_α est le centile d'ordre α de la distribution de $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$, une fonction linéaire des variables aléatoire des khi-deux centrées indépendantes.

Pour le cas particulier où on a une seule variable dépendante PLS(1), on aura $\rho I(Y,t_h) = \rho^2(Y,t_h)$ et $RI = r^2(Y,t_h)$ le carré de coefficient de corrélation simple entre la variable dépendante et le prédicteur t_h, et $\Sigma = \sigma_y^2$. La valeur propre de Σ_{YY} est σ_y^2 et la valeur propre de $-v\Sigma_{YY}$ est $-v\sigma_y^2$ d'où

$$\begin{aligned} \frac{RI(Y,t_h)}{1-RI(Y,t_h)} &= \frac{r^2(y,t_h)}{1-r^2(y,t_h)} \\ &= \frac{V_1}{V_2} \\ &= \frac{\chi_1^2}{\chi_{n-2}^2} \end{aligned}$$

d'où on peut utiliser le test de Fisher comme on peut utiliser le test t de Student.

On rejette H_0, au niveau de signification α si
$\frac{(n-2)r^2(y,t_h)}{1-r^2(y,t_h)} > f^{\alpha}_{(1,n-2)}$ où $f^{\alpha}_{(1,n-2)}$ est le centile d'ordre α de la distribution de Fisher de degrés de liberté (1,n-2) contre l'hypothèse alternative $H_1 : \rho(Y,t_h) > 0$ (test unilatéral) ou le test t de Student :
si $\frac{\sqrt{n-2r(y,t_h)}}{\sqrt{1-r(y,t_h)}} > t^{\alpha}_{n-2}$ où $t^{\alpha}_{(n-2)}$ est le centile d'ordre α de la distribution t de Student centrée $t_{(n-2)}$ à n-2 degrés de liberté.

Chapitre 2

RÉGRESSIONS MULTIPLE ET MULTIVARIÉE: CAS PRATIQUES OÙ N $\geq$ Q

2.1. Introduction

Dans ce chapitre, on s'intéresse à la comparaison des trois méthodes du chapitre 1 sur deux exemples où, le nombre d'observations n est supérieur au nombre de variables explicatives q et la matrice $(X'X)$ est quasi singulière.

Le premier est un cas de régression multiple, il s'agit de l'exemple de Longley (Longley, 1967).

Le second est un exemple de la régression linéaire multivariée, il s'agit de l'exemple sur le Tabac (Waltz, Reid et Colwell, 1948).

2.2. La régression multiple: traitement de l'exemple de Longley

2.2.1. Présentation des données

Cet exemple est célèbre pour le traitement du phénomène de la colinéarité ou la quasi colinéarité entre les variables explicatives du modèle. Cela signifie que la matrice $(X'X)$ est presque non inversible. Les données sont présentées dans le tableau suivant:

TAB. 2.2.1 –. *Les données de l'exemple de Longley*

EMPL	PRICE	GNP	UNEMPL	FORCES	POP	TIME
60,323	83	234,289	2,356	1,59	107,608	1947
61,122	88,5	259,426	2,325	1,456	108,632	1948
60,171	88,2	258,054	3,682	1,616	109,773	1949
61,187	89,5	284,599	3,351	1,65	110,929	1950
63,221	96,2	328,975	2,099	3,099	112,075	1951
63,639	98,1	346,999	1,932	3,594	113,27	1952
64,989	99	365,385	1,87	3,547	115,094	1953
63,761	100	363,112	3,758	3,35	116,219	1954
66,019	101,2	397,469	2,904	3,048	117,388	1955
67,857	104,6	419,18	2,822	2,857	118,734	1956
68,169	108,4	442,769	2,936	2,798	120,445	1957
66,513	110,8	444,546	4,681	2,637	121,95	1958
68,655	112,6	482,704	3,813	2,552	123,366	1959
69,564	114,2	502,601	3,931	2,514	125,368	1960
69,331	115,7	518,173	4,806	2,572	127,852	1961
70,551	116,9	554,894	4,007	2,827	130,081	1962

Il s'agit d'étudier le nombre de travailleurs aux États-Unis (EMPL,en millions) en fonction des variables suivantes:

PRICE: indice des prix;

GNP: produit national brut (milliards de dollars);

UNEMPL: effectif des chômeurs (millions);

FORCES: effectif des forces armées (millions);

POP : effectif de la population (millions);

TIME: date (années).

Dans notre exemple, la matrice $(X'X)$ est quasi singulière. On se trouve en présence d'un phénomène de colinéarité, qui, d'une part, cause des problèmes de

précision numérique lors du calcul des coefficients et, d'autre part, conduit à des variances importantes des coefficients de la régression.

Pour réduire les inconvénients liés à la quasi colinéarité, on supprime une ou des variables explicatives, ou on emploie d'autres alternatives telles que la régression sur les composantes principales, la régression par analyse des valeurs latentes ou la régression par moindres carrés partiels qui font l'objet de notre étude.

La stratégie consiste à étudier l'exemple selon ces trois méthodes alternatives, visant à améliorer la qualité de la régression linéaire et faire une comparaison entre ces alternatives en se basant sur les différents résultats obtenus, en particulier: les écarts types résiduels en fonction des composantes choisies et les corrélations entre les valeurs observées et les valeurs estimées de la variable à expliquer dans le cas de la régression multiple et l'indice de redondance entre les variables à expliquer et les erreurs dans le cas multivarié.

On considère les données centrées réduites.

2.2.2. La régression sur les composantes principales appliquée à l'exemple de Longley

Une application de la régression en composantes principales sur l'exemple de Longley a donné les résultats suivants.

Les valeurs propres de la matrice $(X^{'}X)$ sont:

$\lambda_1 = 4{,}60337710,$

$\lambda_2 = 1{,}17534050,$

$\lambda_3 = 0{,}20342537,$

$\lambda_4 = 0{,}01492826,$

$\lambda_5 = 0{,}00255207,$

$\lambda_6 = 0{,}00037671,$

et la matrice des vecteurs propres (colonne) correspondants est:

variables	t1	t2	t3	t4	t5	t6
x1	0,461835	0,057843	-,149120	-,792874	0,337938	-,135187
x2	0,461504	0,053212	-,277682	0,121621	-,149573	0,818481
x3	0,321317	-,595514	0,728306	-,007646	0,009232	0,107453
x4	0,201510	0,798193	0,561608	0,077255	0,024252	0,017971
x5	0,462279	-,045544	-,195985	0,589745	0,548578	-,311571
x6	0,464940	0,000619	-,128116	0,052287	-,749543	-,450409

A partir de la matrice des vecteurs propres, on calcule les différentes composantes principales qui font l'objet des composantes à considérer dans les modèles.

TAB. 2.2.2 –. *Les coefficients de corrélation simple de y et des composantes, r_{yt_h}, les coefficients de détermination multiples en fonction du nombre de composantes principales retenues, R^2, et les coefficients de régression P_h (PCR-Longley)*

coefficient	1	2	3	4	5	6
r_{yt_h}	0,9562	0,12096	-0,23892	-0,01243	-0,08875	-0,03849
R^2	0,9143	0,9713	0,9860	0,9938	0,9953	0,9955
P_h	0,44565	0,11157	-0,52973	-0,10174	-1,75679	-1,98272

TAB. 2.2.3 –. *Les écarts types des estimateurs des coefficients de régression et la somme des carrés des erreurs en fonction du nombre de composantes principales retenues(PCR-Longley)*

composante	t1	t2	t3	t5	SCE
1	0,03648				1,28621
2	0,03447	0,06822			1,06675
3	0,01594	0,03154	0,07582		0,21050
4	0,01103	0,02182	0,05245	0,46830	0,09235

TAB. 2.2.4 –. *Les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées en fonction du nombre de composantes retenues(PCR-Longley)*

	1	2	3	4	5	6
σ_e	0,2928	0,2667	0,1185	0,0785	0,0684	0,0672
$corr(y,\hat{y})$	0,9562	0,9638	0,9930	0,9969	0,9977	0,9977

L'utilisation de la procédure FORWARD montre à un niveau de signification de 5%, quatre composantes qui ont été retenues soient t1,t2,t3 et t5.

On constate que la première composante explique 91,43% de la variabilité de y, que les deux premières composantes expliquent 97,13% soit une influence partielle de la deuxième composante sur y de 5,71%. La troisième et la cinquième composante représentent respectivement une explication de y évaluée à 1,49% et 0,79%.

Une très forte corrélation positive entre la première composantes retenue et la variable dépendante soit 0,9562, alors, que la corrélation avec les autres composantes est négligeable.

On constate une cohérence des signes entre les coefficients de régression et les coefficients de corrélation ce qui donne un avantage de stabilité des modèles.

La variabilité des écarts types des coefficients de régression est très faible pour les trois premières composantes et la somme des carrés des erreurs diminue en ajoutant une nouvelle composante dans le modèle.

Par ailleurs, il apparaît que les écarts types résiduels baissent au fur et à mesure qu'une composante est présente dans le modèle. La corrélation entre les valeurs observées et les valeurs estimées est très importante en considérant le modèle avec une seule composante, soit 0,9562 alors que l'apport des autres composantes sur cette fonction est très faible.
(Il n'y a pas intérêt à intégrer toutes les composantes qui n'apportent pas beaucoup à l'explication de la variabilité de y et qui sont responsables de l'augmentation de la variance des coefficients de régression.)

2.2.3. La régression par analyse des valeurs latentes appliquée à l'exemple de Longley

Cette méthode consiste comme on l'a déjà montré au premier chapitre, à déterminer des variables latentes à partir de l'analyse en composantes principales sur la variable à expliquer et les variables explicatives, c'est à dire le tableau [y|X]=A.

Pour l'exemple considéré, les sept valeurs propres de la matrice (A'A)sont:

$\lambda_1 = 5{,}53306768$

$\lambda_2 = 1{,}18755464$

$\lambda_3 = 0{,}25221631$

$\lambda_4 = 0{,}01523852$

$\lambda_5 = 0{,}01063626$

$\lambda_6 = 0{,}00102794$

$\lambda_7 = 0{,}00025864$

et la matrice des vecteurs propres correspondants est:

variables	t1	t2	t3	t4	t5	t6	t7
y	0,412723	0,092594	-,404200	0,197527	0,713790	0,295199	0,148505
x1	0,422556	0,033315	-,041548	0,685443	-,538412	0,227541	-,084354
x2	0,423276	0,030039	-,156676	-,188638	-,214812	-,509204	0,674070
x3	0,279152	-,617261	0,677390	0,072977	0,224816	0,063207	0,149478
x4	0,188731	0,776563	0,585988	-,048667	0,104651	0,052908	0,042783
x5	0,421850	-,069241	-,084342	-,667413	-,273938	0,510754	-,169656
x6	0,424784	-,023659	-,031576	-,049330	0,133925	-,578040	-,680810

En présence de la colinéarité, la règle empirique de Webster et al. propose que certaines variables latentes sont qualifiées comme relevant d'une quasi-colinéarité prédictive et d'autres d'une quasi-colinéarité non prédictive, les premières sont

incluses dans l'équation de prédiction et les dernières sont écartées d'où la septième composante doit être éliminée.

Après avoir analysé ces valeurs latentes par le biais de la fonction qui permet

TAB. 2.2.5 –. *Les différentes valeurs latentes obtenues.(LRR-Longley)*

	1	2	3	4	5	6
1	-3,76198	-0,68214	-0,40578	-0,15107	0,06985	0,028175
2	-3,23996	-0,78071	-0,71289	0,09384	-0,13630	0,024016
3	-2,76212	-1,54128	0,49783	-0,00115	0,02208	0,011100
4	-2,40911	-1,26235	0,10208	-0,06132	0,01443	-0,068417
5	-1,54108	1,25292	0,06425	0,07501	-0,10266	0,004691
6	-1,09475	1,92106	0,25504	0,01231	-0,13613	-0,019782
7	-0,65400	1,93037	-0,04604	-0,07641	-0,01207	0,023225
8	-0,15458	0,53618	1,14699	-0,04878	0,05848	0,014478
9	0,18095	0,70141	0,06466	-0,06481	0,15781	-0,050563
10	0,71709	0,58959	-0,43742	0,08077	0,24150	0,021744
11	1,21404	0,45353	-0,51956	0,13341	0,04364	0,014689
12	1,77909	-0,93445	0,76349	0,18211	-0,05204	0,011878
13	2,15654	-0,40144	-0,27470	0,13644	-0,03821	-0,048223
14	2,64632	-0,51177	-0,39421	0,06085	-0,00432	-0,009336
15	3,26097	-1,05153	0,24866	-0,07095	-0,01063	0,047501
16	3,66258	-0,21939	-0,35239	-0,30026	-0,11543	-0,005178

d'estimer les valeurs de la variable à expliquer (voir annexe E), on calcule les corrélations et les écarts types suivants:

TAB. 2.2.6 –. *Les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées en fonction du nombre de valeur latentes retenues(LRR-Longley)*

	1	2	3	4	5	6
σ_e	5,5184	4,9666	1,1692	0,5374	0,1354	0,0831
$corr(y,\hat{y})$	-0,9574	-0,8764	0,4145	0,8519	0,9902	0,9963

Une lecture des tableaux nous indique que la méthode basée sur les valeurs latentes donne des résultats de plus en plus intéressants si le nombre de valeurs latentes s'accroît.

On constate que la régression par analyse des valeurs latentes est mauvaise, en considérant les quatre premières valeurs latentes relativement à la régression sur les composantes principales.

La corrélation entre les valeurs observées et les valeurs estimées sont négativement corrélées pour la première et la deuxième valeur latente et marquent successivement, -0,9574 et -0,8764 alors elle atteint 0,9963, en considérant toutes les valeurs latentes avec un écart-type résiduel de 0,0831.

On remarque que les résultats de la régression par analyse des valeurs latentes, si on considère le modèle avec les cinq ou les six valeurs latentes sont compétitifs aux résultats de la régression sur les composantes principales.

