

NOTES DE STATISTIQUE ET D'INFORMATIQUE

90/1

(Réédition 2003)

LA CORRÉLATION CANONIQUE :
PRINCIPES ET APPLICATION

R. PALM

Faculté universitaire des Sciences agronomiques

GEMBLOUX

(Belgique)

LA CORRÉLATION CANONIQUE : PRINCIPES ET APPLICATION

R. PALM*

RÉSUMÉ

Cette note décrit les différentes étapes de l'interprétation d'une analyse des corrélations canoniques, à partir d'un exemple concret traité par le logiciel SAS. Elle rappelle également les liens qui existent entre l'analyse des corrélations canoniques et d'autres méthodes statistiques d'analyse multivariée.

SUMMARY

This note describes the different stages in interpreting a canonical correlation analysis. It is based on an example analysed by SAS software. Relationship with some other multivariate statistical methods is pointed out, too.

1. INTRODUCTION

L'*analyse des corrélations canoniques*¹ est une technique statistique utilisée pour étudier les relations existant entre deux groupes de variables. Les variables en question sont quantitatives, mais peuvent éventuellement résulter d'un codage du type 0/1, pour décrire, par exemple, une caractéristique qualitative de type présence ou absence.

Le principe de la corrélation canonique est le suivant. Pour chacun des deux groupes de variables, on détermine une nouvelle variable qui est une combinaison linéaire des variables du groupe et qui est appelée première *variable canonique*². Ces deux variables canoniques sont construites de telle sorte que leur coefficient de corrélation, appelé premier *coefficient de corrélation canonique*³ soit maximum. On détermine ensuite, dans chaque groupe, une deuxième variable canonique, non corrélée à la première, en calculant une autre combinaison linéaire des

* Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.

1. En anglais: *canonical correlation analysis, canonical analysis.*

2. En anglais: *canonical variate.*

3. En anglais: *canonical correlation coefficient.*

variables initiales. Ces combinaisons linéaires sont calculées de manière à assurer également la valeur maximum du second coefficient de corrélation canonique. Le processus de détermination des variables canoniques se poursuit jusqu'à ce que le nombre de couples de variables canoniques soit égal au nombre de variables initiales du groupe qui en comporte le moins, pour autant, du moins, que les matrices de corrélation des variables des deux groupes ne soient pas singulières.

L'objectif de cette note est d'illustrer l'interprétation des résultats d'une analyse des corrélations canoniques par un exemple concret et de présenter les relations qui existent entre l'analyse des corrélations canoniques et les autres méthodes statistiques multivariées.

Le paragraphe 2 sera consacré à l'exemple. Nous présenterons d'abord les données utilisées (paragraphe 2.1). Ensuite, nous passerons en revue les trois phases principales de l'interprétation d'une analyse des corrélations canoniques, qui sont l'examen et, éventuellement, les tests de signification des coefficients de corrélation canonique (paragraphe 2.2), l'examen des variables canoniques correspondantes (paragraphe 2.3) et l'analyse de la redondance, dont le but est de mesurer la qualité de la prédiction des variables initiales par les variables canoniques (paragraphe 2.4). Au paragraphe 3, nous examinerons les liens entre l'analyse des corrélations canoniques et les autres méthodes statistiques multivariées. Après un bref rappel théorique relatif au calcul des coefficients et des variables canoniques (paragraphe 3.1), nous envisagerons successivement les liens entre, d'une part, la corrélation canonique et, d'autre part, la régression multiple et la régression multiple multivariée (paragraphe 3.2), l'analyse de la variance multivariée et l'analyse factorielle discriminante (paragraphe 3.3), et l'analyse factorielle des correspondances (paragraphe 3.4). Enfin, quelques conclusions termineront cette note (paragraphe 4).

Les différentes figures illustrant le texte correspondent toutes à des extraits de documents imprimés obtenus par la procédure CANCORR du logiciel SAS. Des informations complémentaires relatives à cette procédure sont données dans le manuel d'utilisation de ce logiciel [X, 1985].

2. INTERPRÉTATION DES RÉSULTATS : UN EXEMPLE CONCRET

2.1. Matrice des données et examen des corrélations simples

L'exemple traité concerne l'étude de la relation existant, pour la chicorée witloof (*Cichorium intybus* L.), entre, d'une part, diverses caractéristiques de la plante et, d'autre part, les caractéristiques du chicon qu'elle produit au forçage [RONCHAI, 1962].

Les variables retenues et les symboles correspondants sont les suivants:

variables relatives à la plante mise en forçage:

PLPOIDSF = poids des feuilles (en grammes);
 PLPOIDSR = poids de la racine (en grammes);
 PLDIAMRC = diamètre de la racine au collet (en millimètres);
 PLDIAMFC = diamètre de la base du feuillage au collet (en millimètres);
 PLHAUTGF = hauteur de la plus grande feuille (en millimètres);
 PLLARGPE = largeur des pétioles (valeur 0 si les pétioles sont étroits et valeur 1 si les pétioles sont larges);
 PLPUBESC = présence (valeur 1) ou absence (valeur 0) de pubescence des feuilles;
 PLLOBES = présence (valeur 1) ou absence (valeur 0) de lobes sur les feuilles;
 PLGAUFRU = présence (valeur 1) ou absence (valeur 0) de gaufrure des feuilles;

variables relatives au chicon produit:

CHICLONG = longueur (en millimètres);
 CHICLARG = largeur (en millimètres);
 CHICPOID = poids (en grammes);
 CHICLAXE = longueur de l'axe (en millimètres);
 CHICQUAL = appréciation de la qualité, en fonction de son degré d'ouverture (valeur 0 si le chicon est de bonne qualité et valeur 1 si le chicon est de qualité inférieure).

