INTRODUCTION A LA REGRESSION DES MOINDRES CARRES PARTIELS AVEC LA PROCEDURE PLS DE SAS

Dominique DESBOIS*

Institut national de la recherche agronomique - Economie et Sociologie Rurales

(INRA-ESR, Nancy)

s/c SCEES-Bureau du RICA, 251 rue de Vaugirard, 75732 PARIS Cedex 15.

E-mail : desbois@jouy.inra.fr, desbois@agriculture.gouv.fr

Résumé : Cet article est une introduction à la mise en oeuvre de la procédure PLS de SAS permettant d'utiliser certaines techniques de régression sur variables latentes. On présente tout d'abord trois modèles de régression sur variables latentes (régression sur composantes principales, analyse des redondances et régression des moindres carrés partiels), puis les options correspondantes de la procédure PLS. Des exemples simples d'utilisation illustrent l'interprétation des résultats fournis par la procédure pour la régression des moindres carrés partiels. *Une bibliothèque de macro-programmes SAS facilitant l'interprétation des résultats paraîtra dans le numéro 25 de la Revue de MODULAD et sera accessible sur le WEB.*

Mots-clés : Régression PLS, Moindres carrés partiels, Logiciel SAS, Régression sur composantes principales, Analyse des redondances, Structures latentes, Chemin de causalité.

* L'auteur remercie les professeurs Gilbert Saporta et Michel Tenenhaus pour les conseils et la documentation scientifique et technique mais reste seul responsable d'éventuelles erreurs commises dans la rédaction de cet article.

I) La régression sur variable latentes

I.1) L'origine de la régression sur variables latentes

Il est d'usage courant dans beaucoup de domaines scientifiques et techniques de vouloir expliquer, contrôler ou prédire le comportement d'une ou plusieurs variables (les *variables expliquées*, encore appelées *réponses* ou *variables endogènes*) par des indicateurs plus faciles à mesurer ou à contrôler (les *variables explicatives*, encore appelées *facteurs* ou respectivement *variables exogènes*). Lorsque ces variables explicatives sont peu nombreuses, qu'elles ne sont pas trop redondantes (*colinéaires*) et que leurs liaisons avec les réponses sont suffisamment bien connues (*modélisées*), la régression linéaire multiple est la technique adéquate pour ajuster un modèle aux données et le transformer en un outil opérationnel d'appréhension de la réalité. Cependant si l'une ou l'autre de ces trois conditions n'est pas remplie, la régression multiple peut s'avérer peu efficiente, voire inadaptée. C'est souvent le cas, par exemple, en chimie où l'on essaie souvent d'estimer les différentes proportions de composants qui constituent un mélange (*spectrogramme*). Les mesures fournies par la spectrographie, exprimées en pourcentages, présentent souvent une grande colinéarité. Les sciences sociales offrent également de nombreux exemples de telles situations : en micro-économie par exemple, certaines modélisations tentent d'estimer le revenu des agents économiques ou les marges des entreprises par l'analyse de certains agrégats comptables ou ratios financiers qui sont généralement très corrélés.

Modèle particulier de régression sur variables latentes, la régression des moindres carrés partiels (MCP) est une méthode statistique permettant de construire des modèles prédictifs lorsque les variables explicatives sont nombreuses et très corrélées. L'accent est mis dans cette méthode sur la prédiction et non sur l'identification, problématique qui relève plutôt des différentes techniques descriptives d'analyse factorielle. La régression MCP a été développée à l'origine par Herman Wold dans les années 60 comme technique économétrique puis fut popularisée auprès des chimistes par son propre fils, Svante Wold – créateur du logiciel SIMCA[1], pour être appliquée au contrôle de processus industriels. La régression MCP est reliée à d'autres techniques multivariées sur variables latentes comme la régression sur composantes principales ou l'analyse de la redondance maximale.

---

[1] *Soft Independant Modelling by Class Analogy*

I.2) Les modèles de régression sur variables latentes

On peut en principe inclure de nombreuses variables explicatives dans un modèle de régression multiple. Cependant, si on laisse le nombre de variables explicatives croître sans contrôle voire dépasser le nombre d'observations, le risque est d'obtenir un modèle qui s'ajuste parfaitement aux données mais dont les estimations obtenues par combinaison des variables explicatives s'avèrent incapables de prédire correctement les valeurs des variables expliquées pour de nouvelles observations. Souvent dans ces situations de saturation du modèle, bien que les variables explicatives a priori soient nombreuses, seul un petit nombre de *variables latentes*[2] permet de rendre compte de la variabilité des variables expliquées.

Par exemple, l'objectif de la régression MCP est d'extraire des « ***composantes*** », variables latentes responsables de la variation des variables explicatives, qui modélisent au mieux le comportement des variables expliquées. Pour prédire les variables expliquées à partir des variables explicatives, la régression MCP procède de manière indirecte en extrayant des composantes T à partir des variables explicatives X pour estimer des composantes U qui permettront de calculer les réponses Y. Cette procédure d'estimation recouvre comme cas particuliers plusieurs techniques statistiques suivant la nature de la variabilité que l'on souhaite expliquer :

- **La régression sur composantes principales (RCP)**[3] - les composantes T sont choisies pour reconstituer de façon optimale la variabilité des facteurs en extrayant les composantes principales du tableau X, directions principales d'allongement du nuage des facteurs. Mais la détermination des composantes principales de X ne tient aucun compte de la forme du nuage des réponses Y à prédire.
- **l'analyse des redondances maximales (RMX)**[4] – Il s'agit en fait d'une régression sur les composantes de l'analyse en composantes principales sur variables instrumentales (ACPVI). Les composantes U sont choisies pour reconstituer de façon optimale la variabilité des réponses en recherchant les directions dans l'espace des facteurs X, qui soient associées à la plus grande part de variabilité des réponses Y.

---

[2] Le terme de « variable latente » utilisé dans cet exposé possède un contenu sémantique particulier à la régression MCP et ne doit pas être confondu avec d'autres acceptions du terme, comme celle définie dans la modélisation formelle d'équations structurelles.

[3] Pour effectuer une régression sur composantes principales, on peut utiliser la procédure REG de SAS.

[4] Pour effectuer une analyse des redondances maximales, on peut utiliser la procédure TRANSREG de SAS.

Cette approche proposée par [van den Wollenberg, 1977] a l'inconvénient de fournir des estimations qui peuvent ne pas être très précises.

- **La régression des moindres carrés partiels (MCP)** - les composantes T et U sont choisies pour obtenir une liaison optimale au sein de chaque paire de composantes. Il s'agit en fait d'une forme robuste de l'analyse des redondances maximales qui recherche les directions de l'espace des variables explicatives liées aux plus fortes variations dans l'espace des variables expliquées mais biaisées vers des directions mieux prédites.

La **décomposition spectrale de l'opérateur d'inertie** permet d'établir un lien entre les trois méthodes qui précèdent. En effet, la régression sur composantes principales est basée sur la décomposition spectrale de l'opérateur d'inertie $X'X$ où X est le tableau des facteurs ; l'analyse des redondances maximales utilise la décomposition spectrale de la matrice $\hat{Y}'\hat{Y}$, où $\hat{Y}$ est le tableau des valeurs estimées des réponses Y ; et la régression selon les moindres carrés partiels s'appuie sur la décomposition spectrale de l'opérateur $X'Y$.

Si le nombre de composantes extraites est supérieur ou égal à la dimension de l'espace des facteurs, la régression MCP fournit alors la solution des moindres carrés ordinaires (MCO). Cependant, la méthode des moindres carrés partiels permet en général de se limiter à un nombre restreint de composantes. Le choix du nombre de composantes à extraire est effectué sur la base de techniques heuristiques cherchant à minimiser la variance résiduelle. Une des approches utilisées consiste à construire le modèle MCP avec un nombre fixé de variables explicatives pour un sous-ensemble particulier des données, puis à tester le modèle obtenu sur un autre sous-ensemble des données en choisissant le nombre de composantes qui minimise l'erreur d'estimation. On peut également choisir le plus petit nombre de composantes associé aux modèles dont la variance résiduelle n'est pas significativement supérieure à celle du modèle à erreur résiduelle minimale [van der Voet, 1994]. Si l'on ne dispose pas d'un échantillon-test, on peut utiliser chaque observation en tant qu'échantillon-test, on effectue alors une *validation croisée*.

I.3) Des modèles aux applications

Les applications des techniques de régression en sciences humaines et sociales présentent un multidéterminisme comportant des variables explicatives parfois si nombreuses qu'il serait vain de vouloir en expliciter toutes les interrelations au sein d'un même modèle. La régression MCP apparaît comme l'une des solutions adaptées à de tels problèmes de spécification du modèle, mais il en existe d'autres, en particulier :

- les techniques factorielles, comme la régression sur composantes principales et l'analyse des redondances maximales ;
- la régression pseudo-orthogonale, conçue à l'origine par les statisticiens (cf. [Hoerl et Kennard, 1970]) pour maîtriser les problèmes induits par la colinéarité en régression ;
- les réseaux neuronaux, technique de reconnaissance des formes s'efforçant de simuler le comportement du cerveau à partir de concepts communs à l'informatique et à la biologie (cf. [Sarle, 1994]) ;

En termes de flexibilité et de robustesse des modèles prédictifs, régression pseudo-orthogonale et réseaux de neurone sont probablement les techniques les mieux placées pour concurrencer la régression MCP. Cependant, aucune d'entre elles n'inclut comme propriété la réduction de la dimensionalité qu'apporte la régression MCP en construisant par combinaison linéaire un petit nombre de composantes susceptibles de modéliser le comportement des variables expliquées (cf. [Franck et Friedman, 1993] pour une comparaison des différentes techniques).

D'autre part, il existe un certain nombre de modifications ou d'extensions de la régression MCP. Par exemple, l'algorithme SIMPLS[5], proposé par de Jong en 1993, donne toujours des résultats très proches du modèle originel des moindres carrés partiels, tout en facilitant considérablement les calculs en particulier si le nombre de variables explicatives est élevé. Le modèle de régression continûment paramétrable, (*continuum regression*), proposé par [Stone et Brooks, 1990], permet de faire varier continûment le modèle de régression suivant les valeurs d'un paramètre α, $0 \leq \alpha \leq 1$, du modèle de la régression multiple ($\alpha = 0$) au modèle de la régression sur composantes principales ($\alpha = 1$) en passant par celui des moindres carrés partiels ($\alpha = 0.5$). En 1992, de Jong et Kiers ont proposé une technique apparentée, la « régression sur covariables principales » (*principal covariates regression*).

---

[5] pour « *Straightforward Implementation of a statistically inspired Modification of the PLS method* »

En dernière analyse, la technique des moindres carrés partiels peut constituer, en chimiométrie comme dans bien d'autres domaines, un outil privilégié d'analyse dans la mesure où il s'avère possible d'interpréter les composantes en termes de facteurs sous-jacents aux variables explicatives. Cependant, il convient d'étudier de façon plus approfondie l'application de cet ensemble de techniques statistiques à la spécification de modèles, sur la base des propositions faites en 1994 par van der Voet concernant une procédure randomisée de comparaison de modèles.

I.4) Pour une utilisation avancée

La régression des moindres carrés partiels demeure une technique de modélisation statistique en constante évolution. On pourra consulter [Geladi et Kowalski, 1986] pour une introduction classique à l'application des moindres carrés partiels en chimiométrie. Parmi les références bibliographiques citées, [Naes et Martens, 1985] et [de Jong, 1993] constituent des présentations techniques plus détaillées de la méthodologie MCP. Une synthèse récente, couvrant un très large champ, est disponible désormais en français [Tenenhaus, 1998].

II) La syntaxe de la procédure PLS de SAS

PLS est une procédure statistique du module STAT de la version 6.11 de SAS permettant à titre expérimental de construire des modèles de régression sur variables latentes selon différentes méthodes, notamment celle des moindres carrés partiels. Les autres méthodes incluses actuellement dans cette procédure comprennent des algorithmes conçus comme des alternatives à la méthode originelle des MCP, tel que la méthode SIMPLS proposée par [de Jong, 1993] et la méthode RLGW proposée par [Ranner et alii, 1994], ainsi que la régression sur composantes principales. L'analyse des redondances maximales devrait également être incluse dans une prochaine version. Les variables explicatives peuvent être spécifiées suivant une syntaxe permettant une modélisation de type GLM. La procédure propose un grand nombre de méthodes de validation croisée pour déterminer le nombre approprié de composantes, avec un test optionnel approprié [van der Voet, 1994]. Les tables de sortie SAS permettent de récupérer les ensembles de données pour la validation croisée et l'information spécifique au modèle comme les valeurs estimées et les *coordonnées pseudo-factorielles* (projection des variables ou des observations selon les pseudo-composantes).

La spécification de la procédure PLS suit la syntaxe présentée ci-dessous, les items entre crochets (< >) étant optionnels :

PROC PLS < ***options*** > ;
CLASS ***variables_nominales*** ;
MODEL ***réponses*** = ***facteurs*** < ***options*** > ;
OUTPUT OUT = ***table_SAS*** < ***options*** > ;

Instruction PROC PLS

PROC PLS < ***options*** > ;

Cette instruction est utilisée pour invoquer la procédure PLS et pour, de façon optionnelle, indiquer les données analysées et la méthode utilisée. Voici les options disponibles :

DATA = ***table_SAS***

indique la table SAS qui contient les données soumises à l'analyse, valeurs des variables explicatives et expliquées.

METHOD = *méthode_d'extraction_des_composantes*

indique la méthode utilisée pour l'extraction des composantes. On peut spécifier l'une des méthodes suivantes :

METHOD = PLS < (*options_PLS*) >

sélectionne les moindres carrés partiels ; c'est l'option par défaut et l'algorithme standard des moindres carrés partiels.

METHOD = SIMPLS

sélectionne la méthode SIMPLS ; cet algorithme est plus efficace que l'algorithme standard des moindres carrés partiels. La méthode SIMPLS est équivalente à la méthode PLS si la réponse est univariée, et donne des résultats très similaires pour les autres cas.

METHOD = PCR

sélectionne la régression sur composantes principales.

Après l'instruction **METHOD = PLS**, vous pouvez indiquer entre parenthèses l'une des *options_PLS* suivantes :

ALGORITHM = *algorithme_PLS*

spécifie l'algorithme utilisé pour calculer les composantes PLS. Sont disponibles les algorithmes suivants

ITER sélectionne l'algorithme itératif NIPALS[6], le plus couramment utilisé.

SVD décomposition en valeurs singulières de l'opérateur $X'Y$, algorithme le meilleur au plan de la précision numérique, mais le pire au plan de l'efficacité algorithmique.

EIG décomposition spectrale (en valeur propres) de l'opérateur $Y'XX'Y$.

RLGW algorithme itératif performant lorsque le nombre de variables explicatives est élevé.

---

[6] pour « *Nonlinear estimation by Iterative Partial Least Squares* »

MAXITER = *nombre*

spécifie le nombre maximum d'itérations pour les algorithmes ITER et RLGW ; la valeur par défaut est 200.

EPSILON = *nombre*

indique le seuil de convergence pour les algorithmes itératifs ITER et RLGW ; la valeur par défaut est 10^{-12}.

CV = *méthode_de_validation_croisée*

indique la méthode utilisée pour effectuer la validation croisée. Par défaut, si cette option n'est pas spécifiée, la procédure ne procède à aucune validation croisée. Les options de validation croisée sont les suivantes :

CV = ONE

sélectionne une validation croisée, effectuée observation par observation ; chaque observation constitue alors un échantillon-test.

CV = SPLIT < (*n*) >

indique que chaque n^e observation est incluse dans l'échantillon-test ; la spécification de *n* est optionnelle, sa valeur par défaut est 1, équivalent à l'option CV=ONE.

CV = BLOCK < (*n*) >

indique que chaque bloc de *n* observations est inclus dans l'échantillon-test ; la spécification de *n* est optionnelle, sa valeur par défaut est 1, équivalent à l'option CV=ONE.

CV = RANDOM < (*options_de_validation_croisée_aléatoire*) >

indique que les observations sont incluses dans l'échantillon-test de façon aléatoire.

CV = TESTSET < (*table_SAS*) >

permet de spécifier une table SAS contenant les observations utilisées comme échantillon-test dans la procédure de validation croisée.

On peut également indiquer, entre parenthèses, après l'instruction CV = RANDOM, les options de validation croisée aléatoire qui suivent :

NITER = *nombre*

indique le nombre d'échantillon-test aléatoire à extraire.

NTEST = *nombre*

indique le nombre d'observations à sélectionner dans chacun des échantillons-tests aléatoires.

SEED = *nombre*

spécifie la valeur du germe utilisé par le générateur de nombres aléatoires.

CVTEST < (*options_du_test_de_validation_croisée*) >

indique que le test de van der Voet pour la comparaison de modèles sera calculé sur chacun des modèles issus de la procédure randomisée de validation croisée ; il est possible de choisir les options suivantes :

PVAL = *nombre*

spécifie la valeur en probabilité du seuil de significativité des différences ; la valeur par défaut est 0,10.

STAT = *statistique_de_test*

spécifie la statistique de test utilisée pour la comparaison de modèles ; on peut indiquer soit **T2** pour la statistique du T^2 de Hotelling, soit **PRESS** pour la somme des carrés des résidus estimés. T2 est la valeur optionnelle prise par défaut

NSAMP = *nombre*

spécifie le nombre de tirages aléatoires à effectuer ; la valeur par défaut est 1 000.

LV = *nombre*

indique le nombre de composantes à extraire ; par défaut, il s'agit du nombre de variables explicatives spécifiées, auquel cas l'analyse est équivalente à une régression des moindres carrés ordinaires des variables expliquées sur les variables explicatives ;

OUTMODEL = *table_SAS*

indique le nom d'une table SAS qui contiendra l'information sur le modèle soumis à l'analyse ;

OUTCVL = *table_SAS*

indique le nom d'une table SAS qui contiendra l'information sur la validation croisée ;

Instruction CLASS

CLASS *variables_nominales* ;

Cette instruction est utilisée pour spécifier les variables nominales utilisées comme variables explicatives pour définir des typologies sur l'ensemble des observations. Le format de ces variables nominales peut être numérique ou alphanumérique. La procédure PLS utilise les valeurs formatées de ces variables nominales pour déterminer les effets correspondant aux groupes dans le modèle. Chaque variable qui ne figure pas dans l'instruction **CLASS** est supposée continue. Les variables continues doivent être codées selon un format numérique.