2.2.4. La régression par moindres carrés partiels appliquée à l'exemple de Longley

Les composantes de la méthode des moindres carrés partiels ont été obtenues par le déroulement de l'algorithme de la fonction PLS(1) programmée en Splus (voir l'annexe B). Cette fonction possède de nombreuses propriétés mathématiques qui justifient l'algorithme et permettent une généralisation au cas de plusieurs variables dépendantes. Son application sur les données de Longley a

révélé les résultats suivants:

TAB. 2.2.7 –. *Les valeurs des composantes PLS, les coefficients de corrélation simple entre y et les composantes PLS,r_{yt_h}, les coefficients de détermination en fonction du nombre de composantes PLS, R^2, les coefficients de régression, P_h (PLS-Longley)*

i	1	2	3	4	5	6
1	-566215000	-1,300901e+20	-0,023266	0,0335007	-7,8805e-04	9,2431e-05
2	-476180664	-1,094044e+20	-0,011557	0,0540126	3,8049e-04	-3,9707e-04
3	-474417262	-1,089992e+20	-0,15182	-0,0129297	-6,675e-05	-1,2437e-06
4	-379348124	-8,715672e+19	-0,101160	0,0106449	-1,0894e-04	3,4100e-04
5	-222055205	-5,101805e+19	0,104385	-0,0123154	8,6488e-05	-4,6700e-04
6	-155255927	-3,567066e+19	0,147380	-0,0297456	2,2391e-05	-2,6431e-04
7	-85773287	-1,970675e+19	0,160741	-0,0162058	-2,8725e-04	2,7676e-05
8	-87125585	-2,001744e+19	-0,029209	-0,0750939	-2,4408e-04	1,2105e-04
9	33826343	7,771735e+18	0,051345	-0,0118405	-1,5389e-04	5,5051e-04
10	112923936	2,594472e+19	0,064213	0,0130959	4,6096e-04	4,6036e-04
11	200298869	4,601945e+19	0,061744	0,0221880	7,4945e-04	9,5694e-05
12	214345665	4,924676e+19	-0,128969	-0,0390288	7,9715e-04	-2,3590e-04
13	349268307	8,024576e+19	-0,018698	0,0197158	7,2445e-04	9,0973e-06
14	424905114	9,762361e+19	-0,023231	0,0258508	3,7507e-04	4,1490e-05
15	488717976	1,122849e+20	-0,10724	-0,0074089	-5,2518e-04	-2,4541e-04
16	622084845	1,429264e+20	0,005345	0,0255601	-1,4223e-03	-1,2838e-04
r_{yt_h}	0,98223	0,98223	0,11048	0,08769	0,06370	0,07403
R^2	0,9648	0,9648	0,9770	0,9847	0,9901	0,9942
P_h	2,668013E-9	1,17376	2,74100	257,04183	106,86211	0

TAB. 2.2.8 –. *Les écarts types des coefficients de régression et la somme des carrés des erreurs en fonction du nombre de composantes PLS retenues(PLS-Longley)*

valeurs	t1	t3	t4	t6	SCE
1	1,36256E-10				0,52842
2	1,14308E-10	0,44710			0,34533
3	9,70966E-11	0,37978	1,11740		0,23000
4	8,12954E-11	0,31797	0,93556	103,91851	0,14780

TAB. 2.2.9 –. *Les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées en fonction du nombre de composantes PLS (PLS-Longley)*

	1	2	3	4	5	6
σ_e	0,1877	0,1517	0,1238	0,0993	0,0761	0,0761
$corr(y,\hat{y})$	0,9822	0,9884	0,9923	0,9951	0,9971	0,9971

L'utilisation de la méthode FORWARD montre que 4 composantes PLS obtenues par la fonction PLS(1) sont significatives à un niveau de signification de 5% soient t1, t3, t4 et t6. Le nombre de composantes concorde avec celui de la méthode de la régression sur les composantes principales.

Un aperçu des différents tableaux obtenus, montre que la première composante PLS est fortement corrélée avec la variable dépendante soit 0,9822 alors que les autres composantes PLS ont une très faible corrélation avec y.

La qualité de l'ajustement est très élevée dans les différents modèles considérés. La première composante PLS explique 96,48% de la variabilité de la variable dépendante pour atteindre une valeur de 98,47% avec les composantes PLS significatives.

Signalons que les écarts types des coefficients de régression sont très proches de zéro pour les deux premières composantes PLS et qu'il existe une cohérence de signe entre les coefficients de corrélation et les coefficients de régression selon les différents composantes PLS.

Le dernier tableau qui décrit les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées selon les différents modèles considérés montre que les écarts types résiduels baissent avec la présence de composantes PLS dans le modèle et que les corrélations sont importantes depuis le premier modèle.

Il est clair que la considération d'une seule composante dans le modèle pour la régression sur les composantes principales et la régression par moindres carrés partiels donne un avantage à la dernière méthode.

Une présentation s'étend des trois méthodes selon un exemple concret où on a trois variables à expliquer ainsi une conclusion et des graphiques sont présentés à la fin du chapitre.

2.3. La régression multivariée: traitement de l'exemple du Tabac

2.3.1. Présentation des données

L'exemple intitulé Tabac est un cas qui traite le problème de la régression linéaire multivariée, où le nombre de variables explicatives est inférieur au nombre d'observations.

Les données suivantes proviennent d'une étude sur le tabac en feuilles et en cigarettes, en vue de relier ses constituants organiques et inorganiques (Waltz, Reid et Colwell, 1948).

TAB. 2.3.1 –. *Les données de l'exemple du Tabac*

Y1	Y2	Y3	X1	X2	X3	X4	X5	X6
1,55	20,05	1,38	2,02	2,90	2,17	0,51	3,47	0,91
1,63	12,58	2,64	2,62	2,78	1,72	0,50	4,57	1,25
1,66	18,56	1,56	2,08	2,68	2,40	0,43	3,52	0,82
1,52	18,56	2,22	2,20	3,17	2,06	0,52	3,69	0,97
1,70	14,02	2,85	2,38	2,52	2,18	0,42	4,01	1,12
1,68	15,64	1,24	2,03	2,56	2,57	0,44	2,79	0,82
1,78	14,52	2,86	2,87	2,67	2,64	0,50	3,92	1,06
1,57	18,52	2,18	1,88	2,58	2,22	0,49	3,58	1,01
1,60	17,84	1,65	1,93	2,26	2,15	0,56	3,57	0,92
1,52	13,38	3,28	2,57	1,74	1,64	0,51	4,38	1,22
1,68	17,55	1,56	1,95	2,15	2,48	0,48	3,28	0,81
1,74	17,97	2,00	2,03	2,00	2,38	0,50	3,31	,98
1,93	14,66	2,88	2,50	2,07	2,32	0,48	3,72	1,04
1,77	17,31	1,36	1,72	2,24	2,25	0,52	3,10	0,78
1,94	14,32	2,66	2,53	1,74	2,64	0,50	3,48	0,93
1,83	15,05	2,43	1,90	1,46	1,97	0,46	3,48	0,90
2,09	15,47	2,42	2,18	,74	2,46	0,48	3,16	0,86
1,72	16,85	2,16	2,16	2,84	2,36	0,49	3,68	0,95
1,49	17,42	2,12	2,14	3,30	2,04	0,48	3,28	1,06
1,52	18,55	1,87	1,98	2,90	2,16	0,48	3,56	0,84
1,64	18,74	2,10	1,89	2,82	2,04	0,53	3,56	1,02
1,40	14,79	2,21	2,07	2,79	2,15	0,52	3,49	1,04
1,78	18,86	2,00	2,08	3,14	2,60	0,50	3,30	0,8
1,93	15,62	2,26	2,21	2,81	2,18	0,44	4,16	0,92
1,53	18,56	2,14	2,00	3,16	2,22	0,51	3,73	1,07

Les variables à expliquer sont définies par:

Y1: taux de consumabilité de la cigarette en pouces par 1000 secondes;

Y2: pourcentage de sucre dans la feuille de tabac;

Y3: pourcentage de nicotine.

Les variables explicatives sont définies par:

X1: pourcentage d'azote;

X2: pourcentage de chlore;

X3: pourcentage de potassium;

X4: pourcentage de phosphore;

X5: pourcentage de calcium;

X6: pourcentage de magnésium.

Il s'agit d'expliquer les variables dépendantes Y_i avec $i = 1,2,3$ à partir des variables explicatives X_j avec $j = 1,2,3,4,5,6$.

On considère les données centrées.

2.3.2. La régression sur les composantes principales appliquée à l'exemple du Tabac

Nous allons traiter l'évolution de la solution de la régression sur les composantes principales en fonction du nombre de composantes. Une application de l'analyse en composantes principales sur l'exemple de Tabac a donné les résultats suivants:

Les valeurs propres de la matrice $(X'X)$ sont:

$\lambda_1 = 0{,}3748$

$\lambda_2 = 0{,}2192$

$\lambda_3 = 0{,}06854$

$\lambda_4 = 0{,}01458$

$\lambda_5 = 0{,}0035$

$\lambda_6 = 0{,}0010$

et la matrice des vecteurs propres correspondants est:

variables	t1	t2	t3	t4	t5	t6
x1	-,005285	0,431402	0,675296	-,519488	-,294579	0,034567
x2	0,979662	-,178433	0,075355	-,045624	-,025773	-,000753
x3	-,082871	-,321131	0,731189	0,506872	0,313762	0,002062
x4	0,004409	0,000626	-,027653	-,040653	0,125409	0,990876
x5	0,176285	0,794969	-,047036	0,577123	0,037384	0,016347
x6	0,047662	0,216686	-,026125	-,369334	0,892742	-,129219

Sur la base de cette matrice, on détermine les valeurs des composantes sur lesquelles on applique la procédure FORWARD multivariée (voir annexe D) (Lazraq et Cléroux, 1988b) afin d'obtenir les différents étapes, les composantes entrantes, les coefficients de redondance partiels, l'indice de redondance et la valeur p.

TAB. 2.3.2 –. *Les différentes étapes, les composantes entrantes, les coefficients de redondance partiels, ri, l'indice de redondance, RI, et la valeur p (PCR-Tabac)*

Étapes	composante entrante	ri	RI	valeur p
1	2	0,45	0,45	0
2	1	0,12	0,57	0,018
3	3	0,07	0,64	0,064
4	4	0,05	0,69	0,07
5	6	0,05	0,73	0,092
6	5	0,00	0,73	0,767

TAB. 2.3.3 –. *Les coefficients de régression selon les composantes principales(PCR-Tabac)*

variables	t1	t2	t3	t4	t5	t6
y1	,00582	-,17422	,26462	,40799	-,59091	-,38092
y2	-2,92115	1,18238	-1,9742	4,03398	14,14027	-,38077
y3	,90885	-,1445	,46340	-,68842	-,95082	1,62394

Les écarts types des coefficients de régression en fonction des deux premières composantes principales retenues sont donnés par le tableau suivant:

TAB. 2.3.4 –. *Les écarts types des coefficients de régression en fonction des 2 premières composantes principales retenues(PCR-Tabac)*

composantes retenues		t2	t1	SCE
1	y1	0,07434		0,66882
	y2	0,68373		56,57008
	y3	0,13625		2,24623
2	y1	0,05848	0,04472	0,39580
	y2	0,61652	0,47152	43,99519
	y3	0,13335	0,10199	2,05836

TAB. 2.3.5 –. *Les coefficients de corrélation simple entre les variables dépendantes et les composantes principales(PCR-Tabac)*

	t1	t2	t3	t4	t5	t6
$r_{y_1 t_h}$	-,63883	,01632	,41494	,29502	-,13414	-,10931
$r_{y_2 t_h}$	,35204	-,66518	-,25137	,23686	,01089	,21240
$r_{y_3 t_h}$	-,16882	,81194	,23148	-,15859	18218	-,05602

TAB. 2.3.6 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre de composantes retenues (PCR-Tabac)*

	1	2	3	4	5	6
$r_{y_1 \hat{y_1}}$	0,01632	0,63904	0,76193	0,81706	0,82434	0,83518
$r_{y_2 \hat{y_2}}$	-0,66518	-0,36892	-0,44630	-0,33067	-0,35592	-0,35305
$r_{y_3 \hat{y_3}}$	0,81194	0,18950	0,28500	0,20851	0,21410	0,18206

TAB. 2.3.7 –. *Les écarts types résiduels en fonction du nombre de composantes retenues(PCR-Tabac)*

	1	2	3	4	5	6
σ_e	0,8966	0,7923	0,7312	0,6744	0,6265	0,6239

TAB. 2.3.8 –. *L'indice de redondance entre les variables dépendantes et les erreurs selon le nombre de composantes retenues (PCR-Tabac)*

	1	2	3	4
$RI(Y,\hat{e})$	0,5471129	0,4272149	0,363941	0,3095239

On constate d'après les tableaux ci-dessus que la première composante explique 45% de la variabilité des variables dépendantes, et que les deux premières composantes expliquent 57% soit 78% de RI, avec un niveau de signification $\alpha = 5\%$ et à niveau de 10% cinq composantes sont significatives avec un apport de qualité d'ajustement évalué à 73%. On remarque aussi que le coefficient de redondance RI a un accroissement dégressif selon les composantes retenues.

Une révision des coefficients de corrélation entre les variables dépendantes et les composantes principales montre une forte corrélation négatives entre y_1 et t_1, y_2 et t_2, puis une forte corrélation positive et significative entre y_3 et t_2, et une corrélation positive entre y_2 et t_1, alors qu'il n'y a pas de corrélation entre y_1 et t_2.

On remarque aussi une incohérence de signes entre les coefficients de régression et les coefficients de corrélation pour toutes les composantes principales à l'exception de la troisième et la quatrième composantes. La somme des carrés des erreurs a baissé en tenant compte des deux premières composantes par rapport à une seule composante dans le modèle et les écarts types des coefficients de régression connaissent des petites fluctuations autour de la valeur zéro.

Les moyennes résiduelles sont quasi nulles et les écarts types résiduels baissent au fur et à mesure que les composantes augmentent dans le modèle.

Les corrélations entre les valeurs observées et les valeurs estimées des variables dépendantes sont adéquatement réparties en considérant les cinq premières composantes soient, 0,824 pour la première variable, -0,356 pour la deuxième et 0,214 pour la troisième. Alors une considération des deux premières composantes donne respectivement les corrélations 0,639,-0,369 et 0,186.

2.3.3. La régression par analyse des valeurs latentes appliquée à l'exemple du Tabac

Cette méthode consiste à introduire les variables dépendantes dans le calcul des valeurs latentes. On considère autant de matrices que de variable dépendantes telles que: $A_1 = [y_1|X]$, $A_2 = [y_2|X]$ et $A_3 = [y_3|X]$.

Pour l'exemple considéré, les valeurs propres calculées à partir de l'analyse en composantes principales appliquée sur les différentes matrices sont données par le tableau suivant:

TAB. 2.3.9 –. *Les différentes valeurs propres obtenues selon chaque variables dépendantes (LRR-Tabac)*

ordre	y1	y2	y3
1	0,38669917	4,38033519	0,46825822
2	0,21922376	0,34734358	0,36011896
3	0,07404447	0,11204295	0,07908671
4	0,01849117	0,05312976	0,03242133
5	0,00689531	0,01208223	0,01246419
6	0,00312920	0,00345532	0,00290188
7	0,00091125	0,00081431	0,00093738

et les coefficients des valeurs latentes correspondants sont:

variables	t1	t2	t3	t4	t5	t6	t7
y1	-,179573	0,005896	0,289825	0,536091	0,730632	0,238286	0,075649
y2	0,981952	-,059015	0,141994	0,101390	0,041683	-,001032	-,010343
y3	0,710186	0,244984	0,286893	-,551273	0,198965	-,097385	0,018403

En utilisant la règle empirique de Webster, Gunst et Mason, (1974), on choisit d'extraire les différentes valeurs latentes dont au moins un coefficient des vecteurs propres associés aux variables dépendantes est inférieur à 0,20 et sa valeur propre inférieure à 0,05 d'où la cinquième, la sixième et la septième valeurs latentes doivent être éliminées.