On remarque la présence, dans chacun des deux groupes de caractéristiques, d'une ou plusieurs variables qualitatives codées sous la forme 0/1. De manière à améliorer la linéarité des relations entre les différentes variables quantitatives, celles-ci ont subi une transformation logarithmique. Par contre, les variables alternatives n'ont pas été transformées. Enfin, signalons encore que les caractéristiques énumérées ci-dessus ont été déterminées sur 1.000 individus. Des informations complémentaires concernant la description des variables et la justification de la transformation logarithmique sont données par RONCHINE [1962].

Les moyennes et les écarts-types des différentes variables sont données dans la figure 1, et les coefficients de corrélation simple entre les variables d'un même groupe et entre les variables de groupes différents sont données dans les figures 2 et 3.

L'examen de la matrice de corrélation des variables relatives aux plantes montre que les corrélations entre les variables correspondant à des mesures proprement dites sont, dans l'ensemble, assez élevées, à l'exception toutefois de la hauteur de la plus grande feuille. Par contre, les variables alternatives sont nettement moins corrélées, entre elles d'une part, et avec les autres variables du groupe, d'autre part.

Variable	Standard		Label
	Mean	Deviation	
PLPOIDSF	5.579349	0.363423	POIDS DES FEUILLES (LOG)
PLPOIDSR	4.730569	0.350282	POIDS DE LA RACINE (LOG)
PLDIAMRC	3.650984	0.143490	DIAM. RACINE AU COLLET (LOG)
PLDIAMFC	3.445512	0.146972	DIAM. BASE DU FEUILLAGE (LOG)
PLHAUTGF	6.258686	0.158010	HAUT. PLUS GRANDE FEUILLE (LOG)
PLLARGPE	0.168000	0.374053	LARGEUR DES PETIOLES (0/1)
PLPUBESC	0.315000	0.464748	PUBESCENCE (0/1)
PLLOBES	0.411000	0.492261	LOBES (0/1)
PLGAUFRU	0.106000	0.307992	GAUFRURE (0/1)
CHICLONG	5.245452	0.198207	LONGUEUR DU CHICON (LOG)
CHICLARG	3.622549	0.175346	LARGEUR DU CHICON (LOG)
CHICPOID	4.605199	0.344265	POIDS DU CHICON (LOG)
CHICLAXE	3.641365	0.577048	LONGUEUR AXE DU CHICON (LOG)
CHICQUAL	0.280000	0.449224	QUALITE DU CHICON (0/1)

Figure 1. Moyennes et écarts-types des différentes variables considérées.

Pour les variables relatives au chicon, les corrélations ne dépassent pas 0,6. Les valeurs les plus élevées concernent la largeur et le poids du chicon (0,59), la longueur de l'axe et le poids du chicon (0,46), et la longueur et le poids du chicon (0,39).

Enfin, la matrice de corrélation entre les variables relatives à la plante et celles relatives au chicon montre que les huit corrélations les plus élevées sont les corrélations entre l'une des quatre premières variables du premier groupe (poids des feuilles, poids de la racine, diamètre de la racine au collet, diamètre de la base du feuillage) et l'une des deux variables suivantes du second groupe : largeur du chicon et poids du chicon. Toutes ces corrélations sont comprises entre 0,37 et 0,55. Quant aux autres corrélations, elles sont toutes inférieures à 0,22. La corrélation la plus élevée de cette matrice étant égale à 0,55, on peut affirmer que le premier coefficient de corrélation canonique sera au moins égal à cette valeur, car le premier coefficient de corrélation canonique est toujours supérieur ou égal à la valeur absolue du coefficient de corrélation le plus grand, en valeur absolue, de la matrice de corrélation entre les variables des groupes différents.

On pourrait éventuellement aussi tester la signification de chacun des coefficients de corrélation. Il suffit pour cela de calculer la valeur critique [DAGNELIE, 1979-1980]:

$$r_{1-\alpha/2} = \frac{t_{1-\alpha/2}}{\sqrt{n-2 + t_{1-\alpha/2}^2}},$$

Correlations Among the PLANTE

	PLPOIDSF	PLPOIDSR	PLDIAMR	PLDIAMFC	PLHAUTGF
PLPOIDSF	1.0000	0.6842	0.6865	0.7848	0.1810
PLPOIDSR	0.6842	1.0000	0.8625	0.8037	0.0880
PLDIAMRC	0.6865	0.8625	1.0000	0.8857	0.0690
PLDIAMFC	0.7848	0.8037	0.8857	1.0000	0.0862
PLHAUTGF	0.1810	0.0880	0.0690	0.0862	1.0000
PLLARGPE	0.0224	0.0207	0.0040	0.0014	0.0497
PLPUBESC	0.0372	0.0603	0.0415	0.0645	0.0453
PLLOBES	0.2373	0.2241	0.2136	0.2169	0.0510
PLGAUFRU	0.3366	0.2720	0.2858	0.2898	0.0136