Instruction MODEL

MODEL *réponses* = *facteurs* < / INTERCEPT > ;

Cette instruction permet de spécifier les *réponses* (variables expliquées) et les *facteurs* (variables explicatives) utilisés pour modéliser leur comportement. Il est possible de simplement lister comme facteurs les variables explicatives, mais on peut également recourir à la syntaxe de spécification de la procédure GLM pour prendre en compte les interactions ou spécifier un polynôme comme facteur. Par défaut, les facteurs sont centrés et le modèle ne comporte pas de terme constant, mais l'option INTERCEPT permet d'inclure un terme constant dans le modèle.

Instruction OUTPUT

OUTPUT

OUT = *table_SAS mots-clés = noms < ... mots-clés = noms >* ;

Cette instruction indique le nom de la table SAS qui contiendra les résultats numériques calculés pour chaque observation, telles que les coordonnées pseudo-factorielles et les différentes valeurs estimées. Sont prévus les mots-clés suivants :

PREDICTED valeurs estimées pour les variables expliquées ;

YRESIDUAL erreurs résiduelles pour les variables expliquées ;

XRESIDUAL erreurs résiduelles pour les variables explicatives ;

XSCORE variables latentes pour les variables explicatives (X-coordonnées pseudo-factorielles, composantes T) ;

YSCORE variables latentes pour les variables explicatives (Y-coordonnées pseudo-factorielles, composantes U) ;

STDY valeurs centrées réduites des Y-estimations ;

STDX valeurs centrées réduites des X-estimations ;

H mesure approchée de l'influence ;

PRESS somme des carrés des erreurs d'estimation ;

T2 somme standardisée des carrés des coordonnées factorielles ;

XQRES somme des carrés des résidus standardisés pour les variables explicatives ;

YQRES somme des carrés des résidus standardisés pour les variables expliquées.

III) Exemples d'utilisation de la procédure PLS

III.1) Modèles de régression d'une réponse univariée

III.1.1) L'ajustement du modèle univarié

Le programme SAS ci-dessous présente un exemple de mise en oeuvre de la procédure PLS pour effectuer une régression des moindres carrés partiels de la « réponse » univariée constituée par l'indice d'octane moteur $Y = \{y\}$ (variable expliquée) sur les « facteurs » que sont les composants $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ d'un mélange de carburants (variables explicatives). L'exemple suivant, extrait de [Cornell, 1990], concerne la composition de douze mélanges de carburants différents ($N = 12$) enregistrée pour déterminer l'influence des sept composants ($M = 7$) suivants sur l'indice d'octane moteur y :

- x_1 Distillation directe
- x_2 Reformat
- x_3 Naphta de craquage thermique
- x_4 Naphta de craquage catalytique
- x_5 Polymère
- x_6 Alkylat
- x_7 Essence naturelle

Le listage du programme SAS permet de visualiser les tableaux de données : il comporte la composition relative en pourcentages de chacun des mélanges selon les composants $x_1 - x_7$, ainsi que son indice d'octane moteur y.

```
data cornell;
input num $ x1 x2 x3 x4 x5 x6 x7 y;
cards;
01    0.00  0.23  0.00  0.00  0.00  0.74  0.03  98.7
02    0.00  0.10  0.00  0.00  0.12  0.74  0.04  97.8
03    0.00  0.00  0.00  0.10  0.12  0.74  0.04  96.6
04    0.00  0.49  0.00  0.00  0.12  0.37  0.02  92.0
05    0.00  0.00  0.00  0.62  0.12  0.18  0.08  86.6
06    0.00  0.62  0.00  0.00  0.00  0.37  0.01  91.2
07    0.17  0.27  0.10  0.38  0.00  0.00  0.08  81.9
08    0.17  0.19  0.10  0.38  0.02  0.06  0.08  83.1
09    0.17  0.21  0.10  0.38  0.00  0.06  0.08  82.4
10    0.17  0.15  0.10  0.38  0.02  0.10  0.08  83.2
11    0.21  0.36  0.12  0.25  0.00  0.00  0.06  81.4
12    0.00  0.00  0.00  0.55  0.00  0.37  0.08  88.1
;
run;
```

L'étude de la matrice des corrélations de l'ensemble de ces variables indique l'existence d'une situation de multicolinéarité au vu des niveaux de corrélation enregistrés. D'une part, il existe une relation linéaire exacte (colinéarité mécanique induite par l'expression en pourcentages de la composition du mélange) entre les variables du tableau X (la somme de chaque ligne du tableau est égale à 1). D'autre part, se manifeste une pseudo-colinéarité relevant des caractéristiques empiriques des observations : la quantité x_3 de naphta de craquage thermique est très directement liée au produit x_1 de la distillation directe. En étudiant les corrélations, on peut déjà identifier au sein du tableau X deux groupes de variables distincts dont les éléments sont positivement corrélés : d'une part l'ensemble $\{x_1, x_3, x_4, x_7\}$, d'autre part l'ensemble $\{x_5, x_6\}$. La variable y est corrélée positivement aux variables du groupe $\{x_5, x_6\}$ et négativement aux variables du groupe $\{x_1, x_3, x_4, x_7\}$. On remarque également que le niveau de corrélation de la variable x_5 avec l'ensemble des autres variables est globalement plus faible.

Tableau T1 : matrice de corrélations entre variables des tableaux X et Y.

	x1	x2	x3	x4	x5	x6	x7	y
x1	1,0000	0,1042	0,9999	0,3707	-0,5480	-0,8046	0,6026	-0,8373
x2		1,0000	0,1008	-0,5369	-0,2926	-0,1913	-0,5900	-0,0708
x3			1,0000	0,3740	-0,5482	-0,8052	0,6071	-0,8380
x4				1,0000	-0,2113	-0,6457	0,9159	-0,7067
x5					1,0000	0,4629	-0,2744	0,4938
x6						1,0000	-0,6564	0,9851
x7							1,0000	-0,7411
y								1,0000

Si la colinéarité mécanique peut être traitée parfois en éliminant une des variables de la composition, la relation de pseudo-colinéarité entre les variables explicatives pose un problème d'estimation des coefficients de la régression de l'indice d'octane selon les composants du mélange, car l'inversion d'une matrice $X'X$ ayant un déterminant voisin de zéro conduit en régression MCO à des estimations numériques perturbées des coefficients dotés alors d'une forte variabilité.

La technique de régression des moindres carrés partiels permet d'éviter que l'estimation des coefficients de régression soit perturbée par cette situation de multicolinéarité. Pour effectuer ce type de régression, nous allons utiliser la procédure PLS de SAS et commenter les résultats de cette analyse. L'algorithme des moindres carrés partiels

extrait des composantes au sein de chacun des espaces de variables associés aux tableaux X et Y qui maximisent la covariance entre les composantes de X (T) et celles de Y (respectivement U).

Pour fixer les notations, rappelons que, d'une manière générale, le modèle de la régression MCP s'écrit :

$$X = TP' + E$$
$$Y = UQ' + F$$

où X : matrice des facteurs Y : matrice des réponses

T : matrice des X-composantes U : matrice des Y-composantes

P : matrice des X-saturations Q : matrice des Y-saturations

E : matrice des X-résidus F : matrice des Y-résidus

Dans le cas d'une réponse univariée, l'algorithme SIMPLS produit des composantes colinéaires à celles issues de la régression des moindres carrés partiels [de Jong, 1993]. C'est donc cette méthode d'extraction des composantes que nous allons utiliser.

III.1.2) Sélection du nombre de pseudo-composantes par validation croisée

La spécification d'un modèle de régression des moindres carrés partiels n'est complète qu'une fois fixé le nombre de pseudo-composantes. Les techniques de validation croisée consistent à diviser l'échantillon en deux ou plusieurs groupes d'individus afin de déterminer ce nombre de pseudo-composantes. L'ajustement du modèle est effectué sur l'ensemble des groupes retenus, formant le jeu d'apprentissage, à l'exception de l'un d'entre-eux, formant l'échantillon-test.

L'efficacité prédictive du modèle est alors évaluée sur les individus du groupe écarté de l'estimation, l'échantillon-test. En répétant ce processus pour chacun des groupes, on peut mesurer l'efficacité prédictive globale d'une spécification particulière du modèle à l'aide de l'indicateur PRESS[7] - somme des carrés des erreurs de prédiction -, statistique basée sur les résidus issus de ce processus itératif.

```
proc pls data=cornell method=pls cv=split;
 model y = x1 - x7;
run;
```

[7] pour « *Predicted REsidual Sum of Squares* »

Les paramètres de la procédure `PLS` indiquent le nom de la table SAS des données (`data=cornell`) et la méthode d'extraction des composantes (`method=pls`). L'instruction `model` permet de spécifier les facteurs utilisés pour modéliser la variable expliquée (`y=x1-x7`).

Avec l'instruction `proc pls`, on a la possibilité de spécifier l'argument de l'option CV= pour indiquer la méthode de validation croisée utilisée afin de déterminer le nombre de pseudo-composantes à retenir. Une des méthodes de validation croisée les plus courantes consiste à exclure une observation toutes les *n* observations (« *split-sample validation* »). On peut utiliser cette méthode en spécifiant CV=SPLIT, avec pour valeur par défaut *n=1* (chaque observation constitue un échantillon-test), comme dans les instructions précédentes.

Les résultats de la procédure de validation croisée montrent que la valeur absolue minimale du PRESS est atteinte avec 5 pseudo-composantes. On note cependant que la valeur atteinte par le PRESS avec 4 pseudo-composantes n'est pas beaucoup plus élevée.

Tableau T2 : résultats initiaux de la validation croisée.

CORNELL, indice d'octane moteur
Regression MCP : validation croisee

The PLS Procedure
Cross Validation for the Number of Latent Variables

Number of Latent Variables	Root Mean PRESS
0	1.0397
1	0.3449
2	0.3148
3	0.2495
4	0.1880
5	0.1843
6	0.3422
7	0.3422

Minimum Root Mean PRESS = 0.184292 for 5 latent variables

Le tableau suivant propose alors le tableau des pourcentages de variance expliquée par le modèle à 5 pseudo-composantes :

Tableau T3 : pourcentages de variance expliquée par le modèle à 5 pseudo-composantes.

CORNELL, indice d'octane moteur
Regression MCP : validation croisee

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	57.3606	57.3606	92.3594	92.3594
2	15.2521	72.6128	5.2749	97.6343
3	19.2130	91.8258	1.4212	99.0556
4	7.9271	99.7528	0.0233	99.0788
5	0.2460	99.9988	0.1683	99.2471

Permettant de tester si cette différence est significative, une statistique a été proposée par [van der Voet, 1994]. Pour calculer cette statistique, il suffit de spécifier l'option CVTEST, à l'instar de l'extrait de programme SAS, ci-dessous :

```
proc pls data=cornell method=pls cv=split cvtest;
   model y = x1 - x7;
 run;
```

ce qui nous donne le tableau de résultats suivant :

Tableau T4 : test de comparaison des modèles MCP.

Regression MCP : validation croisee

The PLS Procedure
Cross Validation for the Number of Latent Variables

Number of Latent Variables	Root Mean PRESS	Test for larger residuals than minimum T2	Prob > T2
0	1.0397	6.9251	0.00100
1	0.3449	2.5711	0.0950
2	0.3148	3.5234	0.0330
3	0.2495	1.0041	0.3990
4	0.1880	0.0254	0.9130
5	0.1843	0	1.0000
6	0.3422	2.3514	0.0160
7	0.3422	2.3514	0.0160

Minimum Root Mean PRESS = 0.184292 for 5 latent variables
Smallest model with p-value > 0.1: 3 latent variables

Ce test de comparaison de modèle s'effectue à partir d'une procédure de randomisation des données. Par défaut, le germe utilisé pour cette randomisation est généré par l'horloge du système d'exploitation. Les résultats de cette procédure figurent dans le tableau suivant. Le modèle minimal avec une probabilité critique $p > 0.1$ est le modèle MCP à 3 pseudo-composantes. La différence avec les modèles à 4 ou 5 pseudo-composantes n'est pas significative. Ainsi dans une seconde étape d'estimation du modèle, ultérieure à la validation croisée, on utilise un modèle MCP à 3 composantes en exécutant les instructions suivantes :

```
proc pls data=cornell method=simpls lv=3 outmodel=estim;
model y=x1-x7;
output out=outmcp xscore=t yscore=u pred=yest yr=yres;
run;
```

La syntaxe de l'instruction `proc pls` permet alors de spécifier le nombre de composantes à extraire (`lv=3`) et le fichier des paramètres du modèle (`outmodel=estim`). L'instruction **OUTPUT** indique les éléments d'information (`xscore=t pred=p yr=e`) que contiendra le fichier des résultats (`out=outmcp`).

III.1.3) L'interprétation des résultats

En premier lieu, la procédure PLS produit un tableau permettant de vérifier comment chaque composante MCP extraite rend compte de la variabilité des facteurs (`Model Effects`) et de la réponse (`Dependent Variables`) au sein du modèle. Ainsi, on peut constater que les trois premières composantes MCP extraites représentent près de 92 % de la variabilité des composants du mélange de carburant (`Model Effects`) et expliquent plus de 99 % de la variabilité de l'indice d'octane moteur (`Dependent Variables`) :

Tableau T5 : parts de variance expliquée par le modèle à 3 composantes MCP.

CORNELL, indice d'octane moteur
Modèle de régression MCP à 3 composantes

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	57.3606	57.3606	92.3594	92.3594
2	15.2521	72.6128	5.2749	97.6343
3	19.2130	91.8258	1.4212	99.0556

Ainsi, le tableau I5 permet de connaître directement les parts de variance expliquée par les composantes MCP :

• dans le modèle à une composante MCP $\quad X_j^* = t_1 p_{1j}' + E_{1j} \quad j = 1,\ldots,7$
$Y^* = u_1 q_1' + F_1$

• la part de variance de Y expliquée par la première X-composante t_1 est égale à :

$R^2(Y,t_1) = 0{,}9236$

• la part de variance de X expliquée par t_1 vaut : $\dfrac{\|t_1\|^2 \times \|p_1\|^2}{(N-1)M} \approx 0{,}5736$

• dans le modèle à deux composantes MCP $\quad X_j^* = t_1 p_{1j}' + t_2 p_{2j}' + E_{2j} \quad j = 1,\ldots,7$
$Y^* = u_1 q_1' + u_2 q_2' + F_2$

• la part de variance de Y expliquée par la seconde X-composante t_2 est égale à :

$R^2(Y,t_2) \approx 0{,}0527$

• la part de variance de Y expliquée par le plan $(t_1 \oplus t_2)$ est égale à :

$R^2(Y,t_1 \oplus t_2) \approx 0{,}9763$

• la part de variance de X expliquée par t_2 vaut : $\dfrac{\|t_2\|^2 \times \|p_2\|^2}{(N-1)M} \approx 0{,}1525$

• la part de variance de X expliquée par le plan $(t_1 \oplus t_2)$ est égale à :

$$\frac{\|t_1\|^2 \times \|p_1\|^2 + \|t_2\|^2 \times \|p_2\|^2}{(N-1)M} \approx 0{,}7261$$

• dans le modèle à trois composantes MCP $\quad X_j^* = t_1 p_{1j}' + t_2 p_{2j}' + t_3 p_{3j}' + E_{3j} \quad j = 1,\ldots,7$
$Y^* = u_1 q_1' + u_2 q_2' + u_3 q_3' + F_3$

• la part de variance de Y expliquée par la troisième X-composante t_3 est égale à :

$R^2(Y,t_3) = 0{,}0142$

• la part de variance de Y expliquée par le sous-espace $(t_1 \oplus t_2 \oplus t_3)$ est égale à :

$R^2(Y,t_1 \oplus t_2 \oplus t_3) \approx 0{,}9906$

• la part de variance de X expliquée par t_3 vaut : $\dfrac{\|t_3\|^2 \times \|p_3\|^2}{(N-1)M} \approx 0{,}1921$

• la part de variance de X expliquée par le sous-espace $(t_1 \oplus t_2 \oplus t_3)$ est égale à :

$$\frac{\|t_1\|^2 \times \|p_1\|^2 + \|t_2\|^2 \times \|p_2\|^2 + \|t_3\|^2 \times \|p_3\|^2}{(N-1)M} \approx 0{,}9183$$

La procédure PLS produit également deux tables SAS : la première contient le fichier des paramètres du modèle (`estim`), la seconde le fichier des résultats (`outmcp`).

En utilisant le macroprogramme (`%prt_scr`) :

```
* Impression des &lv composantes MCP, en X et en Y                    *;
 title3 "Composantes MCP: projections dans l'espace des observations";
 %prt_scr(outmcp);
```

on obtient les projections dans l'espace des observations des deux ensembles de composantes : celles de X (t_h) et celles de Y (respectivement u_h)[8] :

Tableau T6 : projection des observations sur les composantes MCP.

CORNELL, indice d'octane moteur
Modèle de régression MCP univarié à 3 composantes
Composantes MCP: projections des observations

OBS	T1	T2	T3	U1	U2	U3
1	2.06171	1.00975	1.58621	1.55133	0.56252	0.33808
2	2.48720	0.79721	0.10968	1.41332	0.22045	0.04325
3	2.34290	1.13865	-0.17338	1.22931	0.10564	-0.14745
4	2.04748	-1.96069	-0.50286	0.52393	-0.45806	-0.02224
5	-0.06845	-0.26763	-2.96368	-0.30413	-0.27130	-0.21181
6	1.62199	-1.74815	0.97367	0.40125	-0.37667	0.01191
7	-2.21542	-0.21885	0.23815	-1.02485	0.03768	0.08633
8	-2.00365	0.12361	0.11873	-0.84084	0.12012	0.09265
9	-2.09817	0.18257	0.35554	-0.94818	0.05812	0.01754
10	-1.92548	0.39123	0.19699	-0.82550	0.09797	0.01101
11	-2.08681	-0.56599	1.03390	-1.10152	-0.10068	0.02513
12	-0.16329	1.11830	-0.97296	-0.07412	0.00420	-0.24437

Dans le cadre du modèle de régression MCP univarié, les variable u_h pour $h > 1$ correspondent à des résidus normalisés et ne sont donc pas directement interprétables.