Après avoir calculé les valeurs latentes, on estime les différentes prédictions par la méthode de la régression par analyse des valeurs latentes en introduisant valeur par valeur puis on traite l'évolution de la solution de la régression en fonction du nombre de valeurs latentes, les coefficients de redondance entre les variables dépendantes et les résidus obtenus, les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées des variables dépendantes.

TAB. 2.3.10 –. *Les coefficients de corrélation simple entre les valeurs observées et les valeurs estimées selon le nombre de valeurs latentes retenues (LRR-Tabac)*

	1	2	3	4
$r_{y_1\hat{y_1}}$	-0,640866	-0,6409289	-0,4991155	0,1143894
$r_{y_2\hat{y_2}}$	-0,7286437	0,712799	0,7303025	0,7324123
$r_{y_3\hat{y_3}}$	-0,7154362	-0,8260512	-0,04197035	0,8294807

Les coefficients de corrélation simple entre les valeurs observées et les valeurs estimées pour chaque variable dépendante débutent par une forte corrélation négative et se terminent par des corrélations positives. En considérant le modèle avec les 4 valeurs latentes, on constate une forte corrélation positive pour les variables dépendantes y_2 et y_3.

TAB. 2.3.11 –. *Les écarts types résiduels en fonction du nombre de valeurs latentes retenues(LRR-Tabac)*

	1	2	3	4
σ_e	1,546181	1,577551	0,6655349	0,6533762

TAB. 2.3.12 –. *L'indice de redondance entre les variables dépendantes et les erreurs selon le nombre de composantes retenues (LRR-Tabac)*

	1	2	3	4
$RI(Y,\hat{e})$	0,2276706	0,308473	0,1376048	0,08197518

Une lecture des tableaux montre que la régression par analyse des valeurs latentes a donné des bons résultats pour les écarts types résiduels et les indices de redondance entre les variables dépendantes et les résidus. Le modèle avec les 4 valeurs latentes marque une valeur de 0,6533 pour l'écart-type résiduel et 0,0819 pour l'indice de redondance.

En effet, la régression par analyse des valeurs latentes s'avère compétitive en pratique par rapport à la régression sur les composantes principales.

2.3.4. La régression par moindres carrés partiels appliquée à l'exemple du Tabac

Nous avons utilisé la fonction PLS(2) programmée en Splus sur les données de Tabac pour obtenir les composantes PLS et on a appliqué la procédure FORWARD. Pour avoir certains résultats complémentaires à l'étude, on a utilisé la procédure régression en SAS.

Par ces algorithmes, les composantes PLS sont extraites une à une puis on exprime successivement les variables dépendantes en fonction d'une composante PLS, de deux composantes PLS,...etc jusqu'à un nombre q de composantes PLS.

Nous commentons les résultats obtenus.

TAB. 2.3.13 –. *Les différents composantes PLS obtenues à partir de la fonction PLS(2)*

	t1	t2	t3	t4	t5	t6
1	-0,4097	0,0586	-0,0690	0,0451	0,0169	-0,0482
2	0,6471	1,0897	-0,2060	0,0801	0,0036	-0,0154
3	-0,2520	-0,0305	0,0291	-0,1320	-0,0555	-0,0540
4	-0,3620	0,4622	-0,0363	0,0959	0,0269	-0,0830
5	0,3348	0,4355	0,0077	-0,0279	-0,0727	0,0838
6	-0,5879	-0,5114	0,3890	0,1707	-0,0484	-0,0103
7	0,3672	0,7519	0,6335	-0,0951	0,0302	0,0002
8	-0,2016	-0,1302	-0,2212	-0,0793	-0,0075	0,1161
9	0,0229	-0,2933	-0,2502	-0,0757	0,0688	-0,0089
10	1,2208	0,4017	-0,2690	0,1701	0,0134	-0,0346
11	-0,1005	-0,5360	0,0183	-0,1338	-0,0046	-0,0044
12	0,0972	-0,4939	0,0491	0,0100	0,0092	0,0963
13	0,5148	0,0963	0,2471	0,0428	-0,0035	0,0073
14	-0,3441	-0,7290	-0,1864	0,0012	0,0273	-0,0497
15	0,5669	-0,2359	0,4666	-0,0702	0,0278	-0,0074
16	0,5148	-0,7727	-0,3484	0,0705	-0,0375	-0,0361
17	0,9026	-1,1763	0,1723	0,0225	-0,0007	-0,0028
18	-0,2101	0,2254	0,0307	-0,1221	0,0031	0,0172
19	-0,6661	0,3108	0,1297	0,3848	-0,0235	0,0019
20	-0,3951	0,0602	-0,1573	-0,0425	-0,0116	-0,0938
21	-0,3416	0,0099	-0,2616	0,0557	0,0291	0,0520
22	-0,2760	0,0798	-0,0302	0,1171	0,0230	0,0503
23	-0,6988	0,0795	0,2301	-0,1384	0,0193	-0,0439
24	0,0961	0,4949	-0,2249	-0,2746	-0,0463	-0,0534
25	-0,4397	0,3530	-0,1427	-0,0749	0,0129	0,1210

TAB. 2.3.14 –. *Les différentes étapes, les composantes entrantes, l'indice de redondance, RI, et la valeur p(PLS-Tabac)*

Étapes	composante entrante	ri	RI	valeur p
1	1	0,5684209	0,5684209	4,887581e-06
2	2	0,05777757	0,6261984	0,2341253
3	3	0,03974786	0,6659463	0,3292297
4	4	0,03134326	0,6972896	0,3900689
5	5	0,03505641	0,732346	0,3614933
6	6	0,002736064	0,735082	0,8613113

TAB. 2.3.15 –. *Les coefficients de corrélation simple entre les variables dépendantes et les différents composantes PLS(PLS-Tabac)*

	t1	t2	t3	t4	t5	t6
$r_{y_1 t_h}$	0,46163	-0,39485	0,37261	-0,39214	-0,15428	-0,10946
$r_{y_2 t_h}$	-0,75619	-0,21980	-0,20229	-0,17873	0,19338	-0,02390
$r_{y_3 t_h}$	0,74330	0,43996	0,10803	0,09944	-0,01225	0,18731

Seule la première composante PLS est significative au niveau 5% et elle explique 56,8% de la variabilité des variables à expliquer. En effet, une lecture du tableau des coefficients de corrélation entre les variables dépendantes et les différents composantes PLS montre que la corrélation de la première composante PLS avec les variables dépendantes est de l'ordre successivement de 0,462, -0,743 et 0,743. On remarque qu'elle est plus corrélée aux valeurs observées des variables à expliquer que toutes autres composantes PLS.

TAB. 2.3.16 –. *Les coefficients de régression en fonction des composantes PLS (PLS-Tabac)*

variables	t1	t2	t3	t4	t5	t6
y1	0,15119	-0,12841	0,24976	-0,49098	-0,79270	-0,31410
y2	-3,05000	-0,87992	-1,66963	-2,75340	12,20961	-0,84271
y3	0,76417	0,44901	0,22727	0,39056	-0,19381	1,68783

TAB. 2.3.17 –. *Les écarts types des coefficients de régression et la somme des carrés des erreurs en fonction des quatre premières composantes PLS retenues(PLS-Tabac)*

		t1	t2	t3	t4	SCE
1	y1	0,06058				0,52644
	y2	0,55033				43,44508
	y3	0,14340				2,94996
2	y1	0,05547	0,05507			0,42213
	y2	0,53001	0,52618			38,54423
	y3	0,11046	0,10966			1,67416
3	y1	0,05014	0,04978	0,10261		0,32924
	y2	0,51243	0,50873	1,04875		34,39294
	y3	0,11043	0,10963	0,22601		1,59727
4	y1	0,04260	0,04229	0,08718	0,16280	0,22633
	y2	0,49974	0,49614	1,02278	1,90995	31,15289
	y3	0,11083	0,11003	0,22682	0,42356	1,53210

D'après ces deux tableaux on constate que les écarts types des coefficients de régression sont minimes et que les signes des coefficients de régression et des coefficients de corrélation sont cohérents, abstraction faite de la signification des composantes PLS.

TAB. 2.3.18 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre des composantes PLS retenues (PLS-Tabac)*

	1	2	3	4	5	6
$r_{y_1\hat{y}_1}$	0,46163	0,59324	0,71265	0,81344	0,82797	0,83517
$r_{y_2\hat{y}_2}$	-0,75619	-0,71547	-0,47382	-0,32895	-0,35922	-0,35300
$r_{y_3\hat{y}_3}$	0,74330	0,64625	0,29420	0,20981	0,20838	0,18204

TAB. 2.3.19 –. *Les moyennes et les écarts types résiduels en fonction des composantes PLS retenues(PLS-Tabac)*

	1	2	3	4	5	6
σ_e	0,7962876	0,7587324	0,7005736	0,6668940	0,6271584	0,6239511

TAB. 2.3.20 –. *L'indice de redondance entre les variables dépendantes et les erreurs selon le nombre de composantes retenues (PLS-Tabac)*

	1	2	3	4
$RI(Y,\hat{e})$	0,431558	0,391812	0,3340478	0,3027012

Les écarts types résiduels ont une tendance à la baisse avec la présence de composantes PLS dans le modèle et se situant autour de zéro.

La structure des coefficients de corrélation entre les valeurs observées et les valeurs estimées selon les composantes PLS retenues montre que l'optimum de cette corrélation correspond à la première composante PLS.

2.4. Conclusion

Dans les paragraphes précédents, nous avons traité deux exemples de la régression linéaire où le nombre d'observations est supérieur au nombre de variables explicatives et la matrice $(X'X)$ est presque singulière.

Examinons les différentes représentations graphiques obtenues selon les deux exemples. D'après l'exemple de Longley, les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendante se rapprochent, en considérant les modèles optimaux de chaque méthode avec une supériorité du côté de la régression par analyse des valeurs latentes.

Pour l'exemple Tabac, la régression par analyse des valeurs latentes présente une supériorité par rapport aux deux autres méthodes due aux valeurs minimes des écarts types résiduels et des indices de redondance entre les variables dépendantes et les résidus obtenus à partir des différents modèles considérés.

Par ailleurs, pour d'autres critères de comparaison la stabilité des modèles, la signification des composantes, le nombre de composantes dans le modèle, les corrélations entre les valeurs observées et les valeurs estimées des variables à expliquer, l'avantage est donné à la méthode par moindres carrés partiels.

Bref, la compétition entre les trois méthodes est très étroite.

Mais, pour appuyer notre comparaison nous traiterons deux autres exemples où le nombre d'observations est inférieur au nombre des variables explicatives et en considérant, deux jeux de données: un jeu de calibration et un jeu de validation.

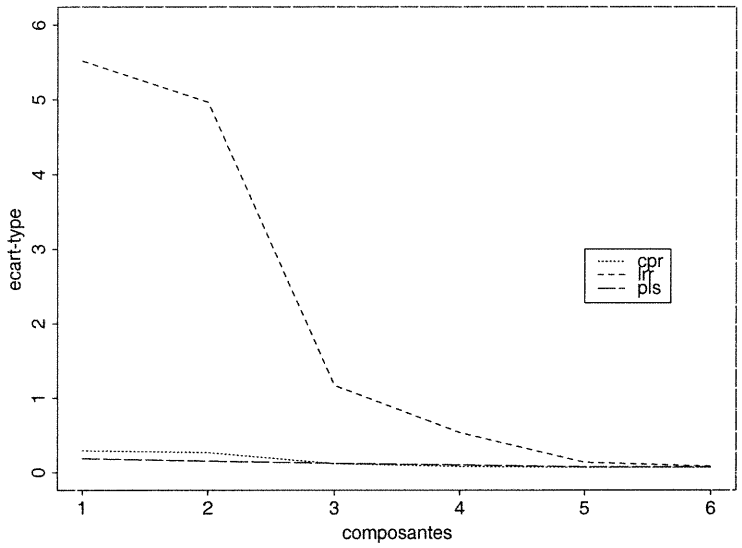

FIG. 2.4.1 –. *Les écarts types résiduels en fonction des composantes selon les trois méthodes (l'exemple de Longley)*

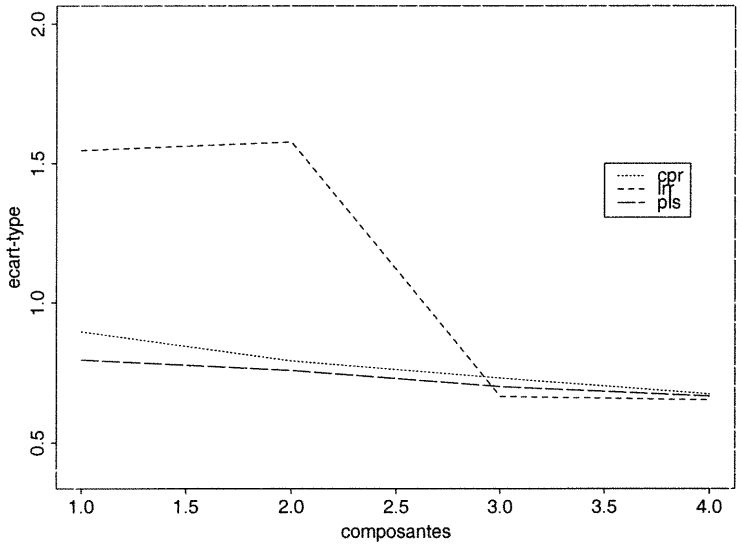

FIG. 2.4.2 –. *Les écarts types résiduels en fonction des composantes selon les trois méthodes (l'exemple du Tabac)*

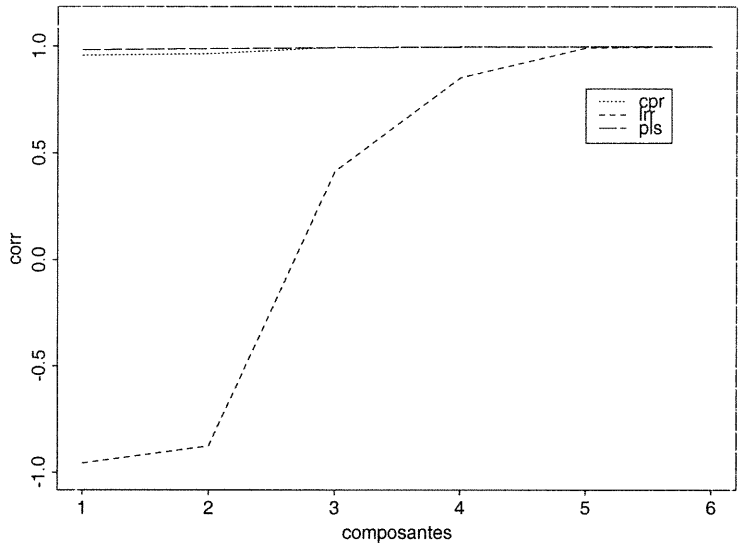

FIG. 2.4.3 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées en fonction des composantes selon les trois méthodes (l'exemple de Longley)*

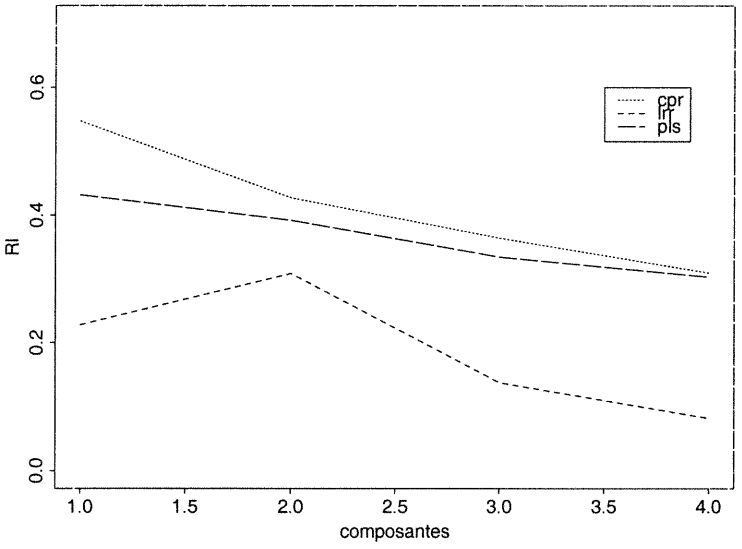

FIG. 2.4.4 –. *L'indice de redondance entre les variables dépendantes et les erreurs en fonction des composantes selon les trois méthodes (l'exemple du Tabac)*

Chapitre 3

RÉGRESSIONS MULTIPLE ET MULTIVARIÉE: CAS PRATIQUES OÙ N<Q

3.1. INTRODUCTION

Dans ce chapitre, on illustre les résultats des trois méthodes à travers deux exemples où le nombre de variables explicatives est largement supérieur au nombre d'observations. Il s'agit de l'exemple de l'indice d'octane dans le cas multiple où on a une seule variable dépendante et l'exemple de Gauchi qui traite un problème de régression multivariée.