Correlations Among the PLANTE

	PLLARGPE	PLPUBESC	PLLOBES	PLGAUFRU
PLPOIDSF	0.0224	0.0372	0.2373	0.3366
PLPOIDSR	0.0207	0.0603	0.2241	0.2720
PLDIAMRC	0.0040	0.0415	0.2136	0.2858
PLDIAMFC	0.0014	0.0645	0.2169	0.2898
PLHAUTGF	0.0497	0.0453	0.0510	0.0136
PLLARGPE	1.0000	0.0638	0.0867	-0.0070
PLPUBESC	0.0638	1.0000	-0.0633	-0.0866
PLLOBES	0.0867	-0.0633	1.0000	0.1349
PLGAUFRU	-0.0070	-0.0866	0.1349	1.0000

Correlations Among the CHICON

	CHICLONG	CHICLARG	CHICPOID	CHICLAXE	CHICQUAL
CHICLONG	1.0000	-0.1145	0.3927	0.2611	-0.3070
CHICLARG	-0.1145	1.0000	0.5896	0.2627	0.1744
CHICPOID	0.3927	0.5896	1.0000	0.4612	0.0897
CHICLAXE	0.2611	0.2627	0.4612	1.0000	-0.0259
CHICQUAL	-0.3070	0.1744	0.0897	-0.0259	1.0000

Figure 2. Coefficients de corrélation simple entre les variables d'un même groupe.

Correlations Between the PLANTE and the CHICON

	CHICLONG	CHICLARG	CHICPOID	CHICLAXE	CHICQUAL
PLPOIDSF	0.0349	0.3712	0.4376	0.2145	0.0234
PLPOIDSR	-0.0676	0.5176	0.5539	0.1723	0.1269
PLDIAMRC	-0.0195	0.5061	0.5272	0.1546	0.0734
PLDIAMFC	-0.0001	0.4579	0.5072	0.1442	0.0425
PLHAUTGF	0.1114	-0.0132	0.0829	0.0816	-0.0701
PLLARGPE	-0.0298	0.0292	0.0171	-0.0181	-0.0002
PLPUBESC	-0.0461	-0.0172	-0.0121	-0.0211	0.0038
PLLOBES	0.0122	0.1156	0.1543	0.0055	0.0042
PLGAUFRU	-0.0660	0.2164	0.1873	0.0962	0.0674

Figure 3. Coefficients de corrélation simple entre les variables de groupes différents.

soit, pour $\alpha = 0,05$ et $n = 1.000$, $r_{1-\alpha/2} = 0,062$, et de rejeter l'hypothèse de nullité du coefficient de corrélation si la corrélation calculée sur les données de l'échantillon est supérieure, en valeur absolue, à cette valeur.

Pour l'exemple qui nous concerne, l'intérêt de ces tests est cependant assez limité, car compte tenu de l'importance de l'effectif, les tests conduisent au rejet de l'hypothèse de nullité pour des coefficients de corrélation dont la valeur est très faible. Bien que significatifs, de tels coefficients sont sans grand intérêt, car on s'intéresse bien plus à l'intensité des liaisons qu'à la recherche de l'indépendance entre les caractères. D'autre part, d'un point de vue théorique, la réalisation de près d'une centaine de tests augmente de façon importante le risque global de première espèce [DAGNELIE, 1982].

2.2. Coefficients de corrélation canonique et tests de signification

La figure 4 reprend les informations qui concernent les coefficients de corrélation canonique et leur test de signification.

Tout d'abord, on vérifie qu'on dispose bien de cinq coefficients de corrélation canonique, puisque le groupe de variables qui comporte le moins de variables est le groupe relatif au chicon, avec précisément cinq variables. On vérifie aussi que les coefficients ont des valeurs progressivement décroissantes et que le premier coefficient est supérieur au coefficient de corrélation simple le plus élevé, en valeur absolue, de la matrice de corrélation entre les variables des groupes différents.

Les coefficients de corrélation canonique donnés dans la première colonne de la figure 4 sont des estimations biaisées des coefficients de corrélation canonique théoriques et LAWLEY [1959] a montré que des estimations moins biaisées pouvaient être obtenues, du moins lorsque les coefficients de corrélation canonique théoriques ne présentent ni des valeurs trop proches les unes des autres, ni des

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.660496	0.655446	0.017836	0.436256
2	0.212107	.	0.030215	0.044989
3	0.174263	.	0.030678	0.030367
4	0.118993	.	0.031191	0.014159
5	0.050299	.	0.031559	0.002530

Eigenvalues of $\text{Inv}(E) * H = \text{CanRs}q / (1 - \text{CanRs}q)$

	Eigenvalue	Difference	Proportion	Cumulative
1	0.7739	0.7267	0.8903	0.8903
2	0.0471	0.0158	0.0542	0.9445
3	0.0313	0.0170	0.0360	0.9806
4	0.0144	0.0118	0.0165	0.9971
5	0.0025		0.0029	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.51333900	15.77	45	4413.7	<.0001
2	0.91058816	2.93	32	3641.5	<.0001
3	0.95348474	2.26	21	2837.6	0.0009
4	0.98334650	1.39	12	1978	0.1632
5	0.99747002	0.50	5	990	0.7747