Pour qu'un modèle PLS soit satisfaisant, les premiers facteurs doivent présenter une forte corrélation entre les deux ensembles de composantes : celles de X et celles de Y (respectivement U).

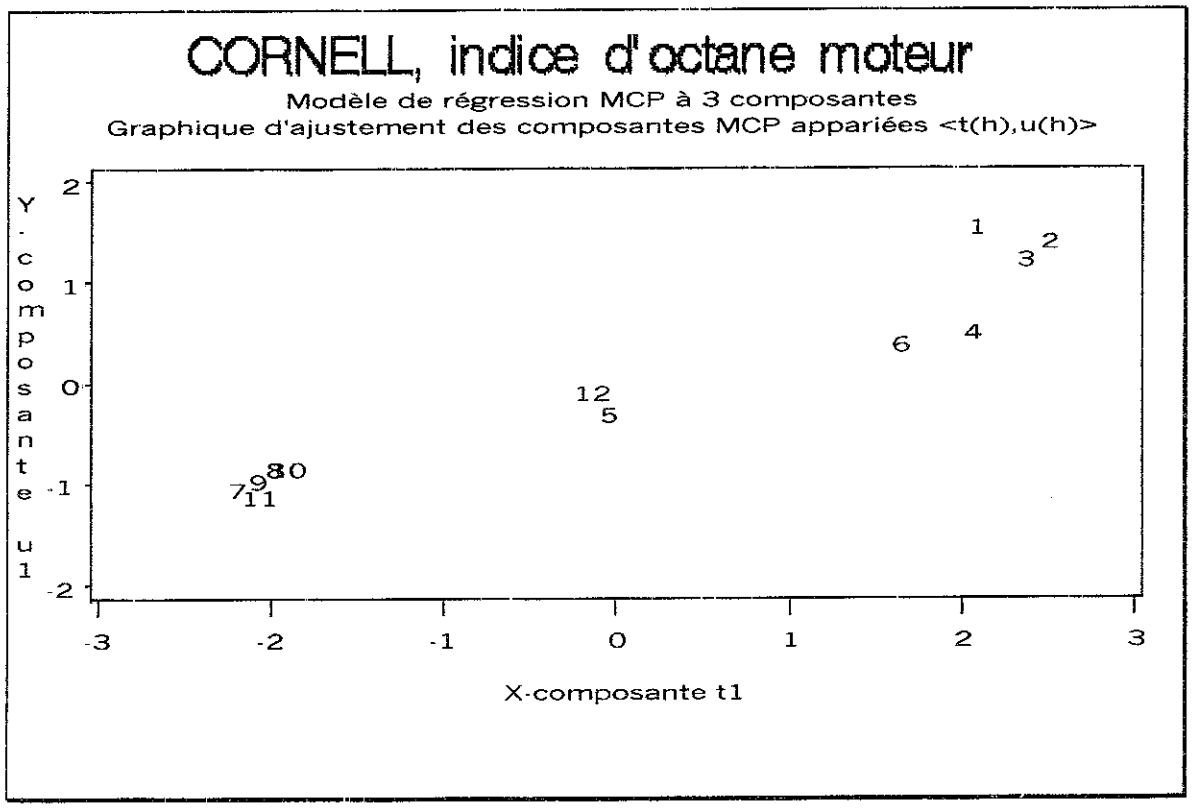
[8] Les composantes u_h extraites par la procédure PLS de SAS sont normées. On obtient les mêmes directions que dans la régression MCP d'une réponse univariée aux facteurs de colinéarité près introduits par l'utilisation de l'algorithme SIMPLS.

Pour le vérifier, il suffit de tracer le diagramme croisant les composantes T de X et U de Y, deux à deux, en utilisant le macroprogramme suivant[9] :

```
* Diagramme croisant les X-composantes avec les Y-composantes          *;
 title3 "Graphique d'ajustement des composantes MCP appariées <t(h),u(h)>";
 %plot_scr(outmcp);
```

Les diagrammes correspondants apparaissent en figures F1 et F2, respectivement.

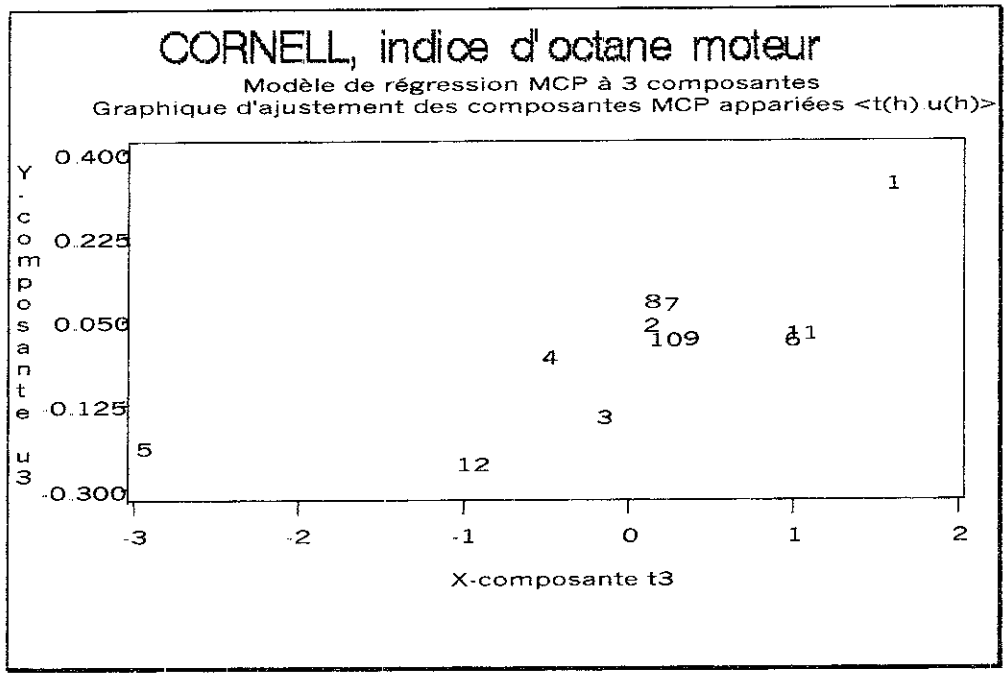
Figure F1 : ajustement des composantes appariées t_1 et u_1.



Les points correspondent aux observations représentées par leur numéro d'ordre. Sur le premier graphique $(t_1 \times u_1)$, on perçoit nettement la structure en trois groupes des observations.

[9] Les macroprogrammes SAS d'aide à l'interprétation des résultats de la procédure PLS seront publiés dans le prochain numéro.

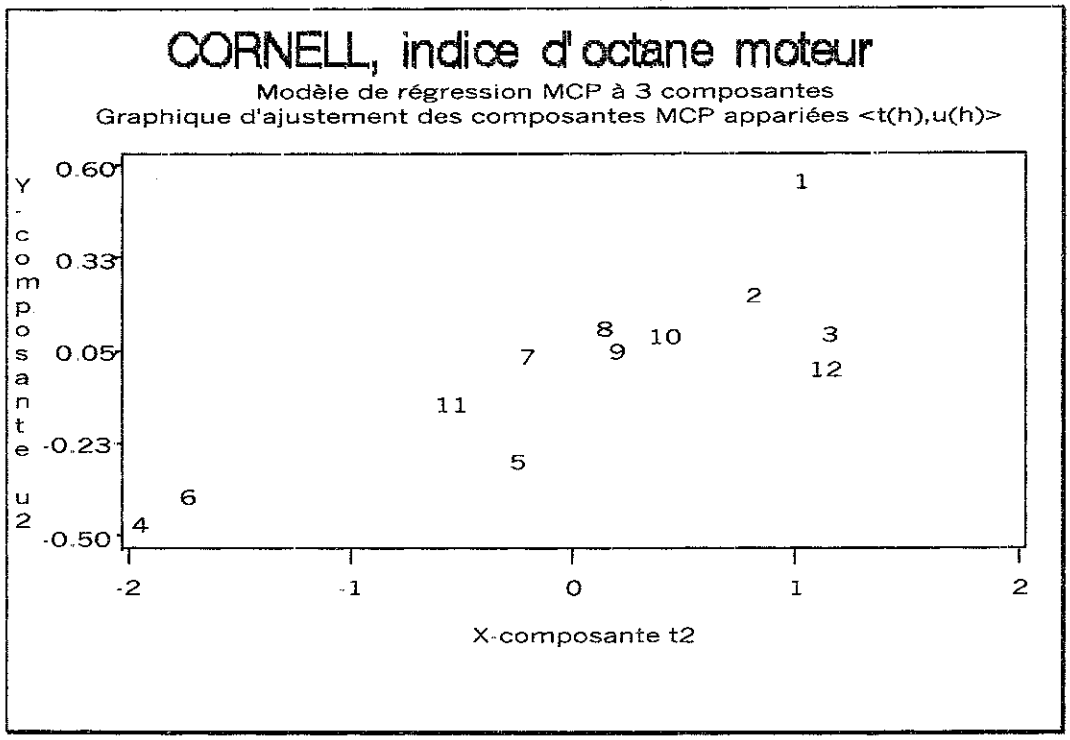
Figure F2 : ajustement des composantes appariées t_2 et u_2.



On remarque dans cet exemple que les diagrammes montrent une grande corrélation entre les composantes T et U, en particulier pour la première composante MCP.

Mais la qualité de cette liaison semble se détériorer au fur et à mesure du processus d'extraction. C'est ainsi que la liaison linéaire est moins évidente entre les X-projections et les Y-projections sur le diagramme $(t_3 \times u_3)$ concernant la troisième composante MCP.

Figure F3 : ajustement des composantes appariées t_3 et u_3.



Une mesure de la qualité de la représentation des variables du tableau X et de la réponse Y par le modèle est fourni par leurs corrélations avec les X-composantes MCP.

```
* Correlation des X et Y-variables par les X-composantes                    *;
  title3 "Corrélations des X et Y-variables avec les X-composantes t(h)";
  %prt_corr(outmcp,lvar=&xvars &yvars,cvar=t1-t&lv,dscorr=corrtab);
```

La macro-instruction qui précède permet d'en imprimer le tableau :

Tableau T7 : corrélations des X et Y-variables avec les X-composantes MCP.

CORNELL, indice d'octane moteur
Modèle de régression MCP univarié à 3 composantes
Qualité de représentation des tableaux X et Y par les X-composantes t(h)

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 12

	T1	T2	T3
X1	**-0.90426**	-0.03575	0.31573
Distillation directe	0.0001	0.9122	0.3174
X2	0.06317	**-0.84366**	0.51971
Reformat	0.8454	0.0006	0.0833
X3	**-0.90586**	-0.03279	0.31315
Naphta de craquage thermique	0.0001	0.9194	0.3216
X4	**-0.70985**	0.23384	**-0.61670**
Naphta de craquage catalytique	0.0097	0.4645	0.0327
X5	0.58676	-0.03821	-0.57289
Polymere	0.0449	0.9061	0.0515
X6	**0.92103**	0.36964	0.12192
Alkylat	0.0001	0.2370	0.7058
X7	**-0.82249**	0.40095	-0.39202
Essence naturelle	0.0010	0.1965	0.2075
Y	**0.96104**	0.22967	0.11922
Indice d'octane moteur	0.0001	0.4727	0.7121

Soit $\lambda_h = \mathrm{var}(t_h)$, la variance des X-composantes, on peut remarquer qu'on a l'égalité suivante :

$$corr(X,t_h) = \sqrt{\lambda_h}\cdot p_h \qquad corr(Y,t_h) = \sqrt{\lambda_h}\cdot r_h$$

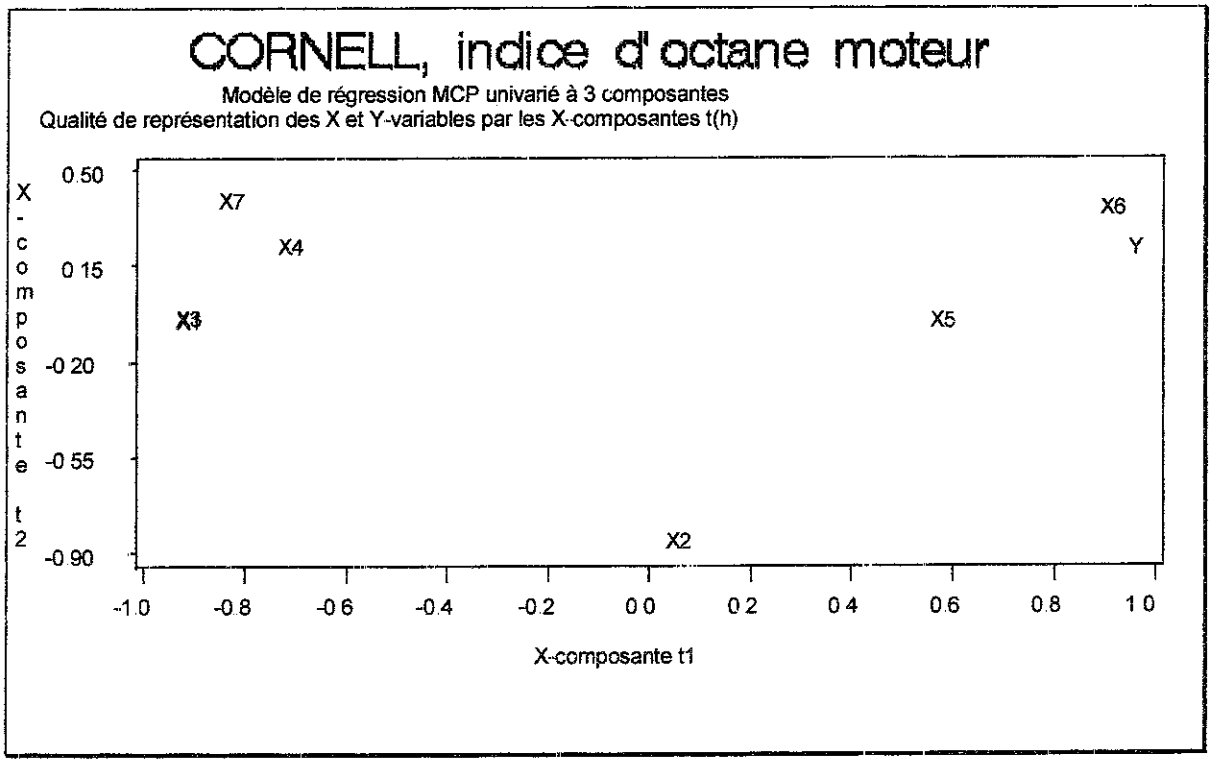
où p_h, r_h sont définis par $E_{h-1} = t_h \cdot p_h' + E_h$, $F_{h-1} = t_h \cdot r_h + F_h$

Afin de juger globalement de la qualité de la représentation des variables du tableau X et de la réponse Y, on peut tracer le graphique des corrélations à l'aide du macroprogramme suivant :

```
* Graphique des corrélations des X et Y -variables avec les X-composantes        *;
title3 "Qualité de représentation des X et Y-variables par les X-composantes t(h)";
%pltcorr(corrtab,max_lv=&lv);
```

ce qui permet d'obtenir la représentation ci-après :

Figure F4 : qualité de la représentation des X et Y-variables par les X-composantes.



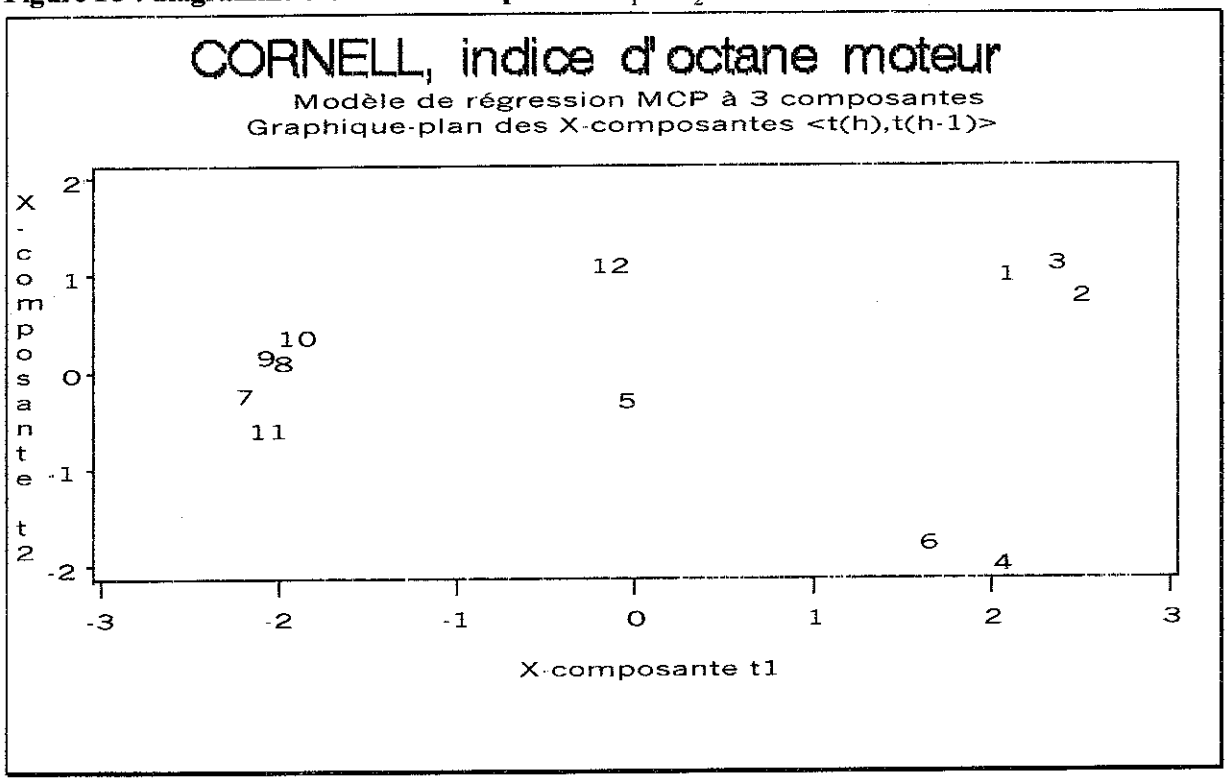
Bien qu'un cercle des corrélations pourrait en rendre l'interprétation plus aisée, ce graphique illustre néanmoins les liaisons internes aux variables du tableau X et celles entre les variables du tableau X et la réponse Y, telles que l'examen de la matrice des corrélations les a révélées : juxtaposition des points x_1 et x_3 ; corrélations positives internes aux groupes de variables $\{x_1, x_3, x_4, x_7\}$ et $\{x_5, x_6\}$; y corrélée positivement au couple $\{x_5, x_6\}$ et négativement aux variables du groupe $\{x_1, x_3, x_4, x_7\}$. La proximité des projections des variables x_6 et y laisse penser que la variable x_6 est un prédicteur important pour l'ajustement du modèle.