La stratégie consiste:

- à prolonger notre comparaison sur l'efficacité des différentes méthodes pour prédire la variable ou les variables dépendantes sur des données supplémentaires. - à déterminer les coefficients de redondance à partir des données de calibration. - à déterminer les écarts types des erreurs à partir des données de validation selon les différents prédicteurs pris dans un ordre de sélection successive par la procédure FORWARD sur les données de calibration pour la méthode de régression sur les composantes principales et la régression par moindres carrés partiels et d'autre part, par la sélection des prédicteurs en se basant sur la règle de Webster, Gunst et Mason (1974) pour la méthode de régression par analyse des valeurs latentes.

3.2. La régression multiple: traitement de l'exemple de l'indice d'octane

3.2.1. Présentation des données de l'indice d'octane

Cet exemple a été élaboré par CAMO AS et a été présenté et développé dans Esbensen, Schönkopf et Midtgard (1994) pages 146 à 153: le cas UOP Guided Wave Inc. Il est particulièrement illustratif de l'utilisation de la régression PLS sur les données de spectroscopie (Tenenhaus, 1998).

Un extrait de données est présenté dans les tableaux suivants:

TAB. 3.2.1 –. *Un extrait de données de calibration -indice d'octane-*

y	X1	X2	X3
88,60	-1,835E-03	-1,496E-03	-1,009E-03
88,80	-1,627E-03	-1,261E-03	-7,482E-04
89,40	-4,649E-04	-3,300E-04	1,807E-04
86,70	-1,443E-03	-1,094E-03	-5,958E-04
91,20	-1,607E-03	-1,155E-03	-5,484E-04
91,30	-5,020E-04	-5,397E-06	6,569E-04
87,40	-1,131E-03	-8,011E-04	-3,231E-04
87,10	-1,969E-03	-1,627E-03	-1,136E-03
87,00	-1,294E-03	-9,517E-04	-4,732E-04
91,80	-9,838E-04	-5,423E-04	4,661E-05
89,10	-1,228E-03	-8,495E-04	-3,259E-04
91,80	-1,518E-03	-1,126E-03	-5,891E-04
86,90	-1,671E-03	-1,302E-03	-7,786E-04
91,70	-1,554E-03	-1,022E-03	-3,313E-04
91,70	-1,581E-03	-1,174E-03	-6,173E-04
87,00	-1,292E-03	-9,211E-04	-4,050E-04
87,00	-1,644E-03	-1,296E-03	-8,091E-04
90,80	-3,835E-04	2,129E-05	5,825E-04
87,20	-8,814E-04	-4,975E-04	3,617E-05
91,40	-1,499E-03	-1,074E-03	-4,891E-04
87,20	-1,506E-03	-1,169E-03	-6,710E-04
92,20	-1,522E-03	-1,076E-03	-4,808E-04
91,80	-1,741E-03	-1,275E-03	-6,383E-04
87,00	-1,410E-03	-1,082E-03	-6,101E-04
89,00	-3,880E-03	-3,539E-03	-3,039E-03
92,40	-4,469E-03	-4,083E-03	-3,547E-03

TAB. 3.2.2 –. *Un extrait de données de validation -indice d'octane-*

y	X1	X2	X3
88,60	1,537E-03	1,848E-03	2,295E-03
88,80	-8,178E-04	-4,507E-04	6,246E-05
91,20	-1,983E-03	-1,575E-03	-1,018E-03
91,80	-1,602E-03	-1,158E-03	-5,636E-04
89,00	-1,928E-03	-1,608E-03	-1,142E-03
91,40	-1,835E-03	-1,458E-03	-9,300E-04
88,60	-1,011E-03	-5,896E-04	-1,142E-05
91,40	-8,832E-04	-3,842E-04	2,836E-04
87,10	-1,461E-03	-1,068E-03	-5,326E-04
91,40	-3,937E-03	-3,527E-03	-2,971E-03
90,30	-4,330E-03	-3,950E-03	-3,428E-03
91,20	-3,821E-03	-3,414E-03	-2,857E-03
91,00	-3,917E-03	-3,515E-03	-2,957E-03

Il s'agit de relier la variable à expliquer y représentant l'indice d'octane par $226 X_j$ variables explicatives qui représentent des valeurs d'absorbance à différentes longueur d'onde.

Cependant, on construit un modèle à partir des données de calibration puis, on teste sa performance sur les données de validation selon les trois méthodes prescrites.

3.2.2. La régression sur les composantes principales appliquée à l'exemple de l'indice d'octane

Une application de la procédure de l'analyse en composantes principales du programme SAS sur l'exemple de l'indice d'octane a révélé les résultats suivants : 25 valeurs propres prises dans un ordre croissant de la matrice $(X'X)$ non nulles dont on présente les cinq premières valeurs:

$\lambda_1 = 4{,}5063$

$\lambda_2 = 0{,}0895$

$\lambda_3 = 0{,}00277$

$\lambda_4 = 0{,}00077$

$\lambda_5 = 0{,}00015$

et un extrait de 4 composantes principales prises parmi les 25 calculées de la matrice des vecteurs propres selon des données de calibration et le résultat d'une validation de la procédure FORWARD au niveau de signification de 10%.

Un ensemble de tableaux a été élaboré à partir des modèles considérant ces quatre composantes principales.

TAB. 3.2.3 –. *Les 4 composantes principales t1,t6,t10 et t12 obtenues à partir des données de calibration(PCR-octane)*

t1	t6	t10	t12
-0,43756	0,007784	-0,003328	0,004143
-0,79821	-0,000633	0,014944	0,011295
0,03268	0,015322	0,006059	0,014730
-2,56830	0,000479	-0,001575	-0,004348
1,94426	0,015470	-0,007666	0,000883
1,80542	-0,017162	-0,000742	0,007583
-1,41663	-0,003091	-0,012349	0,013653
-2,58884	-0,009472	0,002015	-0,007235
-2,94873	-0,005454	0,000883	-0,008253
2,56328	-0,018804	0,003919	0,001729
-0,06800	0,009322	-0,008288	0,000669
2,16437	0,000231	-0,001638	-0,009988
-1,90743	0,009650	-0,018859	0,001679
2,00507	-0,010771	0,002577	0,000229
2,27442	-0,002363	-0,001284	0,000191
-2,49866	-0,013361	0,001042	-0,008333
-2,49858	0,000716	0,000454	-0,009259
1,59182	0,014374	-0,003162	-0,002044
-2,44805	-0,010111	-0,005012	0,000555
2,22295	0,002812	-0,008428	-0,005892
-2,16659	0,001000	0,012101	0,004969
2,67324	-0,018087	0,001153	0,002161
2,32579	0,019321	0,014326	-0,013763
-1,97883	0,012368	0,012571	0,004735
-0,35325	0,001043	0,001001	0,000005
3,07436	-0,000583	-0,000716	-0,000094

TAB. 3.2.4 –. *Les coefficients de corrélation simple de y et des composantes, r_{yt_h}, les coefficients de détermination multiple en fonction du nombre de composantes principales retenues, R^2, et les coefficients de régression P_h (PCR-octane-calibration)*

coefficients	t1	t6	t10	t12
r_{yt_h}	0,99182	-0,05588	0,05491	-0,05129
R^2	0,9837	0,9868	0,9898	0,9925
P_h	0,99893	-10,96887	14,85669	-15,27901

TAB. 3.2.5 –. *Les écarts types des estimateurs des coefficients de régression et la somme des carrés des erreurs(SCE) en fonction du nombre de composantes principales retenues (PCR-octane-calibration)*

composante	t1	t6	t10	t12	SCE
1	0,02625				1,86238
2	0,02411	4,69815			1,50717
3	0,02164	4,21848	5,81477		1,16185
4	0,01907	3,71715	5,12373	5,64172	0,86088

TAB. 3.2.6 –. *Les moyennes, les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendantes en fonction du nombre de composantes retenues (PCR-octane-validation)*

	1	2	3	4
moyennes	7,651999E-15	8,745141E-15	8,745141E-15	1,749028E-14
σ_e	0,2392876	0,2377764	0,2314994	0,2314173
$corr(y,\hat{y})$	0,9875	0,9876	0,9883	0,9883

L'analyse des tableaux montre que les coefficients de régression ont le même signe que les coefficients de corrélation ce qui donne une possibilité d'explication de l'indice d'octane par les composantes principales retenues.

On remarque aussi que la part de la variabilité de la première composante représente 98,37% ce qui implique que la quasi totalité de l'information qui explique l'indice d'octane est incorporée dans la première composante. L'apport partiel rapporté par les autres composantes ne représente respectivement que 0,31%, 0,30% et 0,27%.

Aussi, la somme des carrés des erreurs (SCE) connaît une légère différence entre les différentes composantes présentes dans le modèle.

Après avoir constitué nos modèles à partir des données de calibration, une projection a été faite sur les données de validation, qui a révélé des moyennes et des écarts types résiduels se situant autour de la valeur zéro et une forte corrélation entre les valeurs observées et les valeurs estimées de la variable dépendante (voir tableau 3.2.6).

Il ressort de notre analyse que la considération d'une seule composante donne des qualités d'ajustement très satisfaisantes pour prédire la variable dépendante tout en préservant les variables explicatives et en réduisant la variance des coefficients de régression.

3.2.3. La régression par analyse des valeurs latentes appliquée à l'exemple de l'indice d'octane

Une application du programme SAS de l'analyse en composantes principales sur la matrice [y|X]=A a donné vingt cinq valeurs propres non nulles selon un ordre croissant de grandeur dont voici les cinq premières valeurs:

$\lambda_1 = 9{,}04029469$

$\lambda_2 = 0{,}08953284$

$\lambda_3 = 0{,}03717159$

$\lambda_4 = 0{,}00277449$

$\lambda_5 = 0{,}00077129$

et un extrait de la matrice des vecteurs propres correspondant est donnée par:

variables	t1	t2	t3	t4	t5
y	0,709650	-,016970	0,703699	-,000482	-,002346
x1	-,000037	-,002462	-,000094	0,002411	0,006995
x2	-,000023	-,002514	-,000041	0,002068	0,006779
x3	-,000008	-,002573	0,000007	0,001791	0,006621
x4	0,000008	-,002640	0,000062	0,001592	0,006561
x5	0,000026	-,002717	0,000136	0,001434	0,006569
x6	0,000046	-,002805	0,000236	0,001226	0,006547
x7	0,000070	-,002907	0,000353	0,000832	0,006342

Selon les critères de choix des valeurs latentes proposé par la règle Webster, Gunst et Mason (1974), seulement trois valeurs sont prises en considération.

L'analyse de ces valeurs latentes par la fonction (LRR) programmée en Splus a donné des résultats sur les valeurs estimées de la variable dépendante y, dont on calcule les différents écarts types résiduels et les corrélations. Par la suite, une projection des modèles élaborés à partir des données de calibration sur les données de validation et selon l'incorporation étape par étape des valeurs latentes choisies dans le modèle on obtient les différents moyennes, écarts types résiduels et corrélations afin, d'une part, de tester la stabilité des modèles, et d'autre part, de faire la comparaison avec les autres méthodes.

TAB. 3.2.7 –. *Les différentes valeurs latentes obtenues (LRR-octane)*

	1	2	3
1	-0,85143	-0,05961	-0,23063
2	-0,96360	-0,04251	0,16713
3	0,04760	-0,04318	0,00024
4	-3,70100	0,02938	-0,05442
5	2,67179	-0,15950	-0,09148
6	2,64495	-0,18869	0,07761
7	-2,39282	-0,06235	-0,38131
8	-3,43161	0,00200	0,24129
9	-3,75613	0,02220	0,42702
10	3,53372	-0,15820	-0,10739
11	-0,23623	-0,07252	-0,14098
12	3,25267	-0,18338	0,17499
13	-3,09345	-0,04508	-0,38501
14	3,06947	-0,18821	0,21773
15	3,25924	-0,19025	0,02635
16	-3,43903	0,00682	0,10719
17	-3,43899	0,02314	0,10733
18	2,13961	-0,08115	-0,12082
19	-3,26145	0,00832	0,21220
20	3,01007	-0,11049	-0,14673
21	-3,06315	-0,01336	0,01163
22	3,89505	-0,14483	0,09680
23	3,36640	-0,22333	0,05906
24	-3,07279	-0,00879	-0,26280
25	-0,50827	0,65651	0,00881
26	4,31939	1,22706	-0,01380

TAB. 3.2.8 –. *Les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendantes en fonction du nombre de valeur latentes retenues (LRR-octane-calibration)*

	1	2	3
σ_e	4,154611	4,039619	0,2680672
$corr(y,\hat{y})$	-0,9918182	-0,8836496	0,9917917

TAB. 3.2.9 –. *Les moyennes, les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendantes en fonction du nombre de valeur latentes retenues(LRR-octane-validation)*

	1	2	3
moyennes	3,846153e-08	3,846153e-08	3,846155e-08
σ_e	2,932976	2,705474	0,2337803
$corr(y,\hat{y})$	-0,9874814	-0,6088471	0,9876984

Ces tableaux donnent les moyennes et les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendante pour un nombre croissant de valeurs latentes dans le modèle. On constate que, pour les deux jeux de données le comportement des écarts types résiduels et des corrélations reste stable. Si on prend en considération la première valeur latente, on trouve une forte corrélation négative entre les valeurs estimées et les valeurs observées de la variable dépendante et un écart-type résiduel relativement grand. De même pour le deuxième modèle, on a une corrélation négative inférieure à la première mais avec un écart-type résiduel légèrement petit. En effet, le troisième modèle donne des résultats concurrentiels à ceux de la régression sur les composantes principales. Les moyennes résiduelles sont restées autour de zéro ce qui est souhaitable.