Multivariate Statistics and F Approximations

S=5 M=1.5 N=492

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.51333900	15.77	45	4413.7	<.0001
Pillai's Trace	0.52830165	13.00	45	4950	<.0001
Hotelling-Lawley Trace	0.86917977	19.02	45	3074.3	<.0001
Roy's Greatest Root	0.77385347	85.12	9	990	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Figure 4. Coefficients de corrélation canonique et informations associées.

valeurs trop faibles. Le logiciel SAS donne ces coefficients de corrélation canonique ajustés, pour autant qu'ils puissent être calculés et pour autant aussi qu'un coefficient ajusté ne soit pas supérieur au précédent [X, 1985].

L'erreur-standard approximative des coefficients de corrélation canonique est également donnée dans la figure 4. Elle est égale à:

$$\frac{1 - r_k^2}{\sqrt{n}},$$

r_k étant le $k^{ième}$ coefficient de corrélation canonique observé et n l'effectif de l'échantillon. On retrouve, en fait, la formule de l'erreur-standard relative à un coefficient de corrélation classique [DAGNELIE, 1979-1980].

Les tests de signification des coefficients de corrélation canonique peuvent être réalisés par l'intermédiaire des variables Λ de WILKS [DAGNELIE, 1982]. Pour tester la signification du premier coefficient de corrélation canonique, on émet l'hypothèse nulle suivante:

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_5 = 0,$$

ρ_k ($k = 1, \dots, 5$) étant les k coefficients de corrélation canonique théoriques.

La valeur Λ_{obs} est égale à:

$$\Lambda_{obs} = \prod_{k=1}^5 (1 - r_k^2) = 0,51334,$$

et cette valeur doit être comparée à la valeur théorique Λ_α , relative à la variable de WILKS à $p = 9$, $q = 5$ et $r = 1.000 - 5 - 1 = 994$ degrés de liberté, ou encore à $p = 5$, $q = 9$ et $r = 1.000 - 9 - 1 = 990$ degrés de liberté, en raison de l'identité entre les distributions de WILKS de paramètres p , q et r , d'une part, et q , p et $q + r - p$, d'autre part [DAGNELIE, 1982].

La relation qui existe entre les variables de WILKS et les variables F de SNEDECOR, permet d'affirmer que la quantité:

$$F_{obs} = \frac{1 - \Lambda_{obs}^{1/t}}{\Lambda_{obs}^{1/t}} \frac{st - 2u}{pq},$$

avec:

$$s = r - \frac{p - q + 1}{2}, \quad t = \sqrt{\frac{p^2 q^2 - 4}{p^2 + q^2 - 5}},$$

si $p^2 + q^2 - 5$ est positif, et $t = 1$ sinon, et:

$$u = \frac{pq - 2}{4},$$

est une valeur observée d'une variable aléatoire approximativement distribuée selon une loi F de SNEDECOR à $k_1 = pq$ et $k_2 = st - 2u$ degrés de liberté [RAO, 1952]. On a, par conséquent:

$$F_{obs} = 15,77,$$

les nombres de degrés de liberté étant respectivement égaux à 45 et 4.413,72. La probabilité pour que la variable F de SNEDECOR soit supérieure à F_{obs} étant pratiquement égale à zéro, on rejette l'hypothèse de nullité des cinq coefficients de corrélation canonique. On conclut donc aussi que le premier coefficient de corrélation canonique est significatif.

Pour le test de signification du deuxième coefficient de corrélation canonique, on a :

$$\Lambda_{obs} = \prod_{k=2}^5 (1 - r_k^2) = 0,91059,$$

et les nombres de degrés de liberté de la variable de WILKS correspondante sont égaux à $p = 8$, $q = 4$ et $r = 994$. En utilisant la transformation donnée ci-dessus, on trouve :

$$F_{obs} = 2,93,$$

avec $k_1 = 32$ et $k_2 = 3.641,47$ degrés de liberté et on conclut que le deuxième coefficient de corrélation canonique est significatif.

Des calculs analogues permettent de tester la signification des trois autres coefficients de corrélation canonique. Toutefois, dans le cas des deux derniers coefficients, la formule de passage de la variable de WILKS à la variable de SNEDECOR est une relation exacte. Pour le quatrième coefficient, on a, en effet, une variable de WILKS à $p = 6$, $q = 2$ et $r = 994$ degrés de liberté et on peut déterminer la valeur F_{obs} en utilisant une relation plus simple [DAGNELIE, 1982] :

$$F_{obs} = \frac{r - p + 1}{p} \frac{1 - \sqrt{\Lambda_{obs}}}{\sqrt{\Lambda_{obs}}} = \frac{994 - 6 + 1}{6} \frac{1 - \sqrt{0,98335}}{\sqrt{0,98335}} = 1,39,$$

avec $k_1 = 2p = 12$ et $k_2 = 2(r - p + 1) = 1.978$ degrés de liberté. De même, pour le cinquième coefficient, on a une variable de WILKS à $p = 5$, $q = 1$ et $r = 994$ degrés de liberté, et donc aussi [DAGNELIE, 1982] :

$$F_{obs} = \frac{r - p + 1}{p} \frac{1 - \Lambda_{obs}}{\Lambda_{obs}} = \frac{994 - 5 + 1}{5} \frac{1 - 0,99747}{0,99747} = 0,50,$$

avec $k_1 = p = 5$ et $k_2 = (r - p + 1) = 990$ degrés de liberté.