On peut également projeter les composantes T l'une vis à vis de la suivante pour détecter des irrégularités : par exemple des points influents dont la présence où l'absence peut déterminer des variations importantes dans l'estimation des coefficients du modèle. Pour ce faire, il suffit d'invoquer le macroprogramme suivant :

```
* Diagramme croisant chaque composante X avec la precedente                *;
 title3 "Graphique-plan des X-composantes <t(h),t(h-1)>";
 %plotxscr(outmcp);
```

On obtient ainsi des graphiques croisant les X-projections sur chacune des composantes croisées avec les X-projections de la composante précédente

Figure F5 : diagramme croisant les composantes t_1 et t_2.



Ce graphique-plan $(t_1 \times t_2)$ est l'équivalent d'un premier plan principal du tableau X mais orienté vers l'explication de la réponse y. Il permet de visualiser les projections des douze mélanges de carburant correspondant aux observations. On distingue clairement une possible typologie en trois classes de carburants : $\{1,2,3,4,6\}, \{5,12\}, \{7,8,9,10,11\}$, qui pourrait être effectuée sur la base de leur indice d'octane moteur puisque la composante MCP t_1 est reliée à la composante u_1 correspondant à la valeur centrée réduite y^* de l'indice d'octane. On peut également rechercher des motifs particuliers de configuration des

observations. Un motif incurvé laissera suspecter la présence d'un terme quadratique à ajouter aux spécifications du modèle.

Pour imprimer les tableaux et produire les diagrammes permettant de visualiser la distribution des contributions pour l'ensemble X, il suffit d'exécuter les appels aux macroprogrammes suivants :

```
* Calcul des X-contributions pour chaque composante MCP                    *;
 title3 "Tableau des X-contributions w(h)";
 %get_wts(estim,dsxwts=xwts);

* Diagramme des X-contributions pour les max_lv lieres composante MCP      *;
 title3 "Graphique-plan des X-contributions <w(h),w(h-1)>";
 %plot_wt(xwts,max_lv=&lv);
```

On obtient ainsi tout d'abord le tableau des X-contributions w_h :

Tableau T8 : tableau des X-contributions.

CORNELL, indice d'octane moteur
Modèle de régression MCP à 3 composantes
Tableau des X-contributions w(h)

X_VAR	_LABEL_	W1	W2	W3
X1	Distillation directe	-0.43921	0.20190	0.29046
X2	Reformat	-0.03715	-0.83675	0.45251
X3	Naphta de craquage thermique	-0.43956	0.20746	0.29188
X4	Naphta de craquage catalytique	-0.37071	-0.15598	-0.56878
X5	Polymere	0.25903	-0.44595	-0.44589
X6	Alkylat	0.51673	0.63567	0.10926
X7	Essence naturelle	-0.38876	0.31381	-0.31091

La contribution w_1 peut se calculer directement par la formule :

$$w_1 = \frac{1}{\sqrt{\sum_{j=1}^{M} cor^2(X_j, Y)}} \begin{bmatrix} cor(X_1, Y) \\ \vdots \\ cor(X_M, Y) \end{bmatrix}$$

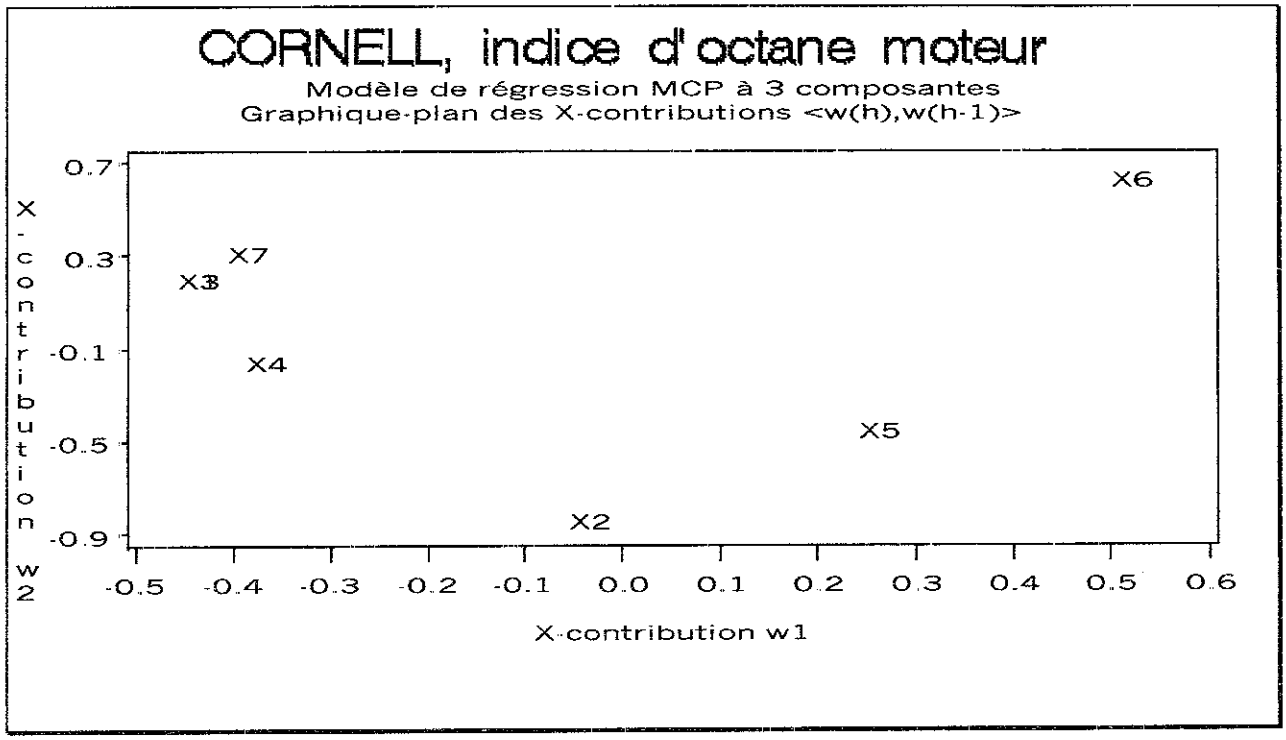
Les X-contributions w_h sont les coefficients de la régression de la composante t_1 sur le tableau X centré-réduit. Ainsi, la composante t_1 s'écrit : $t_1 = E_0 w_1$, c'est à dire :

$t_1 \approx -0,439 \times E_{01} - 0,037 \times E_{02} - 0,440 \times E_{03} - 0,371 \times E_{04} + 0,259 \times E_{05} + 0,517 \times E_{06} - 0,389 \times E_{07}$

où les E_{0j} sont les variables centrées-réduites x_j^*.

Puis, viennent les diagrammes des X-contributions présentées deux à deux :

Figure F6 : diagramme des X-contributions w_1 et w_2.



À l'opposé l'un de l'autre sur le premier axe de ce graphique, on retrouve les deux groupes de variables identifiés lors de l'examen de la matrice des corrélations.

On obtient également le tableau des X-saturations et le graphique correspondant en exécutant les macros-instructions suivantes :

```
* Calcul des X-saturations pour chaque composante MCP                    *;
 title3 "Tableau des X-saturations p(h)";
 %getxload(estim,dsxload=xloads);

* Diagramme des X-saturations pour les max_lv lieres composante MCP      *;
 title3 "Graphique-plan des X-saturations <p(h),p(h-1)>";
 %pltxload(xloads,max_lv=&lv);
```

dont voici le résultat :

Tableau T9 : tableau des X-saturations.

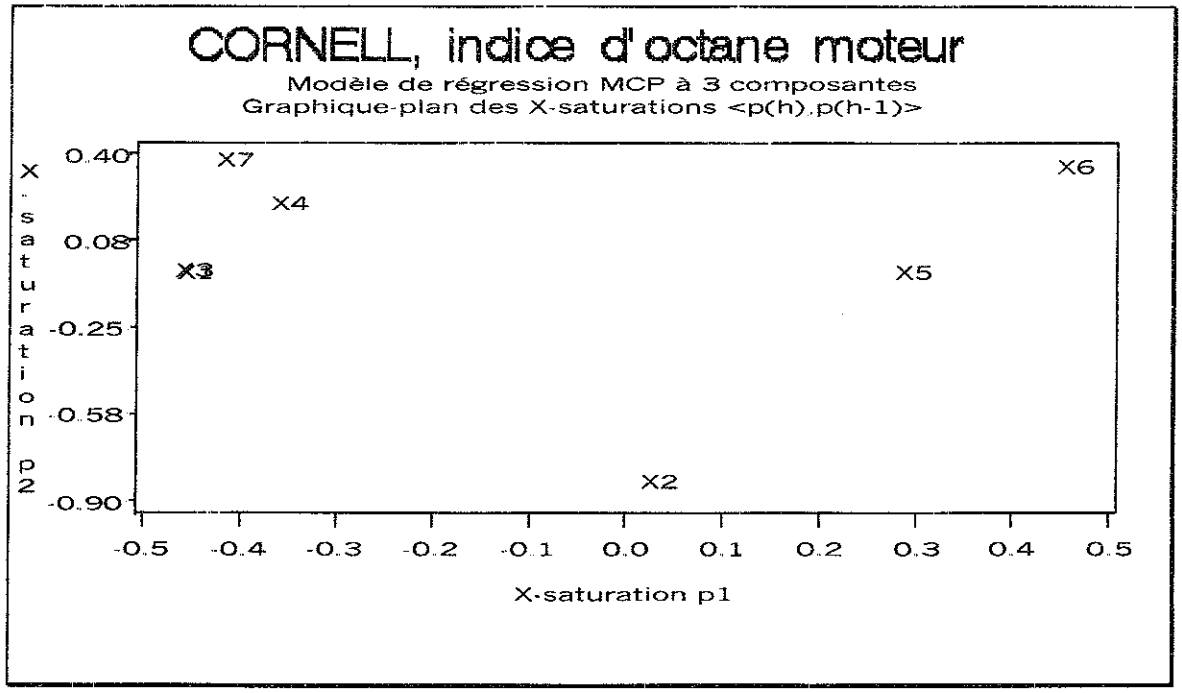
CORNELL, indice d'octane moteur
Modèle de régression MCP à 3 composantes
Tableau des X-saturations p(h)

X_VAR	_LABEL_	P1	P2	P3
X1	Distillation directe	-0.45127	-0.03460	0.27225
X2	Reformat	0.03153	-0.81649	0.44814
X3	Naphta de craquage thermique	-0.45207	-0.03174	0.27002
X4	Naphta de craquage catalytique	-0.35425	0.22631	-0.53177
X5	Polymere	0.29282	-0.03698	-0.49400
X6	Alkylat	0.45964	0.35774	0.10513
X7	Essence naturelle	-0.41046	0.38804	-0.33804

Issues de la décomposition $E_{h-1} = t_h p'_h + E_h$, les X-saturations $p_{h,j}$ sont les coefficients de régression des $E_{h-1,j}$ sur t_h : $E_{0,j} = p_{1,j} t_1 + E_{1,j}$ pour $j = 1, \cdots, 7$. Ainsi, on a par exemple $E_{01} \approx -0{,}451 \times t_1 + E_{11}$

Le diagramme correspondant des X-saturations est en général très semblable à celui des X-contributions :

Figure F7 : diagramme des X-saturations p_1 et p_2.



Les diagrammes de résidus et des quantiles normalisés des résidus permettent de détecter les points atypiques (« *outliers* ») qui pourraient nuire à la qualité de l'ajustement. Ces diagrammes permettent de détecter également les situations d'anormalité,

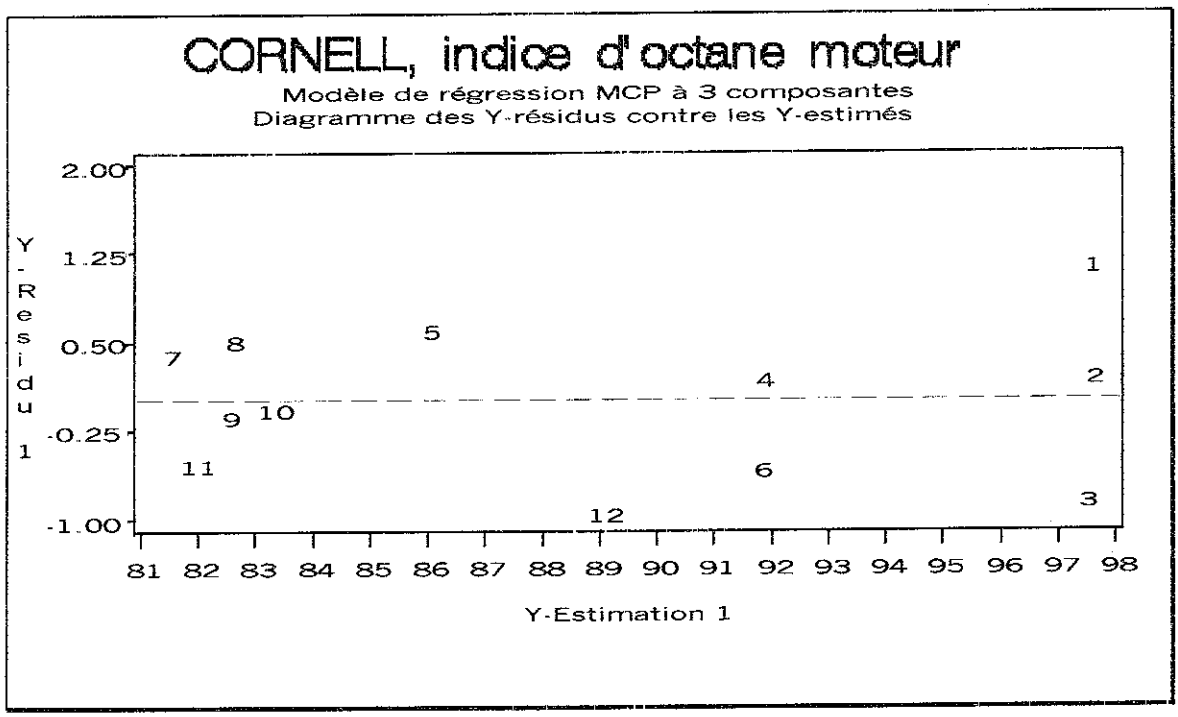
d'autocorrélation et d'hétéroscédasticité qui toutes peuvent être à l'origine de différents problèmes dans la construction d'intervalles de confiance et de tolérance pour les prévisions. Le diagramme idéal des résidus ressemble à un nuage rectangulaire avec une majorité de points situés dans le tiers médian vertical du graphique. Dans un diagramme de normalité idéal, les points sont distribués selon une ligne droite. Pour produire le diagramme des résidus pour chacune des variables expliquées, utilisez la macro-instruction `%res_plot`, et pour un diagramme des quantiles normaux respectivement la macro-instruction `%nor_plot`.

```
* Diagramme des Y-residus contre les Y-estimes                    *;
 title3 "Diagramme des Y-résidus contre les Y-estimés";
 %res_plot(outmcp);

* Droite de Henry: Y-residus contre les scores normaux empiriques Z    *;
 title3 "Droite de Henry: Y-résidus contre scores normaux empiriques";
 %nor_plot(outmcp);
```

Le diagramme des Y-résidus contre les Y-estimés permet de juger de la pertinence de la spécification du modèle :

Figure F8 : diagramme des Y-résidus.