Il ressort de notre analyse que la considération de toutes les valeurs latentes choisies dans le modèle donne des résultats intéressants.

3.2.4. La régression par moindres carrés partiels appliquée à l'exemple de l'indice d'octane

Nous présentons dans cette section l'évolution des solutions de la régression PLS en fonction des composantes retenus. L'utilisation de l'algorithme de la fonction PLS(1) (voir Annexe B) sur les données de calibration sans tenir compte de l'inférence sur les composantes a donné une réduction de variables de 226 à 24 composantes PLS et sur la base de la validation par la méthode FORWARD, nous indiquons à un niveau de signification de 10% les tableaux ci-dessus des composantes choisies, les coefficients de corrélation simple entre y et les composantes, r_{yt_h}, les coefficients de détermination et les coefficients de régression.

TAB. 3.2.10 –. *Les valeurs des composantes obtenus selon les données de calibration*

i	t1	t3	t6	t4
1	-1,9696	-,000081997	0,000006848	-,000038470
2	-3,5930	0,000020708	0,000013513	-,000032803
3	0,1472	0,000012081	-,000002536	-,000055383
4	-11,5612	0,000016012	-,000007903	-,000021526
5	8,7523	-,000068985	0,000006366	0,000024898
6	8,1274	0,000059131	-,000017592	0,000085338
7	-6,3768	-,000089303	-,000019550	0,000039467
8	-11,6535	0,000038851	0,000006836	-,000002527
9	-13,2736	0,000095680	0,000013599	-,000002378
10	11,5388	0,000014940	-,000019238	-,000045985
11	-0,3060	-,000082020	0,000009005	-,000050951
12	9,7432	0,000026382	0,000014917	0,000038571
13	-8,5862	-,000109498	-,000007809	0,000023521
14	9,0261	0,000066284	-,000003961	0,000077750
15	10,2386	0,000023701	0,000005699	0,000035620
16	-11,2476	0,000060048	-,000004609	-,000005433
17	-11,2473	0,000031218	0,000002802	-,000029235
18	7,1657	-,000008081	-,000003852	-,000127056
19	-11,0198	0,000066286	-,000008635	0,000012636
20	10,0067	-,000015482	-,000002087	-,000086457
21	-9,7528	0,000006782	0,000005084	0,000057211
22	12,0338	0,000042886	-,000006502	-,000062469
23	10,4698	-,000040305	0,000019099	0,000125832
24	-8,9076	-,000069936	0,000004928	-,000049728
25	-1,5913	-,000043351	-,000009691	0,000132580
26	13,8369	0,000027967	0,000005269	-,000043023

TAB. 3.2.11 –. *Les coefficients de corrélation simple entre y et les composantes,r_{yt_h}, les coefficients de détermination en fonction du nombre de composantes retenus, R^2,les coefficients de régression, P_h (PLS-octane-calibration)*

r_{yt_h}	0,99182	0,11043	0,0447	0,0262
R^2	0,9837	0,9959	0,9979	0,9986
P_h	0,22191	4189,81	913,791	9117,61

TAB. 3.2.12 –. *Les écarts types des estimateurs des coefficients de régression et la somme des carrés des erreurs en fonction du nombre de composantes PLS retenues(PLS-octane-calibration)*

composantes	t1	t3	t6	t4	SCE
1	0,00583				1,86234
2	0,00299	506,73215			0,46882
3	0,00219	371,07582	1994,84169		0,24048
4	0,00181	307,51665	1653,15821	275,09423	0,15764

Une simple lecture des tableaux ci-dessous montre une cohérence des signes entre les coefficients de régression et les coefficients de corrélation, ce qui donne un avantage pour l'explication des modèles. On constate que la variance des coefficients de régression est stable et que la part de la variance expliquée par la première composante s'élève à 98.37% et des deux premiers composantes est de 99,59%.

Les coefficients de régression permettent le passage pour la construction des composantes t_h en fonction des x_j, puis de prédire la variable dépendante en fonction des t_h de même que le poids calculé d'une variable x_{hj} dans la composante t_h à partir des données de calibration permet d'obtenir les composantes du jeu de validation.

L'incorporation de ces composantes étape par étape dans le modèle selon un ordre de signification croissant nous fournit des résultats sur les moyennes et les

écarts types résiduels.

TAB. 3.2.13 –. *Les moyennes, les écarts types résiduels et les corrélations entre les valeurs observées et les valeurs estimées de la variable dépendante en fonction du nombre de composantes retenus (PLS-octane-validation)*

	1	2	3	4
moyennes	1.093E-15	5,466E-15	-2,186E-15	1,093E-15
σ_e	0,2392397	0,2289259	0,2276878	0,2264736
$corr(y,\hat{y})$	0,98748	0,98855	0,98867	0,98879

On remarque que les écarts types résiduels diminuent avec le nombre de composantes retenues dans le modèle. Une considération d'une seule composante donne des résultats très satisfaisants pour prédire la variable dépendante. En guise de conclusion, on peut considérer que les résultats des écarts types résiduels et des corrélations des modèles optimums des trois méthodes coïncident. La régression sur les composantes principales et la régression par moindres carrés partiels par une seule composante et la régression par analyse des valeurs latentes par trois valeurs latentes considérées.

Une conclusion et des représentations graphiques sont présentées à la fin du chapitre donnent une idée sur les résultats obtenus.

3.3. La régression multivariée: traitement de l'exemple de Gauchi

3.3.1. Présentation des données de Gauchi

Nous tenons à remercier J.P.GAUCHI pour nous avoir autorisé à utiliser cet exemple illustratif de la régression multivariée.

Pour rendre le fichier conforme à notre problème on fixe le nombre d'observations de telle sorte qu'il soit inférieur au nombre de variables explicatives. En revanche, on procède par une méthode de sondage pour le choix de notre fichier à traiter.

Le fichier principal contient 1008 observations dont des valeurs manquantes qu'on élimine. Après, on procède par un tirage aléatoire systématique en se basant sur une table des nombres au hasard pour tirer 25 observations. Ces derniers sont divisés en deux jeux de données: jeu de calibration qui contient 20 observations et jeu de validation avec 5 observations dont voici la présentation d'un extrait.

TAB. 3.3.1 –. *Extrait des données de Gauchi (calibration)*

y1	y2	y3	x1	x2	x3
6,75243	6,72800	4,76565	742,89522	11,73088	19,61087
6,76513	6,82485	4,70715	741,73399	10,27837	19,59923
6,76600	6,76058	4,68892	740,84031	14,57494	19,52608
6,75469	6,73143	4,66440	741,25674	17,82347	19,63210
6,74669	6,72530	4,59742	743,36880	11,95619	19,69824
6,75877	6,73219	4,66114	741,30660	7,98768	19,62354
6,74711	6,72772	4,64444	740,20634	9,97555	19,52026
6,74741	6,84797	4,48766	738,25015	20,58462	19,16288
6,73915	6,86881	4,46422	740,18581	16,05055	19,16207
6,73367	6,93658	4,54218	740,59789	12,87886	19,19089
6,74977	6,82958	4,61528	740,27436	12,35196	19,16617
6,73761	6,92994	4,53203	739,17693	13,56752	19,32780
6,73632	6,87219	4,51581	735,82246	17,34203	19,22112
6,75180	6,90827	4,62534	737,20573	14,36955	19,26591
6,75307	6,92225	4,57433	736,83732	15,76489	19,28052
6,75696	6,93636	4,59798	735,40610	17,20257	19,26710
6,76671	6,95179	4,69323	735,92524	18,06360	19,31378
6,77244	6,96163	4,68689	736,91046	18,38070	19,70008
6,76344	6,95398	4,60686	740,36006	15,51464	19,66435
6,79962	6,99867	4,58210	743,80339	14,16342	19,66533

TAB. 3.3.2 –. *Extrait des données de Gauchi (validation)*

y1	y2	y3	x1	x2	x3
6,81596	7,01954	4,59859	745,80377	14,58131	19,62898
6,81096	7,20492	4,63136	745,66215	18,93670	19,59578
6,79610	7,00514	4,61031	746,32131	19,21152	19,68300
6,79326	6,98884	4,64761	747,85355	11,95202	19,61238
6,78065	6,97886	4,71571	746,23606	11,30239	19,52740

Cet exemple est relatif à la synthèse gaz-gaz d'un composé gazeux X. Il est fabriqué à partir de plusieurs étapes complexes, dans deux réacteurs en parallèle. Le nombre de variables explicatives s'élève à 27 et les variables dépendantes à 16.

Nous abordons dans cette section à travers l'exemple une situation de deux groupes de variables: l'ensemble Y qui représente le groupe des variables dépendantes et X celui des variables indépendantes. Nous présentons l'illustration des trois méthodes prescrites.

On considère les données sont centrées-réduites.

3.3.2. La régression sur les composantes principales appliquée à l'exemple de Gauchi

Les résultats pratiques obtenus par l'analyse en composantes principales sur les variables indépendantes donne 18 valeurs propres, dont les sept premières valeurs prises par ordre croissant sont:

$\lambda_1 = 9{,}92079268$

$\lambda_2 = 7{,}00081991$

$\lambda_3 = 3{,}09771174$

$\lambda_4 = 1{,}78321711$

$\lambda_5 = 1{,}54685225$

$\lambda_6 = 1{,}01655842$

$\lambda_7 = 0{,}69599656$

et un extrait de composantes principales obtenues à partir des vecteurs propres associés aux valeurs propres est présenté dans le tableau suivant:

TAB. 3.3.3 –. *Un extrait des composantes principales obtenues par les vecteurs propres (calibration)*

observations	t1	t2	t3	t4	t5	t6	t7
1	3,73556	-1,58340	-3,90219	0,98579	0,93881	0,17528	-1,71902
2	4,14929	-1,80537	-2,49496	1,31880	0,00569	-1,18384	0,14501
3	3,17389	-2,11286	-1,27638	1,03028	-1,58995	0,36768	2,21692
4	2,28337	-1,19519	0,03594	-0,86045	1,02385	2,00376	1,03501
5	2,76377	-1,67399	1,08148	-0,86630	-1,03468	0,89500	-0,10025
6	2,99943	-2,60524	1,55035	-3,15048	0,33875	-0,97184	-0,74931
7	1,98718	-2,26306	1,51090	-1,45645	-0,52713	-1,10282	0,08532
8	-4,53184	-2,35188	1,08792	1,20426	1,65444	0,55142	0,47033
9	-4,49721	-2,13317	1,12433	1,81436	-0,68309	0,38929	-0,89911
10	-3,01250	-2,21315	0,24245	0,98000	0,63553	-1,05668	-0,12579
11	-2,77437	-2,55748	1,41900	0,60782	-0,52190	-1,04898	0,31112
12	-2,29518	-1,85809	-0,41452	-0,72374	1,44523	1,44910	-0,64582
13	-3,04595	1,47927	-0,90074	-0,90402	-1,78146	1,65330	-0,09611
14	-2,46750	2,04674	-0,86266	0,04581	-0,19013	-1,55849	0,69998
15	-2,33235	2,68501	-1,39127	-1,40298	-0,59262	-0,50213	-0,27570
16	-2,32079	2,91431	-1,57641	-1,33675	-0,79419	0,14145	0,01853
17	-2,55331	3,60797	-1,05545	-0,57704	0,22919	-0,60909	0,25809
18	2,36590	4,71174	0,78730	0,28060	3,31000	-0,09556	0,70752
19	3,32032	3,96162	0,96166	1,98345	-1,46171	0,42022	-0,8777
20	3,05230	2,94624	4,07326	1,02703	-0,40463	0,08294	-0,45900

Afin de calculer l'apport de chaque composante dans l'explication des variables dépendantes nous avons utilisé l'algorithme pas à pas de sélection de variables en régression linéaire multivariée introduit par Lazraq et Cléroux (voir Annexe D).

Au niveau de signification de 5% on détermine les différentes étapes, les composantes entrantes, les coefficients de redondance partiels, ri, l'indice de redondance RI et la valeur p obtenus par la fonction FORWARD.

TAB. 3.3.4 –. *Les différentes étapes, les composantes entrantes, les coefficients de redondance partiels, ri, l'indice de redondance, RI, et la valeur p (PCR-Gauchi-calibration)*

Étapes	composante entrante	ri	RI	valeur p
1	2	0,368	0,368	0,001
2	1	0,314	0,682	0
3	4	0,117	0,799	0,002
4	9	0,035	0,834	0,043

TAB. 3.3.5 –. *Les coefficients de corrélation simple entre les variables dépendantes et les composantes principales choisies(PCR-Gauchi-calibration)*

	t2	t1	t4	t9
$r_{y_1t_h}$	0,50684	0,55724	0,14963	0,08671
$r_{y_2t_h}$	0,71524	-0,40818	0,27566	-0,16029
$r_{y_3t_h}$	0,09239	0,71297	-0,10631	-0,01465
$r_{y_4t_h}$	0,14293	0,90221	-0,02249	0,04362
$r_{y_5t_h}$	-0,36704	0,72384	-0,43357	0,22203
$r_{y_6t_h}$	-0,37300	0,47281	-0,48167	0,31255
$r_{y_7t_h}$	-0,23561	0,65844	-0,41218	0,27579
$r_{y_8t_h}$	-0,06653	-0,24181	0,27168	-0,05223
$r_{y_9t_h}$	0,04550	0,15695	0,40923	0,21778
$r_{y_{10}t_h}$	-0,15756	0,51457	0,33381	-0,32643
$r_{y_{11}t_h}$	-0,69625	0,37218	-0,32913	0,08362
$r_{y_{12}t_h}$	-0,77852	0,38631	-0,31457	0,03010
$r_{y_{13}t_h}$	-0,80335	0,50935	-0,08345	0,07501
$r_{y_{14}t_h}$	0,20994	-0,38978	-0,11484	0,11518
$r_{y_{15}t_h}$	0,43484	-0,31498	-0,20790	-0,01990
$r_{y_{16}t_h}$	0,12847	-0,27373	-0,30289	0,16631

Les coefficients de corrélation entre les variables dépendantes et les composantes principales montrent que globalement, le degré de corrélation décroît d'une composantes à l'autre. On remarque que la première et la deuxième composante sont les plus corrélées avec les variables dépendantes que les deux autres composantes. On constate qu'à un niveau de signification 5% quatre composantes sont significatives et expliquent 83,4% de la variabilité des variables dépendantes. On retient ces quatre composantes.