L'examen des probabilités associées aux différents tests montre que les deux derniers coefficients de corrélation sont non significatifs.

On notera cependant que les différents tests de signification des coefficients de corrélation canonique ne sont qu'approximatifs, en raison du non-respect des conditions d'application. En effet, pour les problèmes d'inférence statistique relatifs à la corrélation canonique, on suppose, d'une façon générale, que les q variables du premier groupe et les p variables du deuxième groupe ont une distribution normale à $q + p$ dimensions, et que l'échantillon est aléatoire et simple [DAGNELIE, 1982]. Cette condition de normalité à $q + p$ dimensions n'est évidemment pas vérifiée dans cet exemple, compte tenu de la présence, dans chacun des deux groupes de variables, d'une ou plusieurs variables alternatives. Cependant,

du fait de la taille élevée de l'échantillon, les tests restent valables, du moins de façon approximative [KSHIRSAGAR, 1972]. Le problème de la multinormalité de la population-parent, en relation avec les tests de signification des coefficients de corrélation canonique, sera encore abordé au paragraphe 3.3.

La figure 4 donne également des informations concernant les valeurs propres de la matrice $E^{-1}H$, ainsi que des informations relatives à différents tests multivariés. Des commentaires à ces sujets seront présentés aux paragraphes 3.2 et 3.3.

2.3. Variables canoniques

La figure 5 donne les *coefficients canoniques*⁴, c'est-à-dire les coefficients qui interviennent dans les combinaisons linéaires constituant les variables canoniques. A partir de ces coefficients et des observations centrées réduites, on peut déterminer la valeur des variables canoniques pour chacun des individus.

A partir des données brutes, qui figurent partiellement en annexe, des moyennes et des écarts-types des variables initiales (figure 1), on peut, à titre d'exemple, calculer, pour le premier individu, les observations centrées réduites relatives aux cinq caractéristiques du chicon. On obtient:

$$\begin{aligned}(5, 1930 - 5, 2455)/0, 1982 &= -0, 2649, \\(3, 5264 - 3, 6225)/0, 1753 &= -0, 5482, \\(4, 2767 - 4, 6052)/0, 3443 &= -0, 9541, \\(4, 4886 - 3, 6414)/0, 5770 &= 1, 4683, \\(0 - 0, 28)/0, 4492 &= -0, 6233.\end{aligned}$$

Pour le premier individu, la première variable canonique relative aux caractéristiques du chicon vaut donc:

$$\begin{aligned}(-0, 4255)(-0, 2649) + (0, 2623)(-0, 5482) + (0, 9353)(-0, 9541) \\+ (-0, 1417)(1, 4683) + (-0, 0822)(-0, 6233) = -1, 0803.\end{aligned}$$

Aux arrondis près, ce résultat correspond bien à la valeur donnée dans l'annexe et égale à $-1, 0805$.

Des calculs similaires peuvent évidemment être réalisés pour tous les individus et pour toutes les variables canoniques et on pourrait vérifier que les variables canoniques ainsi créées présentent les caractéristiques suivantes:

- la corrélation entre la $j^{ième}$ variable canonique d'un groupe et la $k^{ième}$ variable canonique de ce même groupe est nulle;
- la corrélation entre la $j^{ième}$ variable canonique d'un groupe et la $k^{ième}$ variable canonique de l'autre groupe est nulle si j est différent de k et est, par définition, égale au $k^{ième}$ coefficient de corrélation canonique si j est égal à k .

4. En anglais: *canonical coefficients*.

Canonical Correlation Analysis

Standardized Canonical Coefficients for the PLANTE

		PLANTE1	PLANTE2	PLANTE3
PLPOIDSF	POIDS DES FEUILLES (LOG)	-0.0712	0.8434	-1.1129
PLPOIDSR	POIDS DE LA RACINE (LOG)	0.7342	-1.0320	-0.6716
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	0.1725	0.5164	0.1232
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.1530	0.0685	1.5345
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	-0.0481	0.5353	0.0516
PLLARGPE	LARGEUR DES PETIOLES (0/1)	0.0481	-0.1245	0.1115
PLPUBESC	PUBESCENCE (0/1)	-0.0407	-0.2187	-0.0061
PLLOBES	LOBES (0/1)	0.0207	0.0034	0.4133
PLGAUFRU	GAUFRURE (0/1)	0.0898	-0.3182	-0.3268

Standardized Canonical Coefficients for the PLANTE

		PLANTE4	PLANTE5
PLPOIDSF	POIDS DES FEUILLES (LOG)	-0.1467	0.2479
PLPOIDSR	POIDS DE LA RACINE (LOG)	1.1976	-0.7302
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	-1.9892	-0.2967
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.8642	0.7878
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	0.3767	-0.1568
PLLARGPE	LARGEUR DES PETIOLES (0/1)	-0.0593	0.4210
PLPUBESC	PUBESCENCE (0/1)	0.2806	0.5760
PLLOBES	LOBES (0/1)	0.2559	-0.1712
PLGAUFRU	GAUFRURE (0/1)	-0.1910	0.4212