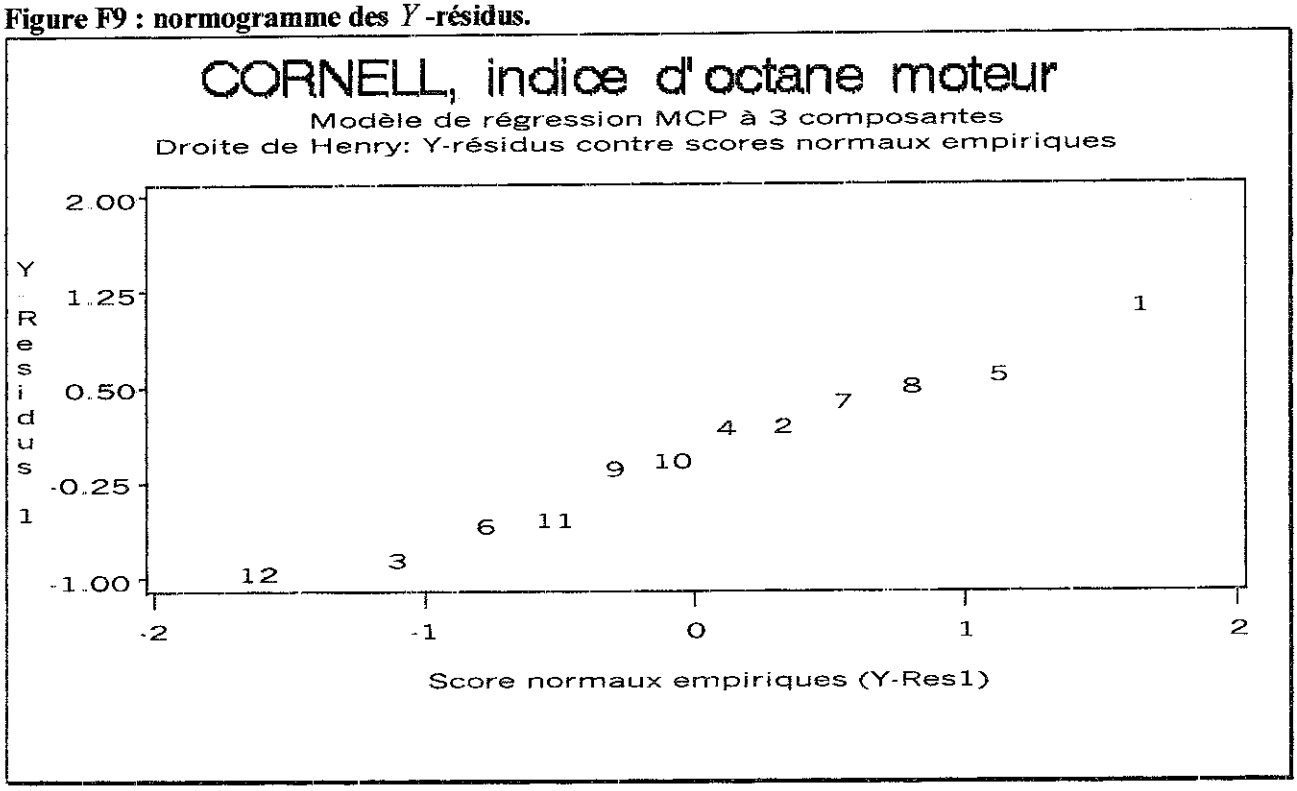


Le diagramme des résidus normalisés permet de contrôler la présence de points atypiques. Ce diagramme des résidus normalisés, ou normogramme, croise les résidus de l'estimation en Y réalisée par le modèle en ordonnée avec les scores normaux empiriques de

ces résidus en abscisse, selon le principe de la Droite de Henry, pour fournir un test graphique de la normalité des résidus.

Le calcul des $e_y^N(i)$, scores normaux empiriques d'ordre N, est effectué selon l'approximation de Blom : $e_y^N(i) = \Phi^{-1}\left(\frac{R(i) - 3/8}{N + 1/4}\right)$ où $R(i)$ est le rang du résidu i et Φ^{-1} l'inverse de la fonction de répartition de la loi normale centrée réduite.

Figure F9 : normogramme des Y-résidus.



La détection de points aberrants peut également se faire en calculant des indices de distance au modèle, soit dans l'espace des X et dans l'espace des Y :

```
* Distances des observations au modele dans l espace des X et des Y     *;
 title3 "Distance des observations au modèle:";
 %get_dmod(outmcp,dsdmod=distmod,id=n);
```

Ce macroprogramme (`%get_dmod`) imprime le tableau des distances au modèle des observations i dans l'espace des X (colonne `DMODX`) et dans l'espace des Y (colonne `DMODY`), ainsi que les diagrammes correspondants.

Tableau T10 : distance des observations au modèle en X et en Y.

CORNELL, indice d'octane moteur
Modèle de régression MCP univarié à 3 composantes
Distance des observations au modèle:
tableau

OBS	DMODX	DMODY
1	0.024894	0.40353
2	0.034676	0.07371
3	0.026940	0.29887
4	0.020872	0.06791
5	0.006717	0.21407
6	0.038532	0.20332
7	0.005624	0.14259
8	0.011610	0.18547
9	0.003334	0.04384
10	0.012052	0.02131
11	0.015605	0.18711
12	0.052551	0.33282

La distance au modèle de régression MCP pour une observation i est définie à partir de l'écart entre les valeurs observées et les valeurs estimées par le modèle de la façon suivante[10] :

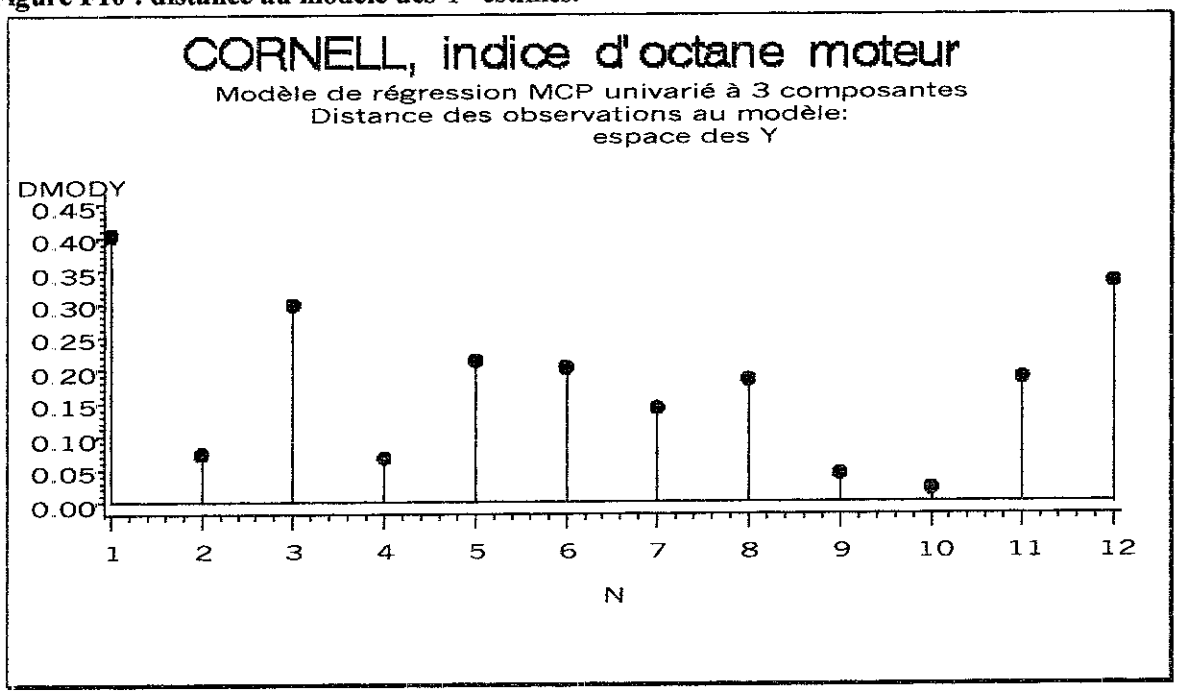
- dans l'espace des X : soit e_{ij} le résidu en l'observation i de la régression de la variable x_j sur les H composantes MCP retenues t_h dans le modèle, alors la distance au modèle de l'observation i est définie par : $$(DModX)_i = \sqrt{\frac{\sum_{j=1}^{M} e_{ij}^2}{N-H-1}}$$

- dans l'espace Y : soit f_i le résidu en l'observation i de la régression de la variable y sur les H composantes MCP retenues t_h dans le modèle, alors la distance au modèle pour l'observation i est définie par : $$(DModY)_i = \frac{|f_i|}{\sqrt{N-H-1}}$$

[10] Le logiciel SIMCA introduit dans cette définition de la distance au modèle un terme correctif et une normalisation pour aboutir à une procédure de validation empirique. Ainsi les distances normalisées au modèle suivraient approximativement une loi de Fisher-Snedecor et permettraient de valider ou de rejeter les estimations ponctuelles par rapport à un seuil critique.

Le diagramme dans l'espace des Y permet de vérifier quelles sont les observations qui s'avèrent les moins bien estimées par le modèle :

Figure F10 : distance au modèle des Y-estimés.



Pour déterminer quels sont les facteurs à exclure du modèle, vous pouvez consulter les coefficients de régression (matrice B) et les indices VIP (importance variable de la projection) pour chacun des facteurs. Les coefficients de régression représentent l'importance de chacun des facteurs dans l'estimation de la réponse. L'indice VIP représente la contribution de chaque facteur à l'ajustement du modèle, à la fois pour les facteurs et pour les réponses. Si un facteur a un coefficient de régression relativement petit en valeur absolue et si son indice VIP a une valeur relativement faible (moins de 0,8) alors la variable explicative peut être exclue du modèle. Les macro-instructions suivantes permettent de produire le tableau des coefficients de régression et des indices VIP.

```
* Calcul des coefficients de regression MCP (matrice B)                    *;
 title3 "Régression MCP: modèle sur variables centrées-réduites";
 %get_bpls(estim,dsout=bmcp);

* Calcul des indices VIP de contribution a la projection                   *;
 title3 "Indices VIP: importance de la contribution à la projection";
 %get_vip(estim,dsvip=indvip);
```

L'exécution des macroprogrammes produit le tableau suivant :

Tableau T11 : tableau des coefficients de régression et des indices VIP.

CORNELL, indice d'octane moteur
Modèle de régression MCP à 3 composantes
Indices VIP: importance de la contribution à la projection
Tableau des coefficients de régression et des indices VIP

X_VAR	B1	VIP
X1	-0.13909	1.13258
X2	-0.20869	0.53904
X3	-0.13756	1.13387
X4	-0.29317	0.96878
X5	-0.03843	0.72939
X6	0.45639	1.37642
X7	-0.14338	1.01629

Ainsi, l'équation de régression MCP s'écrit :

$\hat{y}^* \approx -0,1391x^*_1 - 0,2087x^*_2 - 0,1376x^*_3 - 0,2932x^*_4 - 0,0384x^*_5 + 0,4564^* x_6 + 0,1434^* x_7$

où $\hat{y}^*$ et x^*_j sont les variables centrées-réduites correspondant respectivement à $\hat{y}$ et x_j.

Les valeurs des coefficients de régression et des indices VIP indiquent d'une part l'influence déterminante de la variable x_6 sur l'estimation et d'autre part la faible contribution de la variable x_5. Remarquons également que les signes des coefficients de régression sont cohérents avec les liaisons décelées par l'examen de la matrice des corrélations.

III.2) Modèles de régression d'une réponse multivariée

III.2.1) L'ajustement du modèle multivarié

L'exemple suivant, extrait de [Jackson, 1991], concerne les résultats obtenus par les usagers ($N = 20$) d'un club de gymnastique à trois exercices physiques différents mis en relation avec trois paramètres physiques mesurés sur ces usagers. Il s'agit dans cette étude de relier les résultats des exercices physiques aux paramètres mesurés. L'analyse statistique suivante présente un exemple de mise en oeuvre de la procédure PLS pour effectuer une régression des moindres carrés partiels de la « réponse » multivariée constituée par le tableau $Y = \{y_1, y_2, y_3\}$ des résultats à trois types d'exercice physique (variables expliquées), sur les « facteurs » que sont les trois paramètres physiques $X = \{x_1, x_2, x_3\}$ (variables explicatives) :

- x_1 Poids
- x_2 Tour de taille
- x_3 Pouls
- y_1 Tractions à la barre fixe
- y_2 Flexions
- y_3 Sauts

L'extrait suivant du programme SAS permet de visualiser les données recueillies :

```
data linnerud;
  input num $ x1 x2 x3 y1 y2 y3;
   label x1='Poids' x2='Tour de taille' x3='Pouls'
         y1='Tractions' y2='Flexions' y3='Sauts';
  cards;
01 191     36      50      5       162     60
02 189     37      52      2       110     60
03 193     38      58      12      101     101
04 162     35      62      12      105     37
05 189     35      46      13      155     58
06 182     36      56      4       101     42
07 211     38      56      8       101     38
08 167     34      60      6       125     40
09 176     31      74      15      200     40
10 154     33      56      17      251     250
11 169     34      50      17      120     38
12 166     33      52      13      210     115
13 154     34      64      14      215     105
14 247     46      50      1       50      50
15 193     36      46      6       70      31
16 202     37      62      12      210     120
17 176     37      54      4       60      25
18 157     32      52      11      230     80
19 156     33      54      15      225     73
20 138     33      68      2       110     43
     ;
 run;
```

dont voici les principaux paramètres de tendance centrale et de dispersion :

Tableau T12 : indicateurs de tendance centrale et de dispersion.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Tableau X: indicateurs de tendance centrale et dispersion

X_VAR	MEAN	STD_DEV
X1	178.6	24.6905
X2	35.4	3.2020
X3	56.1	7.2104

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Tableau Y: indicateurs de tendance centrale et dispersion

Y_VAR	MEAN	STD_DEV
Y1	9.45	5.2863
Y2	145.55	62.5666
Y3	70.30	51.2775

L'étude de la matrice des corrélations de l'ensemble de ces variables montre les corrélations suivantes :

Tableau T13 : matrice des corrélations.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Matrice des corrélations des variables d'origine

Correlation Analysis

6 'WITH' Variables: X1 X2 X3 Y1 Y2 Y3
6 'VAR' Variables: X1 X2 X3 Y1 Y2 Y3

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 20

	X1	X2	X3	Y1	Y2	Y3
X1 Poids	1.00000 0.0	**0.87024** 0.0001	-0.36576 0.1128	-0.38969 0.0894	**-0.49308** 0.0272	-0.22630 0.3374
X2 Tour de taille	0.87024 0.0001	1.00000 0.0	-0.35289 0.1270	**-0.55223** 0.0116	**-0.64560** 0.0021	-0.19150 0.4186
X3 Pouls	-0.36576 0.1128	-0.35289 0.1270	1.00000 0.0	0.15065 0.5261	0.22504 0.3401	0.03493 0.8838
Y1 Tractions	-0.38969 0.0894	**-0.55223** 0.0116	0.15065 0.5261	1.00000 0.0	0.69573 0.0007	**0.49576** 0.0262
Y2 Flexions	**-0.49308** 0.0272	**-0.64560** 0.0021	0.22504 0.3401	**0.69573** 0.0007	1.00000 0.0	**0.66921** 0.0013
Y3 Sauts	-0.22630 0.3374	-0.19150 0.4186	0.03493 0.8838	**0.49576** 0.0262	**0.66921** 0.0013	1.00000 0.0

L'examen de la matrice des corrélations montre :

- pour le tableau X, une corrélation positive entre le poids (x_1) et le tour de taille (x_2)et des corrélations négatives entre le pouls (x_3) d'une part, et d'autre part le poids et le tour de taille ;
- pour le tableau Y, des corrélations positives entre les résultats aux exercices de tractions (y_1), flexions (y_2) et sauts (y_3) ;
- pour les relations entre les deux tableaux, on peut voir que les résultats en tractions, flexions et sauts sont corrélés négativement aux mesures de poids, de tour de taille et positivement à la mesure du pouls.

Les estimations des coefficients de régression $\hat{\beta}_j$, fournis par un modèle de régression des moindres carrés ordinaires régressant les « réponses » centrées réduites y_j^* sur les « facteurs » centrés réduits x_j^*[11] sont les suivantes :

Tableau T14 : coefficients de la régression MCO sur variables centrées-réduites.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Régression MCO: modèle sur variables centrées-réduites
Tableau des coefficients de régression de Y* en X*

X_VAR	B1	B2	B3
X1	0.36825	0.28715	-0.25899
X2	-0.88182	-0.88983	0.01460
X3	-0.02585	0.01606	-0.05464

On obtient ainsi les équations :

$$Tractions^* \approx 0{,}3683 \cdot Poids^* - 0{,}8818 \cdot Tour\_de\_taille^* - 0{,}0259 \cdot Pouls^*$$
$$Flexions^* \approx 0{,}2872 \cdot Poids^* - 0{,}8898 \cdot Tour\_de\_taille^* + 0{,}0161 \cdot Pouls^*$$
$$Sauts^* \approx -0{,}2590 \cdot Poids^* + 0{,}0146 \cdot Tour\_de\_taille^* - 0{,}0546 \cdot Pouls^*$$

qui ne s'avèrent pas cohérentes avec les conclusions tirées de l'examen de la matrice des corrélations selon lesquelles nous devrions avoir une régression de la forme $\hat{y}_k^* = \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^*$, avec $\beta_1 < 0$, $\beta_2 > 0$ et $\beta_3 > 0$.

[11] Les coefficient de la régression MCO peuvent être récupérés directement à partir des résultats de la régression MCP effectuée par la procédure PLS. En effet, si H, la dimension du sous-espace des composantes MCP utilisées, est égale à r le rang de l'opérateur V de variance-covariance, alors la solution fournie par la régression MCP est la solution sans biais des MCO. En prenant ici trois composantes MCP, le modèle fournit la solution des moindres carrés ordinaires.

La régression des moindres carrés partiels va permettre d'obtenir des coefficients de régression dont le signe soit cohérent avec les corrélations enregistrées. La mise en oeuvre de la régression MCP peut s'effectuer en utilisant la procédure PLS de SAS de la manière suivante :

```
title1 'LINNERUD, Exercices physiques';
 title2 "Modèle de régression MCP à 3 composante(s)";

 proc pls data=linnerud method=pls lv=3 outmodel=estim;
  model y1 - y3 = x1 - x3;
  output out=outmcp yscore=u xscore=t p=yest1-yest3 yr=yres1-yres3
   xr=xres1-xres3 stdy=sy stdx=sx h=h press=prs t2=scm xqres=xq yqres=yq;
 run;
```

L'algorithme des moindres carrés partiels extrait des composantes au sein de chacun des espaces de variables associés aux tableaux X et Y qui maximisent la covariance entre les composantes de X (T) et celles de Y (respectivement U). Les notations sont similaires au cas de la réponse univariée, rappelons que, d'une manière générale, le modèle de la régression MCP s'écrit :

$$X = TP' + E$$
$$Y = UQ' + F$$

où X : matrice des facteurs Y : matrice des réponses

T : matrice des X-composantes U : matrice des Y-composantes

P : matrice des X-saturations Q : matrice des Y-saturations

E : matrice des X-résidus F : matrice des Y-résidus

Les paramètres de la procédure `PLS` indiquent le nom de la table SAS (`data=linnerud`), la méthode d'extraction des composantes (`method=pls`), ainsi que le nombre de composantes à extraire (`lv=3`) et le fichier des paramètres du modèle (`outmodel=estim`). L'instruction **MODEL** permet de spécifier les facteurs utilisés pour modéliser la variable expliquée (`y1-y3=x1-x3`). L'instruction **OUTPUT** indique les éléments d'information contenus dans le fichier des résultats (`out=outmcp`).