TAB. 3.3.6 –. *Les coefficients de régression selon les composantes principales obtenues des données de calibration(PCR-Gauchi)*

variables	t2	t1	t4	t9
y1	0,00289	0,00267	0,00169	0,00207
y2	0,02541	-0,01218	0,01940	-0,02376
y3	0,00275	0,01782	-0,00627	-0,00182
y4	0,01374	0,07286	-0,00428	0,01750
y5	-0,24162	0,40027	-0,56552	0,61000
y6	-0,17601	0,18742	-0,45034	0,61554
y7	-0,10918	0,25632	-0,37846	0,53339
y8	-0,00121	-0,00369	0,00977	-0,00396
y9	0,00034260	0,00099	0,00611	0,00684
y10	-0,00323	0,00885	0,01354	-0,02789
y11	-0,33406	0,15001	-0,31289	0,16745
y12	-0,32973	0,13744	-0,26398	0,05320
y13	-0,59669	0,31781	-0,12282	0,23252
y14	0,00427	-0,00666	-0,00463	0,00978
y15	0,00386	-0,00235	-0,00366	-0,000737
y16	0,00411	-0,00735	-0,01918	0,02218

Pour donner une idée du comportement des écarts types des coefficients de régression et la somme des carrés des erreurs on les a présentés pour les quatre composantes significatives. Les variables qui ont une somme des carrés des erreurs plus marquante sont y_5, y_6, y_7, y_{11}, y_{12} et y_{13}.

Il est clair que les signes des coefficients de régression concordent parfaitement avec ceux de la corrélation entre les composantes et les variables dépendantes.

TAB. 3.3.7 –. *Les écarts types des coefficients de régression et la somme des carrés des erreurs en fonction des composantes principales retenues (PCR-Gauchi-calibration)*

	t1	t2	t4	t9	SCE
y1	0,000786	0,000935	0,00185	0,00391	0,00175
y2	0,00361	0,00430	0,00853	0,01796	0,03694
y3	0,00443	0,00528	0,01045	0,02202	0,05553
y4	0,00842	0,01003	0,01987	0,04185	0,20061
y5	0,04606	0,05483	0,10864	0,22884	5,99839
y6	0,05677	0,06758	0,13390	0,28203	9,11131
y7	0,05174	0,06159	0,12204	0,25706	7,56917
y8	0,00365	0,00435	0,00862	0,01815	0,03774
y9	0,00142	0,00169	0,00335	0,00707	0,00572
y10	0,00312	0,00371	0,00735	0,01548	0,02746
y11	0,05321	0,06334	0,12550	0,26434	8,00394
y12	0,03496	0,04161	0,08245	0,17368	3,45526
y13	0,04630	0,05512	0,10922	0,23005	6,06204
y14	0,00389	0,00463	0,00918	0,01933	0,04280
y15	0,00157	0,00187	0,00371	0,00782	0,00701
y16	0,00616	0,00733	0,01452	0,03059	0,10718

TAB. 3.3.8 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre de composantes retenues (PCR-calibration)*

	1	2	3	4
$r_{y_1\hat{y}_1}$	0,50684	0,75327	0,76798	0,77286
$r_{y_2\hat{y}_2}$	0,71524	0,17930	0,22957	0,21014
$r_{y_3\hat{y}_3}$	0,09239	0,58959	0,55758	0,55242
$r_{y_4\hat{y}_4}$	0,14293	0,76360	0,74458	0,74477
$r_{y_5\hat{y}_5}$	-0,36704	0,28850	0,19850	0,22215
$r_{y_6\hat{y}_6}$	-0,37300	0,09880	0,00306	0,03810
$r_{y_7\hat{y}_7}$	-0,23561	0,32856	0,24195	0,27137
$r_{y_8\hat{y}_8}$	-0,06653	-0,22365	-0,16643	-0,17123
$r_{y_9\hat{y}_9}$	0,04550	0,14672	0,22364	0,24667
$r_{y_{10}\hat{y}_{10}}$	-0,15756	0,27465	0,33442	0,29568
$r_{y_{11}\hat{y}_{11}}$	-0,69625	-0,19316	-0,25358	-0,24260
$r_{y_{12}\hat{y}_{12}}$	-0,77852	-0,23806	-0,29479	-0,28955
$r_{y_{13}\hat{y}_{13}}$	-0,80335	-0,16374	-0,17686	-0,16733
$r_{y_{14}\hat{y}_{14}}$	0,20994	-0,14709	-0,16664	-0,15267
$r_{y_{15}\hat{y}_{15}}$	0,43484	0,05957	0,01792	0,01558
$r_{y_{16}\hat{y}_{16}}$	0,12847	-0,11605	-0,17284	-0,15309

TAB. 3.3.9 –. *L'indice de redondance entre les variables dépendantes et les erreurs en fonction des composantes retenues (PCR-Gauchi-calibration)*

	1	2	3	4
$RI(Y,\hat{e})$	0,8035026	0,5408991	0,4524626	-

TAB. 3.3.10 –. *Les moyennes et les écarts types résiduels en fonction du nombre de prédicteurs (PCR-Gauchi-validation)*

	1	2	3
moyenne	0	-2,08167E-18	-2,08167E-18
σ_e	0,8454322	0,6448895	0,2623346

L'intérêt du tableau (3.3.8) est de montrer la corrélation qui existe entre les valeurs observées et les valeurs estimées des variables à expliquer. Une lecture montre qu'un certain ajustement se fait entre les corrélations selon l'ajout d'une composante principale dans le modèle. Le modèle avec les quatre premières composantes principales donne des corrélations en général meilleures que les trois premiers modèles. A savoir que cette corrélation en valeur absolue oscille entre 0,77286 pour la première variable y_1 et 0,0381 pour la 6ème variable y_6. La quasi totalité des variables ont une valeur moins que 50%.

En effet, l'indice de redondance entre les variables dépendantes et les erreurs selon le modèle avec quatre composantes n'existe pas.

La projection des données de validation sur les trois premiers modèles a donné des moyennes et des écarts types résiduels connaissant des valeurs optimales. Ces valeurs se situent autour de la valeur zéro.

3.3.3. La régression par analyse des valeurs latentes appliquée à l'exemple de Gauchi

A partir des données de calibration concatinées horizontalement aux variables dépendantes, nous avons formé des matrices, $[y_i|X]$ avec $i = 1,...16$ sur lesquelles on a appliqué l'analyse en composantes principales une à une. Par la suite, nous avons pris le nombre minimal des valeurs latentes associés aux différentes matrices qui sont sélectionnées par la règle de Webster, Gunst et Mason (1974). Il est ressorti que 13 valeurs latentes sont retenues.

A savoir que la méthode de la régression par analyse des valeurs latentes donne des résultats plus performants en introduisant chaque fois une valeur latente dans le modèle jusqu'à ce qu'on atteint le nombre de valeurs latentes considérées par la règle de Webster, Gunst et Mason (1974), puis il y a une stagnation autour des résultats.

Nous commentons les résultats pratiques obtenus par l'application de la fonction LRR programmée en Splus (voir annexe E) sur les différents jeux de données.

Les coefficients de corrélation entre les valeurs observées et les valeurs estimées des variables dépendantes, les indices de redondance entre les variables à expliquer et les erreurs et les écarts types résiduels obtenus selon le nombre des dernières valeurs latentes retenues sont donnés par les tableaux suivants:

TAB. 3.3.11 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre des dernières valeurs latentes retenues (LRR-calibration)*

variables	10	11	12	13
$r_{y_1\hat{y}_1}$	0,4138357	0,4647917	0,5706244	0,6050831
$r_{y_2\hat{y}_2}$	0,0611512	0,1801566	0,3642539	0,4080856
$r_{y_3\hat{y}_3}$	0,180681	0,217263	0,5401179	0,7106136
$r_{y_4\hat{y}_4}$	0,3228186	0,439067	0,4885997	0,6555673
$r_{y_5\hat{y}_5}$	-0,01081258	0,09771182	0,3355996	0,3890038
$r_{y_6\hat{y}_6}$	0,05290177	0,2598218	0,2725738	0,2753261
$r_{y_7\hat{y}_7}$	0,2531396	0,3113119	0,4890584	0,5762289
$r_{y_8\hat{y}_8}$	0,1977209	0,2039034	0,3352736	0,4188452
$r_{y_9\hat{y}_9}$	-0,03575096	0,09582868	0,4734496	0,5386512
$r_{y_{10}\hat{y}_{10}}$	0,1482221	0,3068292	0,3298941	0,4876086
$r_{y_{11}\hat{y}_{11}}$	0,05863762	0,1973923	0,4974258	0,5474977
$r_{y_{12}\hat{y}_{12}}$	-0,3500896	0,1537167	0,1644872	0,6506111
$r_{y_{13}\hat{y}_{13}}$	-0,1175285	0,1426896	0,480918	0,5029557
$r_{y_{14}\hat{y}_{14}}$	0,2268005	0,2279992	0,3774199	0,4175567
$r_{y_{15}\hat{y}_{15}}$	0,09523337	0,273792	0,7461132	0,8410469
$r_{y_{16}\hat{y}_{16}}$	0,2871366	0,4225538	0,5718732	0,6452507

TAB. 3.3.12 –. *L'indice de redondance entre les variables dépendantes et les erreurs en fonction des dernières valeurs latentes retenues (LRR-Gauchi-calibration)*

	10	11	12	13
$RI(Y,\hat{e})$	0,8692612	0,8330856	0,7558933	0,731796

TAB. 3.3.13 –. *Les écarts types résiduels en fonction du nombre des dernières valeurs latentes retenues (LRR-Gauchi-calibration)*

	10	11	12	13
σ_e	2,088507	1,49441	1,296711	0,9986556

D'après les tableaux ci-dessous, on constate que la structure des corrélations entre les valeurs observées et les valeurs estimées des variables dépendantes s'accroît en présence de valeurs latentes dans le modèle. En général, les corrélations avec les 13 valeurs latentes présentent dans le modèle sont supérieures à celles des autres modèles.

On constate aussi que la valeur de l'indice de redondance correspondante au modèle considéré optimal est de 0,7317 c'est à dire que l'apport des erreurs contenues dans la variance des variables dépendantes représente 73,17%.

Les écarts types connaissent une diminution au fur et à mesure que les valeurs latentes sont présentes dans le modèle et marquent 0,9986 pour le dernier modèle.

La projection des données de validation sur les différents modèles calculés selon les données de calibration a provoqué des perturbations au niveau des résultats.

TAB. 3.3.14 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre de composantes retenues (LRR-validation)*

variables	10	11	12	13
$r_{y_1\hat{y}_1}$	0,2296276	0,6212779	-0,006741405	-0,1588675
$r_{y_2\hat{y}_2}$	0,7017276	0,7311785	0,7731249	-0,7333277
$r_{y_3\hat{y}_3}$	0,2025784	0,2198179	0,6909639	-0,2882639
$r_{y_4\hat{y}_4}$	-0,3854098	-0,268905	-0,6852676	-0,7450929
$r_{y_5\hat{y}_5}$	0,7442478	0,7284841	0,9110857	0,2169669
$r_{y_6\hat{y}_6}$	0,4434308	0,3197092	0,7566522	-0,2742578
$r_{y_7\hat{y}_7}$	0,6954674	0,6110793	0,3701051	-0,01292243
$r_{y_8\hat{y}_8}$	0,7637713	0,9150064	0,2586547	-0,4778338
$r_{y_9\hat{y}_9}$	-0,1126305	0,22841	0,5445911	0,6136158
$r_{y_{10}\hat{y}_{10}}$	-0,3048888	-0,2869047	0,3864627	-0,528115
$r_{y_{11}\hat{y}_{11}}$	0,458512	0,4685069	-0,3241568	0,2659279
$r_{y_{12}\hat{y}_{12}}$	0,5387294	0,4792672	-0,4279819	0,02333089
$r_{y_{13}\hat{y}_{13}}$	0,5021014	0,1447875	0,2267845	0,2117116
$r_{y_{14}\hat{y}_{14}}$	0,08870482	0,1580275	-0,4040812	-0,467652
$r_{y_{15}\hat{y}_{15}}$	-0,5641546	-0,5176096	0,5933447	0,7613586
$r_{y_{16}\hat{y}_{16}}$	0,7711375	0,506862	0,6118013	-0,07290919

TAB. 3.3.15 –. *Les moyennes et les écarts types résiduels en fonction du nombre des dernières valeurs latentes retenues (LRR-Gauchi-validation)*

	10	11	12	13
moyennes	1,038787	0,9356123	1,018704	1,110932
σ_e	0,774842	0,7306734	0,7622336	0,8413704

En considérant les 13 valeurs latentes du modèle optimal des données de calibration, on lit sur le tableau (3.3.14) seulement deux valeurs des corrélations fortes y_9 et y_{15} et dix corrélations qui sont négatives. Par ailleurs, on remarque le modèle avec 11 valeurs latentes qui réalise une performance par rapport aux autres modèles avec 11 corrélations positives dont six sont fortes.

Quant aux écarts types résiduels, on observe une amélioration par rapport à ceux des données de calibration et que le modèle avec les onze valeurs latentes marque la valeur minimale, soit 0,7306.

Les moyennes résiduelles sont rendues presque à l'unité ce qui n'est pas souhaitable pour la stabilité de la méthode.

3.3.4. La régression par moindres carrés partiels appliquée à l'exemple de Gauchi

La procédure du travail consiste à utiliser la fonction PLS(2) programmée en Splus sur les données de Gauchi pour avoir les composantes PLS puis on applique la fonction FORWARD qui donne les composantes PLS significatives.

Pour avoir certains résultats complémentaires à l'étude on a utilisé la procédure régression en SAS.

Par ces algorithmes les composantes PLS sont extraites une à une puis on exprime successivement les variables dépendantes en fonction d'une composante PLS, de deux composantes PLS,...etc jusqu'à un nombre q de composantes PLS.

Nous commentons les résultats obtenus.