Standardized Canonical Coefficients for the CHICON

		CHICON1	CHICON2	CHICON3
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.4255	0.6260	0.2323
CHICLARG	LARGEUR DU CHICON (LOG)	0.2623	0.0023	0.0700
CHICPOID	POIDS DU CHICON (LOG)	0.9353	0.0651	0.3452
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	-0.1417	0.3266	-1.0556
CHICQUAL	QUALITE DU CHICON (0/1)	-0.0822	-0.4191	-0.3497

Standardized Canonical Coefficients for the CHICON

		CHICON4	CHICON5
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.6670	-0.8103
CHICLARG	LARGEUR DU CHICON (LOG)	-1.3742	-0.0020
CHICPOID	POIDS DU CHICON (LOG)	1.2570	0.0440
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	-0.0389	0.2095
CHICQUAL	QUALITE DU CHICON (0/1)	-0.0620	-0.9307

Figure 5. Coefficients des variables canoniques.

On peut constater également que, contrairement à l'analyse en composantes principales, les vecteurs des coefficients canoniques ne sont pas de norme 1. Pour la variable CHICON1, c'est-à-dire pour la première variable canonique du groupe de variables décrivant le chicon, on a, par exemple:

$$(-0,4255)^2 + (0,2623)^2 + \dots + (-0,0822)^2 = 1,1515.$$

Cette discordance par rapport à l'analyse en composantes principales résulte du fait qu'en analyse des corrélations canoniques, les vecteurs des coefficients canoniques sont standardisés de manière à obtenir des variables canoniques d'écart-type unitaire.

De même, le produit scalaire des vecteurs relatifs à deux variables canoniques du même groupe n'est pas nul, contrairement à l'analyse en composantes principales. Pour les variables CHICON1 et CHICON2, par exemple, on a:

$$(-0,4255)(0,6260) + (0,2623)(0,0023) + \dots + (-0,0822)(-0,4191) = -0,2167.$$

Les coefficients canoniques représentent les poids des différentes variables initiales centrées réduites dans la détermination de la variable canonique. Afin de distinguer les variables centrées réduites des variables initiales, nous allons adopter la convention suivante: les variables centrées réduites seront désignées par le symbole de la variable initiale correspondante, affecté d'un astérisque. Ainsi, par exemple, PLPOIDSR* représente la variable centrée réduite correspondant au poids de la racine.

L'examen des coefficients canoniques montre, par exemple, que la variable PLANT1 est avant tout une fonction croissante de la variable PLPOIDSR*, alors que la variable PLANT2 est essentiellement une fonction croissante des variables PLPOIDSF*, PLDIAMRC* et PLHAUTGF* et une fonction décroissante de la variable PLPOIDSR*. On voit également que la variable CHICON1 est surtout fonction de la variable CHICPOID* et que CHICON2 est une fonction croissante de CHICLONG* et, dans une mesure moindre, de CHICLAXE* et une fonction décroissante de CHICQUAL*.

Les figures 6 et 7 donnent les corrélations entre les variables canoniques et les variables de départ. On peut constater, par exemple, que la variable canonique PLANTE1 est fortement corrélée aux variables PLPOIDSR, PLDIAMRC et PLDIAMFC et, dans une mesure moindre, à PLPOIDSF. Elle est également moyennement corrélée à deux caractéristiques relatives aux chicons (CHICLARG et CHICPOID). Quant à la variable canonique CHICON1, elle est fortement corrélée à CHICLARG et CHICPOID, d'une part, et moyennement corrélée à PLPOIDSF, PLPOIDSR, PLDIAMRC et PLDIAMFC, d'autre part.

La valeur relativement élevée du premier coefficient de corrélation canonique traduit donc essentiellement les relations entre, d'une part, les quatre premières variables relatives aux plantes (PLPOIDSF, PLPOIDSR, PLDIAMRC et PLDIAMFC) et, d'autre part, les variables CHICLARG et CHICPOID. L'examen de la matrice de corrélation entre les variables des deux groupes (figure 2) avait d'ailleurs déjà mis en évidence l'importance de ces relations.

Canonical Structure

Correlations Between the PLANTE and Their Canonical Variables

		PLANTE1	PLANTE2	PLANTE3
PLPOIDSF	POIDS DES FEUILLES (LOG)	0.6957	0.5252	-0.2838
PLPOIDSR	POIDS DE LA RACINE (LOG)	0.9807	-0.0090	-0.0834
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	0.9178	0.2030	0.1376
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.8638	0.2987	0.2295
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	0.0316	0.6183	-0.0463
PLLARGPE	LARGEUR DES PETIOLES (0/1)	0.0588	-0.1096	0.1157
PLPUBESC	PUBESCENCE (0/1)	0.0097	-0.1801	0.0278
PLLOBES	LOBES (0/1)	0.2548	0.0848	0.3264
PLGAUFRU	GAUFRURE (0/1)	0.3646	-0.1200	-0.3480