III.2.2) Sélection du nombre de pseudo-composantes par validation croisée

À l'instar du modèle de régression MCP univarié, la sélection du nombre pertinent de pseudo-composantes s'effectue par validation croisée en utilisant le test proposé par [van der Voet, 1994]. Pour appliquer le test de van der Voet, il suffit de spécifier l'option `CVTEST` lors du choix de la méthode de validation croisée. La méthode de validation croisée

choisie (SPLIT) est celle du « *split-sample* » où chaque n^e observation est incluse dans l'échantillon-test. Par défaut $n=1$, chaque observation constitue un échantillon-test.

```
proc pls data=linnerud method=pls cv=split cvtest;
model y1 - y3 = x1 - x3;
output out=cpcrout yscore=u xscore=t pred=p yr=e stdy=sy stdx=sx h=h
                   press=prs t2=ssq xqres=xq yqres=yq;
run;
```

Le nombre de pseudo-composantes déterminé par le test de van der Voet est le plus petit nombre de composantes MCP tel que les résidus associés ne soient pas significativement plus grands que les résidus du modèle dont le *PRESS* est minimum.

Tableau T15 : test de comparaison des modèles MCP.

The PLS Procedure
Cross Validation for the Number of Latent Variables

Number of Latent Variables	Root Mean PRESS	Test for larger residuals than minimum T2	Prob > T2
0	1.1153	4.7432	0.2150
1	1.0365	0	1.0000
2	1.1018	3.3460	0.3910
3	1.0769	2.8144	0.4530

Minimum Root Mean PRESS = 1.036531 for 1 latent variable
Smallest model with p-value > 0.1: 0 latent variables

Soit $R_{h,im}$, le i^e résidu individuel pour la réponse m d'un modèle estimé avec h pseudo-composantes, la statistique *PRESS* est définie par : $PRESS = \sum_{im} R^2_{h,im}$. Soit $h_{\min}$, le nombre de pseudo-composantes pour lequel la valeur du PRESS est minimale. La valeur critique du test de van der Voet est basée sur la différence observée entre les carrés des résidus d'estimation :

$$D_{h,im} = R^2_{h,im} - R^2_{h_{\min},im}$$

Une autre valeur critique utilisable pour le test de van der Voet est la quantité :

$$C_h = \sum_{im} D_{h,im}$$

distribuée suivant un T^2 de Hotelling et qui représente la différence entre les valeurs de la statistique PRESS pour le modèle à h pseudo-composantes et celui à h_{min} pseudo-composantes.

On peut également utiliser la statistique :

$$C_h = d'_h S_h d_h$$

où $d_h = \sum_i d_{h,i}$ somme des vecteurs de différences

$$d_{h,i} = \left\{ D_{h,i1}, \ldots, D_{h,im}, \ldots, D_{h,iM} \right\}$$

et $S_h = \sum_i d_{h,i} d'_{h,i}$ matrice des somme de carrés et produits.

Théoriquement, le seuil de significativité du test de van de Voet s'obtient en comparant la statistique C_h avec une distribution de valeurs, résultat de permutations aléatoires entre les résidus au carré $R^2_{h,im}$ et $R^2_{h_{\min},im}$. En pratique, le seuil de significativité est déterminé de manière approchée par une simulation de Monte-Carlo, comme proportion de valeurs critiques simulées qui soient supérieures à la valeur C_h.

Dans l'option CVTEST spécifiée pour comparer différents modèles en appliquant le test de van der Voet, le nombre de pseudo-composantes sélectionnées par défaut est le nombre minimum de pseudo-composantes associées à un seuil de significativité approché supérieur à 10 %.

Tableau T16 : pourcentages de variance expliquée par le modèle à 1 pseudo-composante.

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	69.4781	69.4781	20.9447	20.9447

Les résultats de la validation croisée montrent selon le test de van der Voet que seule la première composante MCP peut être jugée significative. Cette première pseudo-composante t_1 permet de reconstituer 69,5 % de la variabilité des variables explicatives (tableau X) et 20,9 % des variables expliquées (tableau Y).

III.2.3) L'interprétation des résultats

En premier lieu, la procédure PLS produit un tableau permettant de vérifier comment chaque composante MCP extraite rend compte de la variabilité des facteurs (`Model Effects`) et de la réponse (`Dependent Variables`) au sein du modèle.

Tableau T17 : parts de variance expliquée par les composantes MCP.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	69.4781	69.4781	20.9447	20.9447
2	22.6694	92.1475	2.9491	23.8938
3	7.8525	100.0000	3.7718	27.6656

Ainsi, on peut constater que t_1, la première composante MCP extraite, représente près de 21 % de la variabilité des résultats et explique plus de 69 % de la variabilité des mesures. La deuxième pseudo-composante t_2 explique près de 3 % de la variabilité des résultats et prend en compte près de 23 % de la variance des mesures. La troisième et dernière composante MCP t_3 explique près de 4 % de la variance des résultats et prend en compte la variabilité résiduelle (près de 8 %) des mesures. On voit donc que la relation entre les deux tableaux est essentiellement déterminée par la première composante MCP t_1.

La procédure PLS produit également deux tables SAS : la première contient le fichier des paramètres du modèle (`estim`), la seconde le fichier des résultats (`outmcp`)

Les X-contributions w_h sont extraites à chaque étape $h$comme vecteurs propres de l'opérateur $V'_{h-1}V_{h-1}$ où $V_{h-1} = F'_{h-1}E_{h-1}$ est la matrice de variance-covariance des résidus constitués à l'étape précédente $h-1$. On peut les lister en utilisant le macroprogramme `%get_wts` :

Tableau T18 : coordonnées des X-contributions.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Tableau des X-contributions w(h)

X_VAR	_LABEL_	W1	W2	W3
X1	Poids	-0.59846	0.58405	0.65747
X2	Tour de taille	-0.78255	-0.70767	-0.28706
X3	Pouls	0.24235	-0.84279	0.69666

ce qui permet d'écrire les équations exprimant les pseudo-composantes t_h en fonction des x_j^*, variables centrées-réduites du tableau X :

$$t_1 = E_0 w_1 \approx -0{,}598x_1^* - 0{,}783x_2^* + 0{,}242x_3^*$$
$$t_2 = E_1 w_2 = E_0\left(I - w_1 p_1'\right)w_2 = E_0 \widetilde{w}_2$$
$$t_3 = E_2 w_3 = E_0\left(I - w_1 p_1'\right)\left(I - w_2 p_2'\right)w_3 = E_0 \widetilde{w}_3$$

en utilisant le listage des X-saturations obtenu par l'appel au macroprogramme `%getxload` :

Tableau T19 : coordonnées des X-saturations.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Tableau des X-saturations p(h)

X_VAR	_LABEL_	P1	P2	P3
X1	Poids	-0.65638	-0.01588	0.65747
X2	Tour de taille	-0.66634	-0.28469	-0.28706
X3	Pouls	0.35378	-0.95849	0.69666

Les X-saturations p_h permettent également d'écrire la décomposition orthogonale du X-tableau centré-réduit E_0 selon les X-composantes MCP t_h :

$$E_0 = t_1 p_1' + t_2 p_2' + t_3 p_3'$$

qui a permis de calculer la part de variance du tableau X expliquée par chacune des composantes MCP en appliquant le théorème d'Huyghens.

Les Y-saturations q_h sont extraites à chaque étape $h$comme vecteurs propres de l'opérateur $V_{h-1}V_{h-1}'$, où $V_{h-1} = F_{h-1}'E_{h-1}$ est la matrice de variance-covariance des résidus de l'estimation du modèle de régression MCP à l'étape précédente $h-1$. On obtient leur tableau par l'appel au macroprogramme `%getyload` :

Tableau T20 : coordonnées des Y-saturations.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Tableau des Y-saturations q(h)

Y_VAR	Q1	Q2	Q3
Y1	0.61331	0.74852	0.68860
Y2	0.74697	0.64705	0.65710
Y3	0.25669	0.14509	-0.30666

Les Y-saturations q_h permettent de calculer les Y-composantes MCP u_h selon la formule :

$$u_h = F_{h-1} q_h$$

En utilisant le macroprogramme `%prt_scr`, on obtient le tableau des X-composantes MCP t_h et des Y-composantes MCP u_h :

Tableau T21 : coordonnées des X et Y-composantes MCP.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Composantes MCP: projections des observations

OBS	T1	T2	T3	U1	U2	U3
1	-0.65222	0.73677	0.13052	-0.37145	-0.13831	-0.28462
2	-0.78092	0.20766	-0.13395	-1.34032	-1.03165	-0.99155
3	-0.92061	-0.64938	-0.04769	-0.08235	0.48238	0.31235
4	0.69842	-0.84724	-0.34579	-0.35497	-0.52817	0.05929
5	-0.49379	1.41128	0.18177	0.46312	0.83119	0.40533
6	-0.23241	-0.08929	-0.02473	-1.30584	-1.18749	-0.87323
7	-1.42412	-0.09555	0.57184	-0.86179	0.00857	0.21975
8	0.75441	-0.26243	0.03223	-0.79728	-1.19254	-0.74484
9	1.74000	-0.81596	1.55724	1.14230	0.32734	0.94256
10	1.17946	0.20786	-0.33289	3.03444	2.03364	0.40873
11	0.36982	0.87298	-0.20109	0.40922	0.51451	0.45672
12	0.75415	0.86996	-0.00228	1.40508	0.89003	0.24529
13	1.20395	-0.94310	-0.33623	1.53074	0.81311	0.86780
14	-4.45355	-0.94683	-0.25544	-2.22273	0.15336	0.36362
15	-0.83515	1.21322	0.08261	-1.49897	-0.93182	-1.01734
16	-0.75991	-0.64928	0.66716	1.31409	1.57696	1.26984
17	-0.39860	-0.25335	-0.56422	-1.88043	-1.57022	-1.07300
18	1.21670	0.97519	-0.09238	1.23662	0.46585	0.15820
19	1.06376	0.46464	-0.31899	1.60596	1.04298	0.90368
20	1.97061	-1.40713	-0.56769	-1.42542	-2.55972	-1.62859

À chaque étape du processus d'extraction, les composantes MCP se déduisent des vecteurs propres de l'opérateur $V'V$, où $V = F'E$ est la matrice de variance-covariance des résidus des équations générales du modèle à l'étape précédente, selon les formules suivantes :

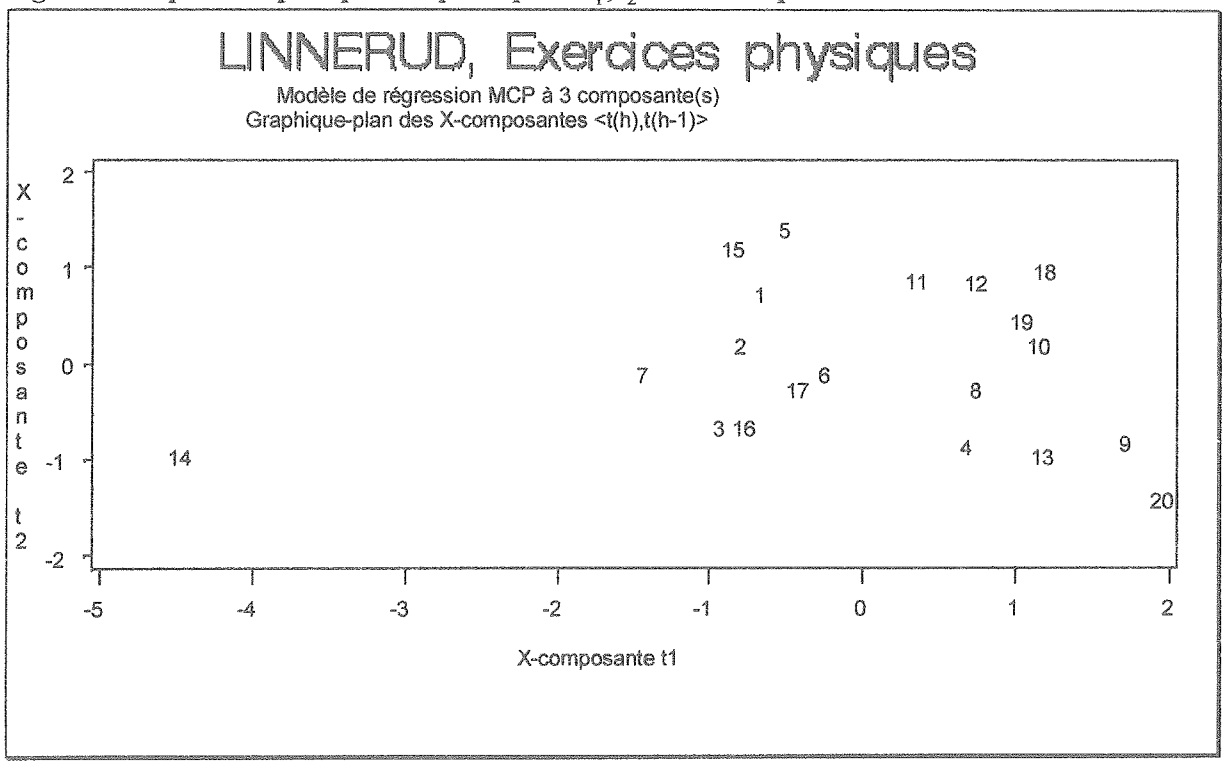
$$\begin{aligned} t_h &= E_{h-1} w_h \\ u_h &= F_{h-1} q_h \end{aligned} \qquad \text{avec} \quad w_h \text{ vecteur propre de } V'_{h-1}V_{h-1} \text{ et } q_h \text{ vecteur propre de } V_{h-1}V'_{h-1}$$

Pour sa part, le logiciel SIMCA normalise chacune des composantes u_h pour que le coefficient de régression de u_h sur t_h soit égal à 1 : $u_h = \frac{1}{b_h} F_{h-1} q_h$ avec $b_h = \frac{t'_h F_{h-1} q_h}{t'_h t_h}$

coefficient de la régression de $F_{h-1} q_h$ sur t_h.

Obtenu grâce au macroprogramme `%plotxscr`, le graphique-plan des deux premières X-composantes MCP $< t_1, t_2 >$ est l'analogue d'un premier plan factoriel d'une analyse en composantes principales du tableau X orientée vers l'explication du tableau Y :

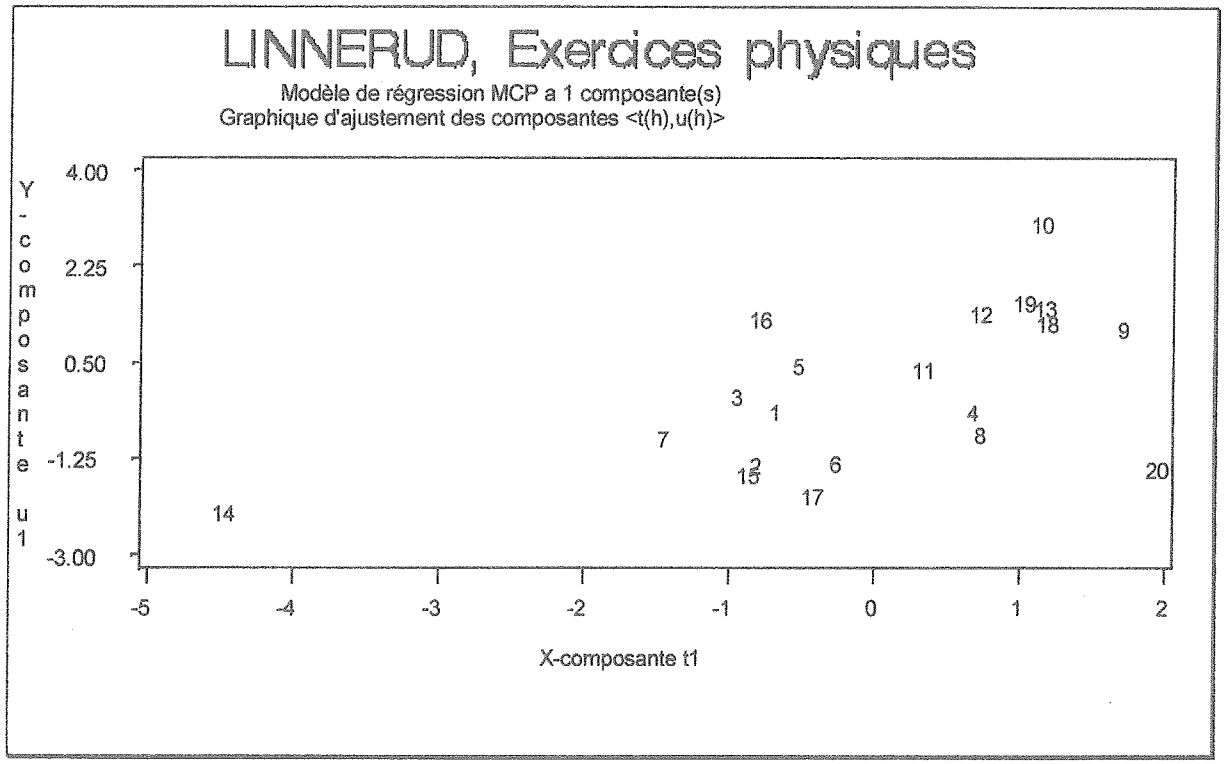
Figure F11 : premier plan pseudo-principal $< t_1, t_2 >$ dans l'espace des X.



Sur ce premier plan pseudo-principal, l'individu 14 apparaît comme particulièrement atypique. De fait, c'est le plus gros de l'échantillon (le poids et le tour de taille sont maximum) et ses résultats sont parmi les plus faibles du groupe (1 seule traction, 50 flexions et 50 sauts).