TAB. 3.3.16 –. *Un extrait des composantes PLS obtenues à partir de la fonction PLS(2)*

composantes	t1	t2	t3	t4	t5
1	3,9412522	-0,02190075	-2,55011473	2,14235279	0,84601234
2	4,4005346	-0,02463965	-1,69716601	1,88683023	0,14782810
3	3,7890909	0,80134116	-0,68968463	0,96759849	0,13062028
4	2,6713381	0,46353854	-0,25324409	-0,49235524	-0,919322461
5	3,3105529	0,84714988	0,89580671	-1,23727398	0,33381739
6	4,0887292	1,80515877	-0,05095741	-2,87728904	-0,72673388
7	2,9526151	1,72926684	0,41908778	-2,13919543	0,02986586
8	-3,0349106	3,46741271	1,19155531	1,10365645	-1,12497746
9	-3,1520456	3,31048503	2,04577544	1,09835813	0,92439605
10	-1,7601367	2,84666722	0,38069866	0,99463481	-0,26437443
11	-1,3453515	3,16593730	1,32734608	0,32876780	-0,26276463
12	-1,2054827	2,45142528	-0,05678329	0,74824893	-0,01909950
13	-3,1798929	0,06211609	-0,63720901	-1,16461602	1,40759212
14	-3,0835481	-1,03539503	-1,23785613	-0,09092302	-0,27672026
15	-3,0950493	-1,35335741	-1,84131777	-1,24377450	0,35560279
16	-3,1798687	-1,55874939	-1,84269636	-1,11907490	0,50028192
17	-3,8114830	-2,35638677	-1,53837461	-0,39506674	-0,66453578
18	-0,2450705	-5,54448297	0,15303174	1,05833758	-2,43596999
19	0,8438225	-4,96400376	2,03659053	0,49480411	2,44521158
20	1,0949042	-4,09158311	3,94551178	-0,06402046	-0,42673006

TAB. 3.3.17 –. *Les différentes étapes, les composantes entrants, l'indice de redondance, RI, et la valeur p*

Étapes	composantes entrants	ri	RI	valeur p
1	1	0,56	0,56	0
2	2	0,156	0,719	0
3	4	0,086	0,805	0,004
4	3	0,052	0,857	0,007

TAB. 3.3.18 –. *Les coefficients de corrélation simple entre les variables dépendantes et les composantes PLS(PLS-Gauchi-calibration)*

coefficients	t1	t2	t4	t3
$r_{y_1 t_h}$	0,29282	-0,71231	0,04299	0,23592
$r_{y_2 t_h}$	-0,68271	-0,51846	0,20153	0,21090
$r_{y_3 t_h}$	0,62351	-0,37163	0,07893	-0,49489
$r_{y_4 t_h}$	0,76528	-0,50481	0,03493	0,00561
$r_{y_5 t_h}$	0,83594	0,08292	-0,36308	-0,27577
$r_{y_6 t_h}$	0,61372	0,19909	-0,41236	-0,26654
$r_{y_7 t_h}$	0,72086	-0,01024	-0,43524	-0,12062
$r_{y_8 t_h}$	-0,20864	0,11331	0,46258	-0,02560
$r_{y_9 t_h}$	0,11238	-0,14676	0,53628	-0,20064
$r_{y_{10} t_h}$	0,51718	-0,10655	0,51563	-0,06600
$r_{y_{11} t_h}$	0,64495	0,51334	-0,19415	-0,36231
$r_{y_{12} t_h}$	0,68661	0,57922	-0,21640	-0,14173
$r_{y_{13} t_h}$	0,79541	0,53471	-0,10365	0,12778
$r_{y_{14} t_h}$	-0,43437	-0,03420	0,08095	-0,26589
$r_{y_{15} t_h}$	-0,44651	-0,24853	-0,04207	-0,65916
$r_{y_{16} t_h}$	-0,28026	0,02119	-0,16863	-0,60800

Une description des tableaux montre qu'au niveau de signification 5%, quatre composantes PLS sont significatives et mesurent 85,7% de la variabilité des variables à expliquer. L'apport de la première composante marque 56%. On constate aussi que les signes des coefficients de régression concordent parfaitement avec ceux de la corrélation entre les composantes et les variables dépendantes. La méthode par moindres carrés a donné de bons résultats au niveau des corrélations

TAB. 3.3.19 –. *Les coefficients de régression selon les composantes PLS obtenues des données de calibration (PLS-Gauchi)*

variables	t1	t2	t4	t3
y1	0,00146	-0,00400	0,00048941	0,00221
y2	-0,02120	-0,01808	0,01427	0,01231
y3	0,01621	-0,01086	0,00468	-0,02420
y4	0,06431	-0,04766	0,00669	0,00088
y5	0,48099	0,05360	-0,47637	-0,29835
y6	0,25313	0,09225	-0,38782	-0,20670
y7	0,29198	-0,00466	-0,40199	-0,09186
y8	-0,00331	0,00202	0,01674	-0,00076
y9	0,000739	-0,00109	0,00805	-0,00248
y10	0,00926	-0,00214	0,02104	-0,00222
y11	0,27048	0,24185	-0,18565	-0,28569
y12	0,25418	0,24088	-0,18267	-0,09865
y13	0,51639	0,38998	-0,15344	0,15597
y14	-0,00772	-0,00068	0,00328	-0,00889
y15	-0,00347	-0,00217	-0,00074	-0,00962
y16	-0,00783	0,000665	-0,01074	-0,03193

entre les valeurs observées et les valeurs estimées des variables dépendantes et des prédictions des écarts types résiduels.

En effet, la méthode est très satisfaisante pour prédire les variables à expliquer.

TAB. 3.3.20 –. *Les écarts types des coefficients de régression et la somme des carrés des erreurs en fonction des trois premières composantes PLS (PLS-Gauchi-calibration)*

	t1	t2	t4	SCE
y1	0,00079	0,00089	0,00181	0,00176
y2	0,00368	0,00413	0,00839	0,03767
y3	0,00444	0,00499	0,01013	0,05498
y4	0,00836	0,00939	0,01906	0,19461
y5	0,05799	0,06514	0,13222	9,36682
y6	0,06632	0,07450	0,15122	12,25215
y7	0,05461	0,06135	0,12452	8,30713
y8	0,00339	0,00381	0,00773	0,03200
y9	0,00136	0,00152	0,00309	0,00512
y10	0,00302	0,00339	0,00688	0,02539
y11	0,05576	0,06264	0,12714	8,66007
y12	0,03539	0,03976	0,08070	3,48929
y13	0,04315	0,04847	0,09839	5,18613
y14	0,00398	0,00448	0,00909	0,04423
y15	0,00167	0,00187	0,00380	0,00773
y16	0,00660	0,00741	0,01504	0,12123

TAB. 3.3.21 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées selon le nombre de composantes retenues (PLS-calibration)*

	1	2	3	4
$r_{y_1\hat{y}_1}$	0,29282	0,77015	0,77135	0,80662
$r_{y_2\hat{y}_2}$	0,68271	0,85726	0,88063	0,90554
$r_{y_3\hat{y}_3}$	0,62351	0,72586	0,73014	0,88205
$r_{y_4\hat{y}_4}$	0,76528	0,91678	0,91745	0,91746
$r_{y_5\hat{y}_5}$	0,83594	0,84004	0,91515	0,95580
$r_{y_6\hat{y}_6}$	0,61372	0,64520	0,76572	0,81078
$r_{y_7\hat{y}_7}$	0,72086	0,72093	0,84213	0,85072
$r_{y_8\hat{y}_8}$	0,20864	0,23743	0,51995	0,52058
$r_{y_9\hat{y}_9}$	0,11238	0,18484	0,56724	0,60168
$r_{y_{10}\hat{y}_{10}}$	0,51718	0,52805	0,73804	0,74099
$r_{y_{11}\hat{y}_{11}}$	0,64495	0,82431	0,84686	0,92111
$r_{y_{12}\hat{y}_{12}}$	0,68661	0,89829	0,92399	0,93480
$r_{y_{13}\hat{y}_{13}}$	0,79541	0,95843	0,96402	0,97245
$r_{y_{14}\hat{y}_{14}}$	0,43437	0,43571	0,44317	0,51681
$r_{y_{15}\hat{y}_{15}}$	0,44651	0,51102	0,51275	0,83511
$r_{y_{16}\hat{y}_{16}}$	0,28026	0,28106	0,32777	0,69072

TAB. 3.3.22 –. *L'indice de redondance entre les variables dépendantes et les erreurs en fonction des composantes PLS retenues(PLS-Gauchi-calibration)*

	1	2	3	4
$RI(Y,\hat{e})$	0,660522	0,5224108	0,4323746	-

TAB. 3.3.23 –. *Les moyennes et les écarts types résiduels en fonction du nombre de composantes (PLS-Gauchi-validation)*

	1	2	3
moyennes	0	2,775558E-18	-8,32667E-18
σ_e	0,6462785	0,5646051	0,2553080

3.4. Conclusion

Principalement, la comparaison est basée d'une part sur les corrélations entre les valeurs observées et les valeurs estimées des variables à expliquer pour le cas multiple et de l'indice de redondance entre les variables dépendantes et les résidus dans le cas multivarié, et d'autre part, sur les écarts types résiduels en fonction des composantes considérées.

Dans notre cas pratique, examinons les graphiques fournis par "Splus". Les figures 3.4.1 et 3.4.3, provenant de l'exemple d'octane montrent que les prédictions des écarts types résiduels ainsi que les corrélations entre les valeurs observées et les valeurs estimées de l'indice d'octane donnent une idée sur la qualité prédictive selon les trois méthodes. De ces graphiques, les trois méthodes donnent des résultats qui se rapprochent pour les modèles optimaux de chaque méthode.

Par ailleurs, la régression par analyse des valeurs latentes donne de mauvaise régression en considérant la première et/ou la deuxième valeur latente dans le modèle.

D'après notre description de données et les représentations graphiques pour les trois méthodes, l'avantage de la régression par moindres carrés partiels par rapport aux deux autres méthodes est clair.

En effet, l'exemple de Gauchi illustre bien la supériorité de la régression par moindres carrés partiels par rapport aux deux autres méthodes.

La figure 3.4.4. présente les indices de redondance entre les variables dépendantes et les résidus. Cette représentation montre que l'apport des résidus contenus dans les variables dépendantes est minime pour la régression par moindres carrés partiels par rapport aux deux autres méthodes.

La figure 3.4.2. montre que la prédiction des écarts types résiduels selon les trois méthodes est favorable pour la régression par moindres carrés partiels.

En résumé, la méthode de la régression par moindres carrés partiels s'avère être une méthode plus efficace dans notre contexte que les deux autres méthodes.

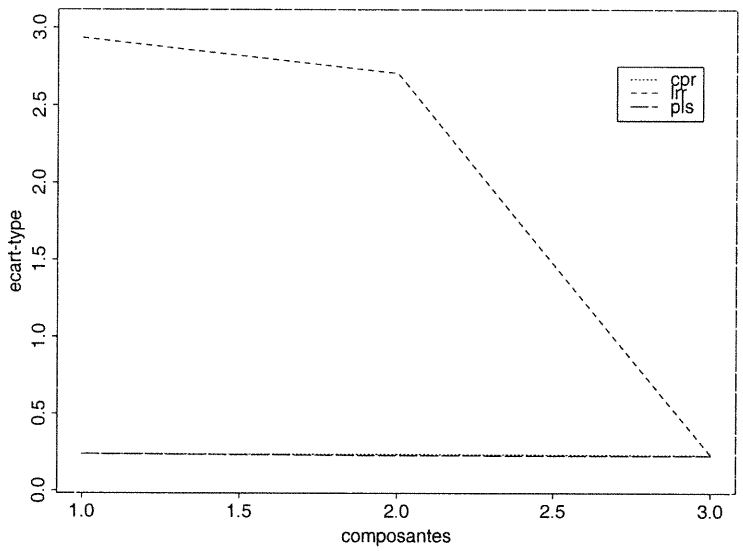

FIG. 3.4.1 –. *Les écarts types résiduels en fonction des composantes selon les trois méthodes (l'exemple d'octane)*

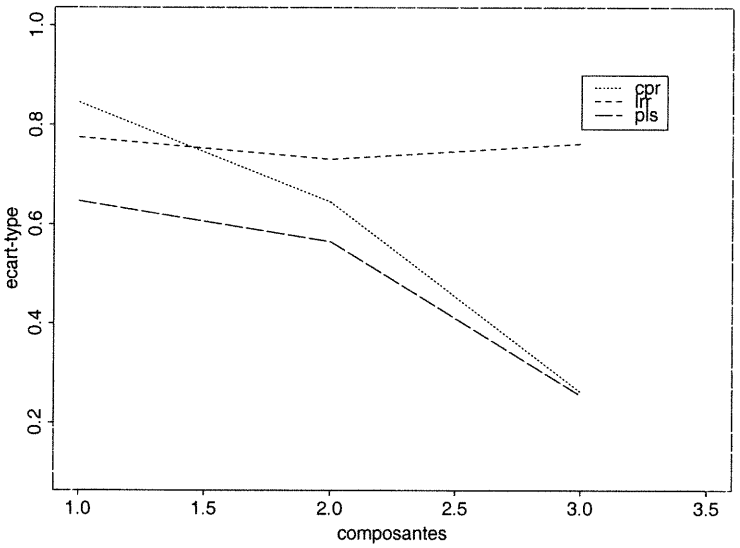

FIG. 3.4.2 –. *Les écarts types résiduels en fonction des composantes selon les trois méthodes (l'exemple de Gauchi)*

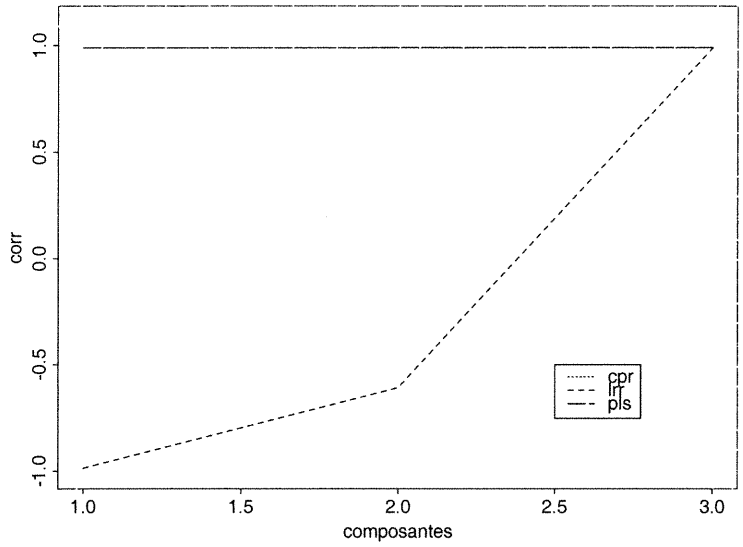

FIG. 3.4.3 –. *Les coefficients de corrélation entre les valeurs observées et les valeurs estimées en fonction des composantes selon les trois méthodes (l'exemple d'octane)*