Correlations Between the PLANTE and Their Canonical Variables

		PLANTE4	PLANTE5
PLPOIDSF	POIDS DES FEUILLES (LOG)	0.0590	0.2665
PLPOIDSR	POIDS DE LA RACINE (LOG)	0.1302	-0.0775
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	-0.2541	0.0401
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.0003	0.2420
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	0.4131	-0.0847
PLLARGPE	LARGEUR DES PETIOLES (0/1)	0.0157	0.4225
PLPUBESC	PUBESCENCE (0/1)	0.3341	0.5738
PLLOBES	LOBES (0/1)	0.2226	-0.1197
PLGAUFRU	GAUFRURE (0/1)	-0.2169	0.3714

Correlations Between the CHICON and Their Canonical Variables

		CHICON1	CHICON2	CHICON3
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.1000	0.8652	0.1916
CHICLARG	LARGEUR DU CHICON (LOG)	0.8109	-0.0182	-0.0914
CHICPOID	POIDS DU CHICON (LOG)	0.8501	0.4254	-0.0405
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	0.2496	0.5315	-0.8083
CHICQUAL	QUALITE DU CHICON (0/1)	0.1817	-0.6135	-0.3505

Correlations Between the CHICON and Their Canonical Variables

		CHICON4	CHICON5
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.0072	-0.4523
CHICLARG	LARGEUR DU CHICON (LOG)	-0.5777	0.0094
CHICPOID	POIDS DU CHICON (LOG)	0.1613	-0.2622
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	0.0073	0.0418
CHICQUAL	QUALITE DU CHICON (0/1)	0.0168	-0.6837

Figure 6. Corrélations entre les variables de départ et les variables canoniques.

Correlations Between the PLANTE and the Canonical Variables
of the CHICON

		CHICON1	CHICON2	CHICON3
PLPOIDSF	POIDS DES FEUILLES (LOG)	0.4595	0.1114	-0.0495
PLPOIDSR	POIDS DE LA RACINE (LOG)	0.6477	-0.0019	-0.0145
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	0.6062	0.0431	0.0240
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.5706	0.0634	0.0400
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	0.0208	0.1311	-0.0081
PLLARGPE	LARGEUR DES PETIOLES (0/1)	0.0389	-0.0233	0.0202
PLPUBESC	PUBESCENCE (0/1)	0.0064	-0.0382	0.0048
PLLOBES	LOBES (0/1)	0.1683	0.0180	0.0569
PLGAUFRU	GAUFRURE (0/1)	0.2408	-0.0255	-0.0606

Correlations Between the PLANTE and the Canonical Variables
of the CHICON

		CHICON4	CHICON5
PLPOIDSF	POIDS DES FEUILLES (LOG)	0.0070	0.0134
PLPOIDSR	POIDS DE LA RACINE (LOG)	0.0155	-0.0039
PLDIAMRC	DIAM. RACINE AU COLLET (LOG)	-0.0302	0.0020
PLDIAMFC	DIAM. BASE DU FEUILLAGE (LOG)	0.0000	0.0122
PLHAUTGF	HAUT. PLUS GRANDE FEUILLE (LOG)	0.0492	-0.0043
PLLARGPE	LARGEUR DES PETIOLES (0/1)	0.0019	0.0212
PLPUBESC	PUBESCENCE (0/1)	0.0398	0.0289
PLLOBES	LOBES (0/1)	0.0265	-0.0060
PLGAUFRU	GAUFRURE (0/1)	-0.0258	0.0187

Correlations Between the CHICON and the Canonical Variables
of the PLANTE

		PLANTE1	PLANTE2	PLANTE3
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.0660	0.1835	0.0334
CHICLARG	LARGEUR DU CHICON (LOG)	0.5356	-0.0039	-0.0159
CHICPOID	POIDS DU CHICON (LOG)	0.5615	0.0902	-0.0071
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	0.1648	0.1127	-0.1409
CHICQUAL	QUALITE DU CHICON (0/1)	0.1200	-0.1301	-0.0611

Correlations Between the CHICON and the Canonical Variables
of the PLANTE

		PLANTE4	PLANTE5
CHICLONG	LONGUEUR DU CHICON (LOG)	-0.0009	-0.0228
CHICLARG	LARGEUR DU CHICON (LOG)	-0.0687	0.0005
CHICPOID	POIDS DU CHICON (LOG)	0.0192	-0.0132
CHICLAXE	LONGUEUR AXE DU CHICON (LOG)	0.0009	0.0021
CHICQUAL	QUALITE DU CHICON (0/1)	0.0020	-0.0344

Figure 7. Corrélations entre les variables de départ et les variables canoniques (suite).

Par contre les variables canoniques PLANT2 et PLANT3 ne sont que faiblement corrélées aux variables du groupe "chicon" et les variables canoniques CHICON2 et CHICON3 ne sont, elles aussi, que faiblement corrélées aux variables du groupe "plante". Parmi toutes ces corrélations croisées, la plus élevée n'atteint, en effet, que 0,18. Ceci explique les valeurs relativement faibles, quoique significatives, des deuxième et troisième coefficients de corrélation canonique.