Pour vérifier la corrélation entre les deux facteurs de la première composante MCP, on utilise le macroprogramme `plot_scr` :

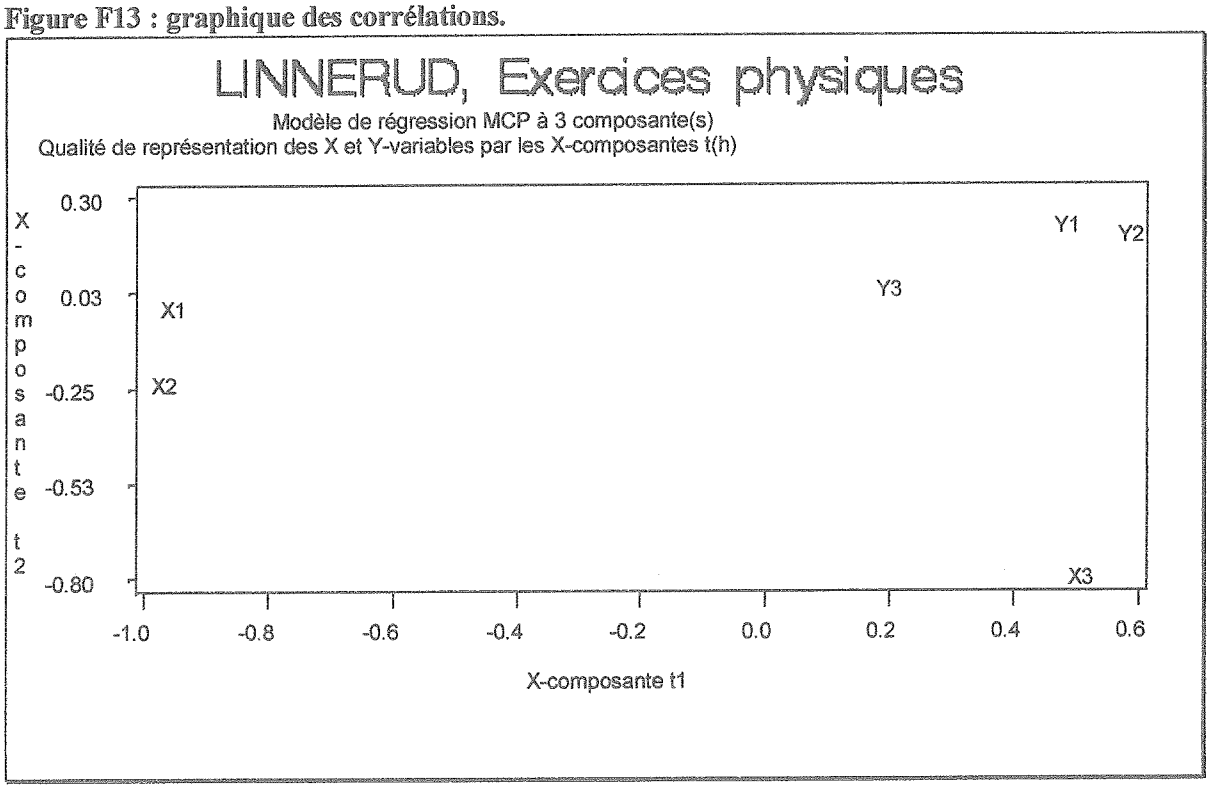
Figure F12 : liaison entre les deux facteurs t_1 et u_1 de la première composante MCP.



Ce graphique permet de mettre en évidence l'influence particulière que peuvent avoir les individus 10, 14 et 20 sur les résultats de la régression MCP. L'individu 14 a déjà été signalé pour ses caractéristiques saturo-pondérales atypiques et ses faibles résultats. Les individus 10 et 20 partagent des caractéristiques physiques semblables mais obtiennent des résultats diamétralement opposés : l'individu 10 réalise les meilleurs scores en tractions, flexions et sauts alors que les performances de l'individu 20 à ces exercices se situent parmi les plus médiocres. Refaire l'analyse en mettant ces observations en supplémentaire permettrait de déterminer leur influence sur la valeur des coefficients de régression.

Le graphique des corrélations des variables avec les deux premières X-composantes générées par le macroprogramme `%pltcorr` traduit bien les liaisons internes à chaque groupe de variables comme les relations entre variables des deux tableaux X et Y :

Figure F13 : graphique des corrélations.



L'examen des corrélations entre les variables des tableaux X et Y et les composantes MCP, obtenues par l'intermédiaire du macroprogramme `%prt_corr`, permet de donner une interprétation de la variabilité expliquée ou prise en compte par ces pseudo-composantes :

Tableau T22 : matrice des corrélations entre variables et composantes MCP.

LINNERUD, Exercices physiques
Modèle de régression MCP à 3 composante(s)
Corrélations des X et Y-variables avec les X-composantes t(h)

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 20

	T1	T2	T3
X1	**-0.94763**	-0.01309	0.31911
Poids	0.0001	0.9563	0.1703
X2	**-0.96201**	-0.23478	-0.13933
Tour de taille	0.0001	0.3191	0.5580
X3	**0.51076**	**-0.79044**	0.33813
Pouls	0.0214	0.0001	0.1448
Y1	**0.48616**	0.22264	0.23163
Tractions	0.0297	0.3454	0.3258
Y2	**0.59211**	0.19246	0.22104
Flexions	0.0059	0.4163	0.3490
Y3	0.20347	0.04316	-0.10315
Sauts	0.3896	0.8566	0.6652

Corrélations des Y-variables avec les Y-composantes u(h)

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 20

	U1	U2	U3
Y1	**0.88016**	0.78202	0.76486
Tractions	0.0001	0.0001	0.0001
Y2	**0.93966**	0.70498	0.60908
Flexions	0.0001	0.0005	0.0044
Y3	0.74074	0.69278	0.40984
Sauts	0.0002	0.0007	0.0727

On peut en déduire que la pseudo-composante t_1 est un indicateur saturo-pondéral qui classe les individus en fonction de leur poids et de leur tour de taille, corrélé au pouls. Elle traduit essentiellement un potentiel physique global. La pseudo-composante t_2 est négativement corrélée au pouls, opposant le poids au tour de taille. Elle permet ainsi de différencier des individus de même poids mais de corpulence dissemblable (de façon schématique, distinguer les « grands maigres » des « petits gros »). La pseudo-composante t_3 n'apporte guère d'information.

La pseudo-composante u_1 semble pouvoir être interprétée comme un indicateur de performance composite des trois types d'exercices. Cependant cet indicateur accorde plus

d'importance aux tractions et aux flexions qu'aux sauts. Les pseudo-composantes u_2 et u_3 apparaissent largement redondantes avec la première Y-composante.

Les macroprogrammes `%get_bpls` et `%get_vip` permettent d'obtenir les coefficients de régression MCP sur les variables centrées réduites, ainsi que les indices VIP. Nous avons vu que le modèle de régression MCP à trois pseudo-composantes nous ramène à la solution sans biais des moindres carrés ordinaires. Le modèle de régression MCP à deux pseudo-composantes donne les résultats suivants :

Tableau T23 : coefficients de régression sur variables normées, modèle à 2 composantes.

LINNERUD, Exercices physiques
Modèle de régression MCP à 2 composante(s)
Indices VIP: importance de la contribution à la projection
Tableau des coefficients de régression et des indices VIP

OBS	X_VAR	B1	B2	B3	VIP
1	X1	-0.07777	-0.13847	-0.06036	1.03352
2	X2	-0.49893	-0.52445	-0.15592	1.34009
3	X3	-0.13219	-0.08542	-0.00729	0.64611

Les valeurs des coefficients VIP suggèrent que la mesure du pouls (x_3) n'apparaît pas déterminante pour l'estimation de Y. D'autre part, les valeurs des coefficients de régression, si elles diffèrent notablement en valeur absolue, sont toutes du même signe, ce qui ne semble pas cohérent avec les corrélations observées. Les résultats aux exercices relèvent d'un phénomène unidimensionnel lié à la première composante MCP, la deuxième composante MCP n'ayant de liaison au sein du tableau X qu'avec la variable x_3, jugée peu importante pour la projection du tableau Y sur le tableau X. Il serait donc plus approprié de n'utiliser que la première composante MCP.

Avec une seule composante MCP, on obtient les résultats suivants :

Tableau T24 : coefficients de régression sur variables normées, modèle à 1 composante.

LINNERUD, Exercices physiques
Modèle de régression MCP à 1 composante(s)
Indices VIP: importance de la contribution à la projection
Tableau des coefficients de régression et des indices VIP

X_VAR	B1	B2	B3	VIP
X1	-0.20152	-0.24545	-0.08434	1.04849
X2	-0.26351	-0.32094	-0.11029	1.25885
X3	0.08161	0.09939	0.03415	0.74769

Les coefficients de régression obtenus paraissent cohérents avec les liaisons enregistrées au sein et entre les tableaux étudiés, tant du point de vue des signes que des

valeurs absolues. Les équations de régression MCP sur les variables centrées réduites s'écrivent :

$$\begin{aligned} Tractions^* &\approx -0,2015 \cdot Poids^* - 0,2635 \cdot Tour\_de\_taille^* + 0,0816 \cdot Pouls^* \\ Flexions^* &\approx -0,2455 \cdot Poids^* - 0,3209 \cdot Tour\_de\_taille^* + 0,0994 \cdot Pouls^* \\ Sauts^* &\approx -0,0843 \cdot Poids^* - 0,1103 \cdot Tour\_de\_taille^* + 0,0342 \cdot Pouls^* \end{aligned}$$

Pour obtenir les équations de régression sur les variables d'origine, il suffit d'utiliser le macroprogramme `%get_apls` dont voici les résultats :

Tableau T25 : coefficients de régression sur variables d'origine, modèle à 1 composante.

LINNERUD, Exercices physiques
Modèle de régression MCP à 1 composante(s)
Régression MCP: modèle sur variables d'origine
Régression MCP de Y sur 1 composante(s)
Modèle sur variables d'origine: Y en fonction de X

Y_VAR	CONSTANT	X1	X2	X3
Y1	29.200	-0.04315	-0.43505	0.05983
Y2	430.251	-0.62197	-6.27125	0.86246
Y3	150.481	-0.17517	-1.76618	0.24290

ce qui donne le modèle MCP suivant :

$$\begin{aligned} Tractions &\approx 29,200 - 0,0432 \cdot Poids - 0,4351 \cdot Tour\_de\_taille + 0,0598 \cdot Pouls \\ Flexions &\approx 430,251 - 0,6220 \cdot Poids - 6,2713 \cdot Tour\_de\_taille + 0,8625 \cdot Pouls \\ Sauts &\approx 150,481 - 0,1752 \cdot Poids - 1,7662 \cdot Tour\_de\_taille + 0,2429 \cdot Pouls \end{aligned}$$

Les diagrammes comparant les valeurs estimées et les valeurs résiduelles, obtenus grâce au macroprogramme `%res_plot`, confirment que les individus 10, 14 et 20 posent un problème d'estimation (cf. les diagrammes suivants concernant y_1 puis y_3, le diagramme pour y_2 étant similaire à celui de y_1) :

Figure F14 : diagramme des valeurs estimées et résiduelles de la variable y_1.

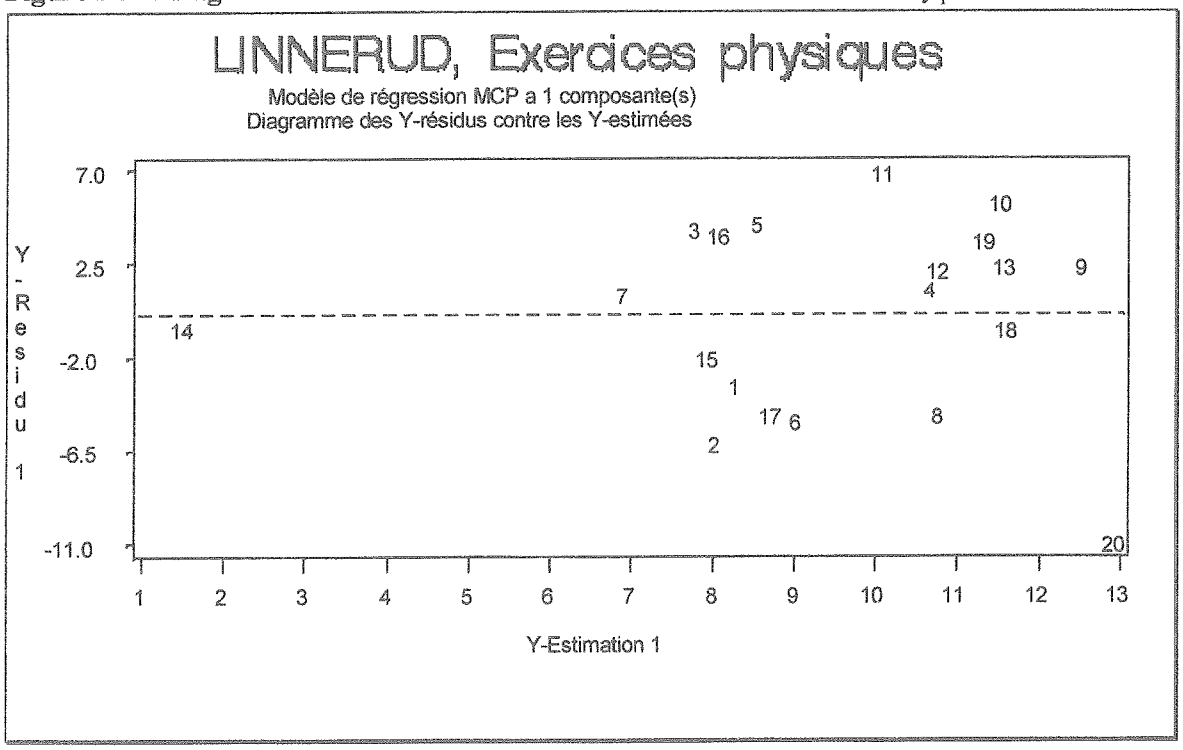
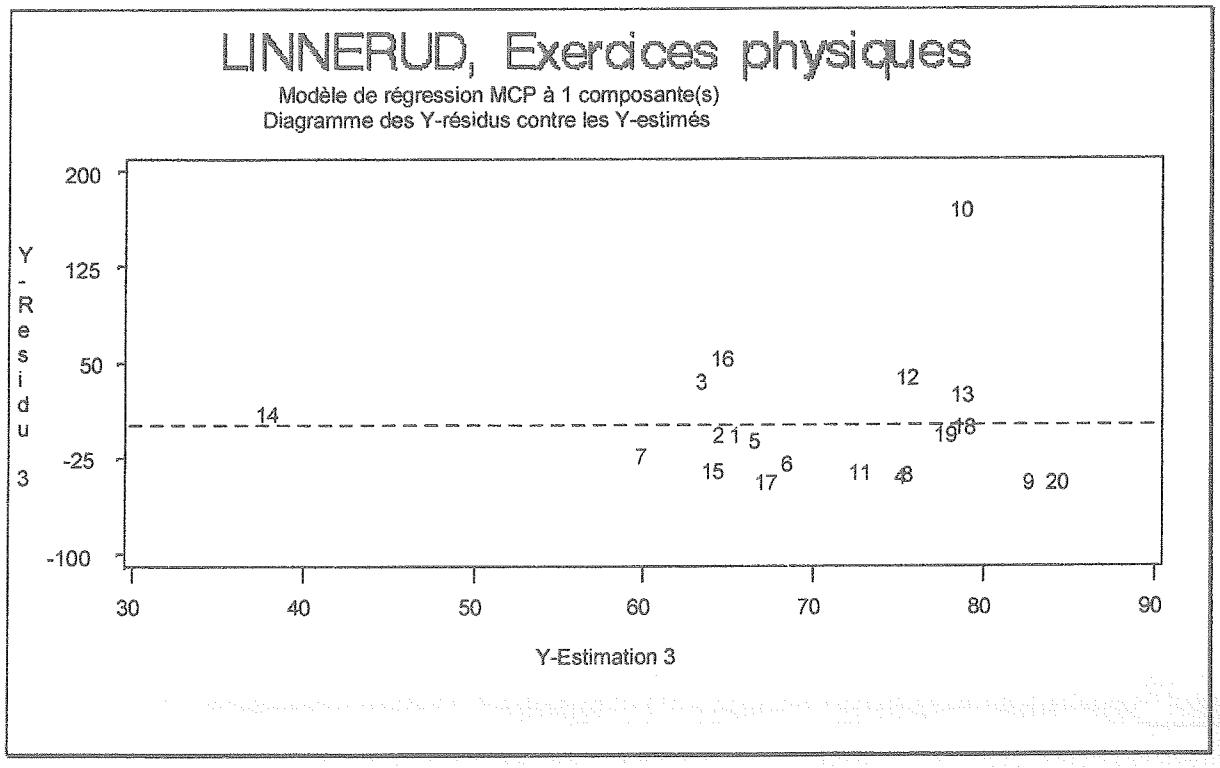
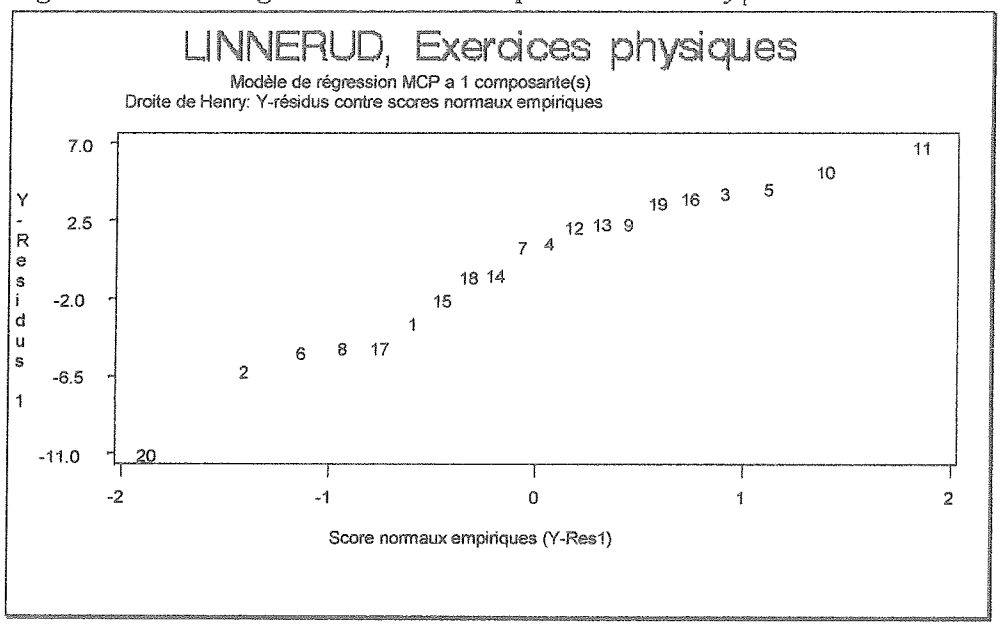


Figure F15 : diagramme des valeurs résiduelles et estimées pour la variable y_3.