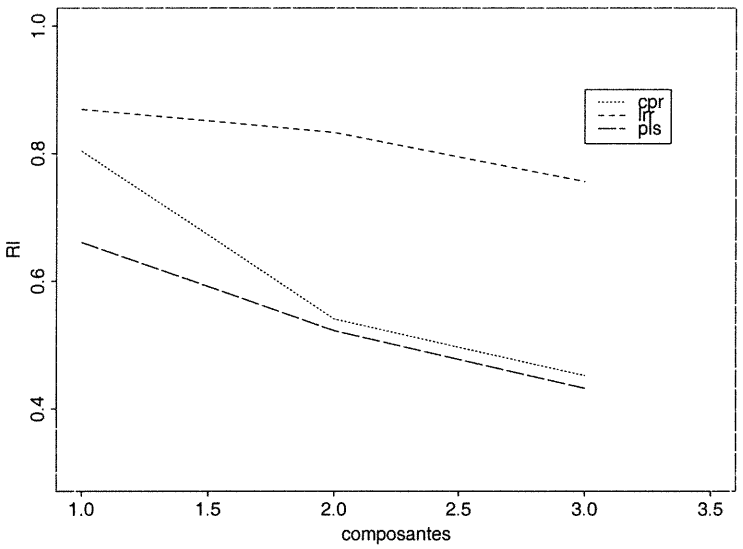

FIG. 3.4.4 –. *L'indice de redondance entre les variables dépendantes et les erreurs en fonction des composantes selon les trois méthodes(l'exemple de Gauchi-calibration)*

Annexe A

LES PROGRAMMES EN SAS

Programmes SAS qui donnent les composantes principales dans le cas de la régression en composantes principales ou les valeurs latentes dans le cas de la régression par analyse des valeurs latentes.

```
OPTIONS linesize=80 nonumber nodate formdlim='-';
filename données "fichier des données";
data ....;
infile données;
input variable ou variables dépendantes + les variables explicatives;
run;
proc print data= .... ;
run;
proc princomp (cov) data=.... out=prin;
var variables explicatives pour le cas PCR et la variable dépendante + les variables ex-
plicatives pour le cas LRR;
run;
proc print data=prin noobs;
var variable dépendante prin1 prin2 .....;
run;
```

Programmes SAS qui appliquent la procédure FORWARD et la régression de la variable dépendante sur les variables explicatives dans le cas multiple (pour les méthodes PCR et PLS).

```
OPTIONS linesize=80 nonumber nodate formdlim='-';
filename donnees "nom de fichier des données";
data ....;
infile donnees;
input la variable et les composantes ;
run;
proc print data=....;
run;
title" Forward";
proc reg data=....;
model la variable dépendante = les composantes ... / SELECTION=forward SLE = α ;
run;
title"La matrice de corrélation";
proc corr data=....;
var la variable et les composantes;
run;
title "la régression en fonction de la première composante";
proc reg data=....;
model la variable dépendante=t1 / R;
output out=new1 P=yhat1 R=residu1;
run;
proc corr data=new1;
var y yhat1;
run;
title "La régression en fonction des deux premières composantes";
proc reg data=....;
model la variable dépendante =t1 t2 / R;
output out=new2 P=yhat2 R=residu2;
```

```
run;
proc corr data=new2;
var y yhat2;
run;
...     ...
...     ...
data concat;
merge new1 new2 ...;
run;
title"La matrice des residus";
proc print data=concat noobs;
var residu1 residu2 ... ;
run;
title"Les moyennes et les écarts-type résiduels";
proc means data=concat ;
var residu1 residu2 ...;
run;
```

Programmes SAS qui appliquent la régression des variables dépendantes sur les variables explicatives dans le cas multivarié (pour les méthodes PCR et PLS).

```
filename donnees "nom de fichier contenant les variables dépendantes et les composantes";
data .....;
infile donnees;
input #1 les variables dépendantes
#2 les composantes;
run;
proc print data = .....;
run;
title"La matrice de corrélation";
proc corr data=.... noprob nosimple;
var composantes;
```

```
with variables dépendantes;
run;
title "la régression multivariée";
proc reg data=....;
model la 1ère varaiable dépendante=t1 / R;
output out=new1 P=prev1 R=residus1 ;
run; proc reg data=....;
model la 2ème varaiable dépendante=t1 / R;
output out=new2 P=prev2 R=residus1 ;
run; proc reg data=....;
model la 3ème varaiable dépendante=t1 / R;
output out=new3 P=prev3 R=residus1 ;
run; ...      ...
...      ...
(autant de output que de variables dépendantes)
run;
data concat1;
set new1 new2 new3 ..... ;
run;
title"correlation entre y et yhat";
proc corr data=concat1;
var les variables dépendantes;
with prev1 prev2 prev3....;
run;
title"Les corrélations entre y et yhat (2)";
proc reg data=....;
model la 1ère variable dépendante=t1 t2/ R;
output out=new17 P=prev17 R=residus2 ;
run; proc reg data=....;
model la 2ème variable dépendante=t1 t2/ R;
output out=new18 P=prev18 R=residus2 ;
run; proc reg data=....;
model la 3ème variable dépendante=t1 t2/ R;
```

```
output out=new19 P=prev19 R=residus2 ;
run; ...      ...
...      ...
run;
data concat2;
set new17 new18 new19 ..... ;
run;
proc corr data=concat2;
var les variables dépendantes;
with prev17 prev18 prev19 ......;
run;
...      ...
...      ...
(on répéte tous les étapes une fois qu'on ajoute une composante dans le modèle.)
data concat;
merge concat1 concat2 ..... ;
run;
title"La matrice des résidus";
proc print data=concat noobs;
var residus1 residus2 ..... ;
run;
title"Les moyennes et les écarts-type résiduels";
proc means data=concat ;
var residus1 residus2 .....;
run;
```

Annexe B

LE PROGRAMME EN "SPLUS" QUI CALCULE LES COMPOSANTES PLS DE LA RÉGRESSION PAR MOINDRES CARRÉS PARTIELS DANS LE CAS MULTIPLE (FONCTION PLS(1))

```
pls1 <- function(YX, n, q, pr)
{ #function pls1 cherchant avec RI (exactement)le nombre de composantes statistiquement significatives.
# dans une procedure pls.
#n :nombre d'observations.
#p :nombre de variables Y.
#q :nombre de variables X.
#YX :tableau de donnees n*(p+q).
#Pr :la valeur limite alpha sous H0.
dq <- 2:(1 + q)
s <- var(YX)
YX <- scale(YX, center = T,scale = F)
Y <- YX[, 1]
SL <- 0
pvalue <- 0
h <- 0
mi <- min(n - 1, q)
while(pvalue <pr && h <mi) {
h <- h + 1
tam <- s[, 1]
sxy <- tam[-1]
```

```
sxx <- s[dq, dq]
U <- sxy %*% t(sxy)
valp <- trace.mat(U)
a <- sxy
th <- YX[, dq] %*% a
cat("\n")
cat("h=")
print(h)
cat("\n")
cat("th")
print(th)
r <- cor.test(Y, th, alt = "g")
ri <- -(r$estimate)^2
names(ri) <- NULL
cat("ri")
print(ri)
SL <- SL + ri
cat("totalRI = ")
print(SL)
v <- var(th)[1, 1]
pvalue <- r$p.value
print(pvalue)
ph <- -sxx%*%a/v
YX[,dq] <- -YX[,dq] - th%*%t(ph)
s <- var(YX)
}
if(h <q)
print(h - 1)
else print(h)
cat(":composantes PLS suffisent")
cat("\n")
}
```

Annexe C

LE PROGRAMME EN "SPLUS" QUI CALCULE LES COMPOSANTES DE LA RÉGRESSION PAR MOINDRES CARRÉS PARTIELS DANS LE CAS MULTIVARIÉ(FONCTION PLS(2))

```
pls2<- function(YX, n, p, q, pr)
{#function pls cherchant avec RI (exactement)le nombre de composantes statistiquement
significatives
# dans une procedure pls avec plusieurs y
#n :nombre d'observations
#p :nombre de variables Y
#q :nombre de variables X
#YX :tableau de donnees n*(p+q)
#Pr :la valeur limite alpha sous H0
m <- p + q
dp <- 1:p
dq <- (p + 1):m
s <- var(YX)
syy <- s[dp, dp]
lambda <- eigen(syy, symmetric = T)$values
YX <- scale(YX, center = T, scale = F)
# ric <- rv("rvi", YX, p)
# cat("RI complet:")
# print(ric)
# cat("\n")
YXh <-   matrix(rep(1, n * (1 + p)), nrow = n)
```

```
YXh[, 1:p] <-   YX[, 1:p]
SL <-   0
pvalue <-    0
h <-   0
mi <-    min(n - 1, q)
while(pvalue <pr && h <mi) {
h <- h + 1
syx <- s[dp, dq]
sxy <- t(syx)
sxx <- s[dq, dq]
U <- sxy %*% syx
H <- eigen(U, symmetric = T)
valp <- H$values[1]
ah <- H$vectors[, 1]
th <- YX[, dq] %*% ah
YXh[, p + 1] <- th
S <- var(YXh)
cat("\n")
cat("h=")
print(h)
cat("\n")
print(th)
cat("ah")
print(ah)
ri <- rv("rvi", S, p, T)
cat("ri=")
print(ri)
SL <- SL + ri
cat("total RI=")
print(SL)
v <- var(th)[1, 1]
pvalue <- test(ri, n, p, lambda)
cat("pvalue")
```

```
print(pvalue)
ph <- sxx %*% ah/v
YX[, dq] <- YX[, dq] - th %*% t(ph)
cat("\n")
s <- var(YX)
} if(h <q)
print(h - 1)
else print(h)
cat(":composantes PLS suffisent ")
cat("\n")
}
```

(Cette fonction fait appelle à la fonction test.)

Annexe D

LA FONCTION FORWARD EN "SPLUS" QUI TRAITE LES DONNÉES MULTIVARIÉES

```
function(x, n, p, q)
{
#function forward
#x:the data matrix of dimension n*(p+q)
#p:the dimension of the independant vector
#q:the dimension of the dependant vector
#n:the number of observations
s <- var(x)
s11 <-  s[1:p, 1:p]
tr <-  trace.mat(s11)
ino <-  0
tab <-  c(rep(1, q))
h <-  0
cat("var", "RI", "pvalue", sep = " ", "\n")
while(ino <q) {
rm <-  0
ino <-  ino + 1
for(j in (1:q)) {
k <-  j + p
lv <-  match(k, tab)
if(is.na(lv)) {
tab[ino] <- k
```

```
ri <-   ric(tab, ino, s, p, tr)
if(rm <ri) {
rm <- ri
kk <- p + j
}
}
}
rip <-   (rm - h)/(1 - h)
h <-   rm
tab[ino] <-   kk
if(rip <1) {
if(ino == 1)
s113 <-   s11
else s113 <-   ssp(s11, s, tab, p, ino)
prob <-   test(rip, s113, n, p, ino)
}
else prob <-   1
im <-   tab[ino] - p
im <-   round(im, 0)
rm <-   round(rm, 3)
prob <-   round(prob, 3)
cat(im, rm, prob,sep = "", "\n")
}
}
```

(Cette fonction fait appelle à d'autres fonctions qui se trouvent dans le répertoire cléroux/lazraq/lazraq/Step. Ces fonctions sont: trace.mat, ssp, ric et test.)

Annexe E

LA FONCTION LRR EN "SPLUS" QUI CALCULE LES RÉSULTATS DE LA RÉGRESSION PAR ANALYSE DES VALEURS LATENTES

lrr <- function(yd,X, V, lambda, l)

{

#function lrr

#X:la matrice des données

#V:la matrice des vecteurs propres associés aux valeurs propres de la matrice ($A^{'}A$ **avec** $A = [yd|X]$

#yd:la variable dépendante

#lambda:le vecteur des valeurs propres

#l:le nombre de valeurs latentes retenu

yd <- $scale(yd, center = T, scale = (FouT))$

X <- $scale(X, center = T, scale = (FouT))$

for(k in (1:l)) {

A1 <- $V[1,1:k]^2/lambda[1:k]$

A <- sum(A1)

q <- $A1/A$

y <- $-X\% * \%V[-1,1:k]$

for(j in (1:k)) {

y[, j] <- $y[,j]/V[1,j]$

}

yc <- $y\% * \%q$

res <- $(yd - yc)$

```
m <-sqrt((res)^2)
r1 <-mean((res)^2)
e <- sqrt(r1)
corr <- cor(yd, yc)
cat("\n")
print(q)
cat("\n")
cat("e:")
print(e)
cat("\n")
print(yc)
cat("\n")
cat("corr :")
print(corr)
cat("\n")
cat("res :")
print(res)
}
}
```

BIBLIOGRAPHIE

[1] DRAPPER N.R., SMITH H., *Applied regression analysis*, 3rd Ed., Wiley New York, (1998).

[2] ESBENSEN K., SCHÖNKOPF S. AND MIDTGAARD T., *Multivariate analysis in practice*, CAMO, Olev tryggvasons gt .24, N-7011 trondheim, Norway (1994).

[3] GAUCHI J-P., *Utilisation de la régression PLS pour l'analyse des plans d'expérience en chimie de formulation*, Rev.Statistique Appliquée, XLIII(1), 65-89 (1995).

[4] GLEASON T. C., *On redundancy in canonical analysis*, Psychological Bulletin, 83, 1004-1006 (1976).

[5] LAZRAQ A. ET CLÉROUX R., *Etude comparative de différentes mesures de liaison entre deux vecteurs aléatoires*, Statistique et Analyse des données 13, 15 (1988a).

[6] LAZRAQ A. ET CLÉROUX R., *Un algorithme pas à pas de selection de variables en régression linéaire multivariée*, Statistique et Analyse des données 13, 39 (1988b).

[7] LAZRAQ A. ET CLÉROUX R., *The PLS multivariate regression model:testing the significance of successive PLS components*, Jour. of Chemometrics, 15, 6, 523-536 (2001).

[8] LONGLEY J.W., *An Apparaisal of Least Squares Programs for the Electronic computer from the Point of View of the User*, Journal of the American Statistical Association, 62, 819-841 (1967).

[9] MARDIA K.V., KENT J.T. AND BIBBY J.M., *Multivariate Analysis*, Academic Press, London (1979).

[10] PALM R., *Les méthodes d'analyse factorielle: principes et applications*, Biom. Praxim. 34, p. 35-80 (1994).

[11] RAWLINGS J.O., PANTULA S.G., DICKEY D.A., *Applied regression analysis: A reserch tool*, 2nd Ed., Springer-verlag New york, Inc. (1998).

[12] STEWART D. AND LOVE W., *A General Canonical Correlation*, Index, Psycho. Bull., 70, p. 160-163 (1968).

[13] TENENHAUS M., *La régression PLS:théorie et pratique*, Edition Technip, Paris (1998).

[14] TENENHAUS M., GAUCHI J-P. AND MENARDO C., *Régression PLS et application*, Rev.Statistique Appliquée, XLIII(1), 7-63 (1995).

[15] VIGNEAU E., BERTRAND D. AND QANNARI E.M., *Application of latent root regression for calibration in near-infrared spectroscopy. Comparison with principal component regression and partial least squares*, Chemometrics and Intelligent laboratory System, 35, 231-238 (1996).

[16] WALTZ W.G., REID W.A. AND COLWELL W.E., *Sugar and Nicotine in Cured Bright Tobacco as Related to Mineral Element Composition*, Proc. Soil Sci. Soc. Am., 13, 385-387 (1948).

[17] WEBSTER J.T., GUNST R.F. AND MASON R.L., *Latent root regression analysis*, Technometrics, 16,513-522 (1974).