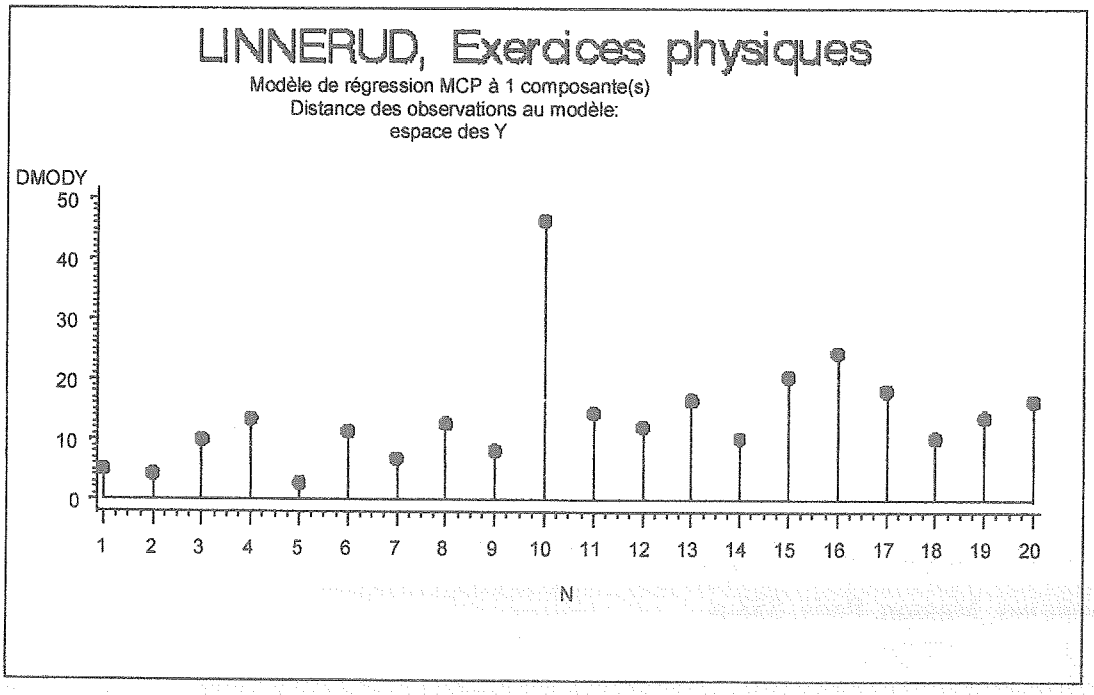
De façon cohérente, le normogramme obtenu grâce au macroprogramme `%nor_plot`, semble indiquer que les valeurs extrêmes sont mal estimées par le modèle :

Figure F16 : normogramme des résidus pour la variable y_1.



Le calcul des distances au modèle, effectué par le macroprogramme `%get_dmod`, montre que dans l'espace des Y c'est l'observation n°10 qui est la moins bien estimée. La mise en éléments supplémentaires de ces trois observations (10, 14 et 20) devrait permettre d'améliorer l'estimation pour les autres observations.

Figure F17 : distances des observations au modèle dans l'espace Y.



IV) Références bibliographiques

Cornell J.A. (1990) *Experiments with Mixtures*, Wiley, New-York.

Davies P.T., Tso M.K.-S. (1993). « Procedures for Reduced Rank Regression », *Applied Statistics*, 31 (3), pp. 244-255.

Jackson J.E. (1991) *A User's Guide to Principal Components*, Wiley, New-York.

Franck I.E., Friedman J.H. (1993). « A Statistical View of Some Chemometrics Regression Tools », *Technometrics*, 35 (2), pp. 109-135.

Hoerl, A., Kennard R. (1970). « Ridge Regression: Biased Estimation for Non-orthogonal Problems », *Technometrics*, 12, pp. 55-67.

de Jong, S. et Kiers, H. (1992), « Principal Covariates Regression », *Chemometrics and Intelligent Laboratory Systems*, 14, pp. 155-164.

de Jong, S. (1993), « SIMPLS: An Alternative Approach to Partial Least Squares Regression », *Chemometrics and Intelligent Laboratory Systems*, 18, pp. 251-263.

Malinvaud E. (1981). *Méthodes statistiques de l'économétrie*, Dunod, Paris, 846 p.

Massy W.F. (1965). « Principal Components Regression in Exploratory Statistical Research », *Journal of the American Statistical Association*, 60, pp. 234-246.

Ranner, Lindgren, Geladi et Wold. (1994), « A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects », *Journal of Chemometrics*, 8, pp. 111-125.

Sarle W.S. (1994), « Neural Networks and Statistical Models », in *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., pp. 1538-1550.

Stone M.et Beooks R. (1990), « Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares, and Principal Components Regression », *Journal of the Royal Statistical Society, Series B*, 52(2)pp. 237-269.

Tenenhaus, M., Gauchi J.-P., Ménardo C. (1995), « Régression PLS et applications », *Revue de Statistique Appliquée*, XLIII (1), pp. 7-63.

Tenenhaus, M. (1998), *La régression PLS. Théorie et pratique*, Éditions Technip, 254 p.

Tobias R. (1994), « An Introduction to Partial Least Squares Regression », in *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., pp. 1250-1257.

Tomassone R., Dervin C., Masson J.-P. (1993). *BIOMÉTRIE, Modélisation des phénomènes biologiques*, Masson, Paris, 553 p.

van den Wollenberg, A.L. (1977), « Redundancy Analysis - An Alternative to Canonical Correlation Analysis », *Psychometrika*, 42, pp. 207-219.

van der Voet, H. (1994), « Comparing the Predictive Accuracy of Models Using a Simple Randomization Test », *Chemometrics and Intelligent Laboratory Systems*, 25, pp. 313-323.

Wold, H. (1966), « Estimation of Principal Components and Related Models by Iterative Least Squares », in *Multivariate Analysis*, éd. P.R. Krishnaiah, Academic Press, New York, pp. 391-420.

Wold, S. (1994), « PLS for Multivariate Linear Modeling », in *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*, éd. H. van de Watterbeemd, Verlag-Chemie, Weinheim-Germany.

ANNEXE I : Exemple CORNELL

AI.1) Texte-source du programme SAS.

```
data cornell;
 input num $ x1 x2 x3 x4 x5 x6 x7 y;
 label x1='Distillation directe'
       x2='Reformat'
       x3='Naphta de craquage thermique'
       x4='Naphta de craquage catalytique'
       x5='Polymere'
       x6='Alkylat'
       x7='Essence naturelle'
       y ='Indice d'octane moteur';
 cards;
01    0.00  0.23  0.00  0.00  0.00  0.74  0.03  98.7
02    0.00  0.10  0.00  0.00  0.12  0.74  0.04  97.8
03    0.00  0.00  0.00  0.10  0.12  0.74  0.04  96.6
04    0.00  0.49  0.00  0.00  0.12  0.37  0.02  92.0
05    0.00  0.00  0.00  0.62  0.12  0.18  0.08  86.6
06    0.00  0.62  0.00  0.00  0.00  0.37  0.01  91.2
07    0.17  0.27  0.10  0.38  0.00  0.00  0.08  81.9
08    0.17  0.19  0.10  0.38  0.02  0.06  0.08  83.1
09    0.17  0.21  0.10  0.38  0.00  0.06  0.08  82.4
10    0.17  0.15  0.10  0.38  0.02  0.10  0.08  83.2
11    0.21  0.36  0.12  0.25  0.00  0.00  0.06  81.4
12    0.00  0.00  0.00  0.55  0.00  0.37  0.08  88.1
 ;
 run;
* Declaration des macro-variables globales                                    *;
 %global xvars yvars predname resname xscrname yscrname
   num_x num_y lv;
 %let xvars=x1 x2 x3 x4 x5 x6 x7;
 %let yvars=y;
 %let ypred=yest1;
 %let yres=yres1;
 %let xscrname=t;
 %let yscrname=u;
 %let predname=yest;
 %let resname=res;
 %let num_y=1;
 %let num_x=7;
  * Nombre de composantes du modele                                           *;
 %let lv=3;
 title1 "CORNELL, indice d'octane moteur";
 title2 "Modèle de régression MCP à &lv composantes";
 proc pls data=cornell method=pls lv=&lv outmodel=estim;
  model y = x1 - x7;
  output out=outmcp yscore=u xscore=t p=yest1 yr=yres1
   xr=xres1-xres7 stdy=sy stdx=sx h=h press=prs t2=scm xqres=xq yqres=yq;
 run;
```

AI.2) Listing des résultats de la régression MCP univariée

The SAS System 15:32 Wednesday, July 9, 1997 3

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	57.3606	57.3606	92.3594	92.3594
2	15.2521	72.6128	5.2749	97.6343
3	19.2130	91.8258	1.4212	99.0556

OBS	NUM	X1	X2	X3	X4	X5	X6	X7	Y	P	E
1	01	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.7	97.5586	1.14136
2	02	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.8	97.5915	0.20849
3	03	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.6	97.4453	-0.84534
4	04	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.0	91.8079	0.19207
5	05	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.6	85.9945	0.60548
6	06	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.2	91.7751	-0.57507
7	07	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.9	81.4967	0.40330
8	08	0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.1	82.5754	0.52458
9	09	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.4	82.5240	-0.12399
10	10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.2	83.2603	-0.06027
11	11	0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.4	81.9292	-0.52924
12	12	0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.1	89.0414	-0.94136

OBS	T1	T2	T3	U1	U2	U3
1	2.06171	1.00975	1.58621	1.55133	0.56252	0.33808
2	2.48720	0.79721	0.10968	1.41332	0.22045	0.04325
3	2.34290	1.13865	-0.17338	1.22931	0.10564	-0.14745
4	2.04748	-1.96069	-0.50286	0.52393	-0.45806	-0.02224
5	-0.06845	-0.26763	-2.96368	-0.30413	-0.27130	-0.21181
6	1.62199	-1.74815	0.97367	0.40125	-0.37667	0.01191
7	-2.21542	-0.21885	0.23815	-1.02485	0.03768	0.08633
8	-2.00365	0.12361	0.11873	-0.84084	0.12012	0.09265
9	-2.09817	0.18257	0.35554	-0.94818	0.05812	0.01754
10	-1.92548	0.39123	0.19699	-0.82550	0.09797	0.01101
11	-2.08681	-0.56599	1.03390	-1.10152	-0.10068	0.02513
12	-0.16329	1.11830	-0.97296	-0.07412	0.00420	-0.24437

ANNEXE II : Exemple LINNERUD

AII.1) Texte-source du programme SAS.

```
data linnerud;
input num $ x1 x2 x3 y1 y2 y3;
cards;
01 191     36      50      5       162     60
02 189     37      52      2       110     60
03 193     38      58      12      101     101
04 162     35      62      12      105     37
05 189     35      46      13      155     58
06 182     36      56      4       101     42
07 211     38      56      8       101     38
08 167     34      60      6       125     40
09 176     31      74      15      200     40
10 154     33      56      17      251     250
11 169     34      50      17      120     38
12 166     33      52      13      210     115
13 154     34      64      14      215     105
14 247     46      50      1       50      50
15 193     36      46      6       70      31
16 202     37      62      12      210     120
17 176     37      54      4       60      25
18 157     32      52      11      230     80
19 156     33      54      15      225     73
20 138     33      68      2       110     43
;
run;
proc pls data=linnerud method= pls lv=3;
model y1 - y3 = x1 - x3;
output out=cpcrout yscore=u xscore=t pred=p yr=e stdy=sy stdx=sx h=h
                   press=prs t2=ssq xqres=xq yqres=yq;
run;
proc print data=cpcrout;
run;
```

AII.2) Listing des résultats de la régression MCP multivariée

The PLS Procedure
Percent Variation Accounted For

Number of Latent Variables	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	69.4781	69.4781	20.9447	20.9447
2	22.6694	92.1475	2.9491	23.8938
3	7.8525	100.0000	3.7718	27.6656

OBS	NUM	X1	X2	X3	Y1	Y2	Y3	SX	SY	P	E	H	PRS	T1
1	01	191	36	50	5	162	60	0.50222	-0.84180	9.6698	-4.66975	1.0353	132.433	-0.65222
2	02	189	37	52	2	110	60	0.42121	-1.40931	8.0183	-6.01832	0.7209	-21.563	-0.78092
3	03	193	38	58	12	101	101	0.58322	0.48238	6.7642	5.23584	1.3215	-16.286	-0.92061
4	04	162	35	62	12	105	37	-0.67232	0.48238	8.6117	3.38828	1.3752	-9.031	0.69842
5	05	189	35	46	13	155	58	0.42121	0.67155	11.0437	1.95629	2.3186	-1.484	-0.49379
6	06	182	36	56	4	101	42	0.13770	-1.03097	8.8465	-4.84646	0.1126	-5.461	-0.23241
7	07	211	38	56	8	101	38	1.31225	-0.27430	8.2212	-0.22124	2.4143	0.156	-1.42412
8	08	167	34	60	6	125	40	-0.46982	-0.65263	10.4997	-4.49969	0.6890	-14.470	0.75441
9	09	176	31	74	15	200	40	-0.10530	1.04989	15.3115	-0.31151	6.1684	0.060	1.74000
10	10	154	33	56	17	251	250	-0.99633	1.42823	11.0064	5.99364	1.5951	-10.071	1.17946
11	11	169	34	50	17	120	38	-0.38881	1.42823	10.8469	6.15313	0.9893	575.294	0.36982
12	12	166	33	52	13	210	115	-0.51032	0.67155	12.0283	0.97172	1.3756	-2.587	0.75415
13	13	154	34	64	14	215	105	-0.99633	0.86072	9.3989	4.60108	2.5020	-3.063	1.20395
14	14	247	46	50	1	50	50	2.77030	-1.59848	-0.4734	1.47342	20.8458	-0.074	-4.45355
15	15	193	36	46	6	70	31	0.58322	-0.65263	9.9032	-3.90324	2.2262	3.183	-0.83515
16	16	202	37	62	12	210	120	0.94773	0.48238	8.8538	3.14621	1.4941	-6.367	-0.75991
17	17	176	37	54	4	60	25	-0.10530	-1.03097	6.9555	-2.95545	0.5914	-7.233	-0.39860
18	18	157	32	52	11	230	80	-0.87483	0.29321	12.7745	-1.77453	2.4899	1.191	1.21670
19	19	156	33	54	15	225	73	-0.91533	1.04989	11.2019	3.79805	1.4992	-7.608	1.06376
20	20	138	33	68	2	110	43	-1.64436	-1.40931	9.5175	-7.51746	6.2356	1.436	1.97061

OBS	T2	T3	U1	U2	U3	SSQ	XQ	YQ
1	0.73677	0.13052	-0.37145	-0.13831	-0.28462	1.0746	6.3483E-33	0.8841
2	0.20766	-0.13395	-1.34032	-1.03165	-0.99155	0.4321	1.6833E-34	1.3685
3	-0.64938	-0.04769	-0.08235	0.48238	0.31235	1.0363	1.3563E-32	1.5729
4	-0.84724	-0.34579	-0.35497	-0.52817	0.05929	1.7971	7.0186E-33	1.3502
5	1.41128	0.18177	0.46312	0.83119	0.40533	3.1859	3.0084E-32	0.1826
6	-0.08929	-0.02473	-1.30584	-1.18749	-0.87323	0.0402	8.6506E-34	1.4524
7	-0.09555	0.57184	-0.86179	0.00857	0.21975	2.3746	3.5652E-34	0.2274
8	-0.26243	0.03223	-0.79728	-1.19254	-0.74484	0.3787	1.365E-34	1.5320
9	-0.81596	1.55724	1.14230	0.32734	0.94256	12.7256	9.7602E-32	0.3485
10	0.20786	-0.33289	3.03444	2.03364	0.40873	1.2014	1.3723E-32	13.5935
11	0.87298	-0.20109	0.40922	0.51451	0.45672	1.3579	7.2313E-33	2.4003
12	0.86996	-0.00228	1.40508	0.89003	0.24529	1.3857	6.253E-33	0.8205
13	-0.94310	-0.33623	1.53074	0.81311	0.86780	2.4832	1.9269E-32	1.9718
14	-0.94683	-0.25544	-2.22273	0.15336	0.36362	11.1109	1.1758E-32	0.5345
15	1.21322	0.08261	-1.49897	-0.93182	-1.01734	2.5279	1.6257E-32	2.4338
16	-0.64928	0.66716	1.31409	1.57696	1.26984	2.7863	9.6062E-33	3.3368
17	-0.25335	-0.56422	-1.88043	-1.57022	-1.07300	1.5220	1.2042E-32	1.9729
18	0.97519	-0.09238	1.23662	0.46585	0.15820	2.1448	2.1613E-32	0.5580
19	0.46464	-0.31899	1.60596	1.04298	0.90368	1.2923	9.6217E-34	1.3097
20	-1.40713	-0.56769	-1.42542	-2.55972	-1.62859	6.1425	2.5439E-32	3.3802