2.4. Analyse de la redondance

L'*analyse de la redondance*⁵ a pour objectif d'examiner dans quelle mesure les variables de départ peuvent être estimées à partir des variables canoniques [COOLEY et LOHNES, 1971; X, 1985]. Pour l'exemple considéré, les figures 8 et 9 donnent les informations relatives à cette analyse.

L'examen de la première partie de la figure 8 montre que 35,96 ou 36 % de la variabilité des variables du groupe "plante" sont expliqués par la première variable canonique de ce groupe et que 15,69 ou 16 % de cette même variabilité sont expliqués par la première variable canonique du groupe "chicon".

La valeur 0,3596 est en fait égale à la moyenne des carrés des coefficients de corrélation de la variable PLANTE1 et des variables initiales du groupe "plante" (figure 6):

$$(0,6957^2 + 0,9807^2 + \dots + 0,3646^2)/9 = 0,3596.$$

De même, la valeur 0,1569 est égale à la moyenne des carrés des coefficients de corrélation de la variable CHICON1 et des variables initiales du groupe "plante" (figure 7):

$$(0,4595^2 + 0,6477^2 + \dots + 0,2408^2)/9 = 0,1569.$$

Cette valeur est aussi égale à la part de la variabilité des variables du groupe "plante" expliquée par la variable canonique PLANTE1, multipliée par le carré du premier coefficient de corrélation canonique:

$$(0,3596)(0,4363) = 0,1569,$$

et on peut, par conséquent, donner à cette valeur l'interprétation suivante: par l'intermédiaire du premier couple de variables canoniques (PLANTE1 et CHICON1), les variables du groupe "chicon" expliquent 16 % de la variabilité des variables du groupe "plante".

On ne perdra cependant pas de vue que la valeur 0,1569 ne représente que la proportion moyenne de la variance des variables du groupe "plante" expliquée par l'intermédiaire du premier couple de variables canoniques, par les variables du groupe "chicon" et que des différences importantes peuvent exister entre les variables. Ainsi, la figure 9 montre que les coefficients de détermination individuels des variables du groupe "plante" et de CHICON1 varient de 0, pour la variable PLPUBESC, à 0,4196, pour la variable PLPOIDSR. De façon plus générale, on constate que, pour les quatre premières variables, le coefficient de

5. En anglais: *redundancy analysis*.

Standardized Variance of the PLANTE Explained by					
Their Own			The Opposite		
Canonical Variables			Canonical Variables		
Canonical Variable	Cumulative	Canonical	Cumulative	Canonical	Cumulative
Number	Proportion	Proportion	R-Square	Proportion	Proportion
1	0.3596	0.3596	0.4363	0.1569	0.1569
2	0.0950	0.4546	0.0450	0.0043	0.1612
3	0.0448	0.4994	0.0304	0.0014	0.1625
4	0.0516	0.5510	0.0142	0.0007	0.1633
5	0.0894	0.6403	0.0025	0.0002	0.1635

Standardized Variance of the CHICON Explained by					
Their Own			The Opposite		
Canonical Variables			Canonical Variables		
Canonical Variable	Cumulative	Canonical	Cumulative	Canonical	Cumulative
Number	Proportion	Proportion	R-Square	Proportion	Proportion
1	0.2971	0.2971	0.4363	0.1296	0.1296
2	0.3178	0.6149	0.0450	0.0143	0.1439
3	0.1646	0.7795	0.0304	0.0050	0.1489
4	0.0720	0.8515	0.0142	0.0010	0.1499
5	0.1485	1.0000	0.0025	0.0004	0.1503

Figure 8. Analyse de la redondance.

détermination varie de 0,21 à 0,42, alors que, pour les cinq dernières variables, il est très faible. Cette constatation est directement liée au commentaire donné à la fin du paragraphe 2.3.

La seconde partie de la figure 8 montre que, en moyenne, 29,71 ou 30 % de la variabilité des variables du groupe "chicon" sont expliqués par la variable CHICON1 et que 12,96 ou 13 % de cette variabilité sont expliqués, en moyenne, par la variable PLANTE1. Cette dernière valeur montre que par l'intermédiaire du premier couple de variables canoniques (PLANTE1 et CHICON1), les variables du groupe "plante" expliquent, en moyenne 13 % de la variabilité des variables du groupe "chicon". Toutefois, ici aussi, des différences importantes existent entre les variables: pour les variables CHICLARG et CHICPOID, les coefficients de détermination sont respectivement égaux à 0,29 et 0,32, alors que, pour les trois autres variables, ils sont pratiquement nuls (figure 9). Ceci confirme, encore une fois, le commentaire de la fin du paragraphe 2.3.

Canonical Redundancy Analysis

Squared Multiple Correlations Between the PLANTE
and the First M Canonical Variables of the CHICON

M	1	2	3
PLPOIDSF POIDS DES FEUILLES (LOG)	0.2112	0.2236	0.2260
PLPOIDSR POIDS DE LA RACINE (LOG)	0.4196	0.4196	0.4198
PLDIAMRC DIAM. RACINE AU COLLET (LOG)	0.3674	0.3693	0.3699
PLDIAMFC DIAM. BASE DU FEUILLAGE (LOG)	0.3255	0.3296	0.3312
PLHAUTGF HAUT. PLUS GRANDE FEUILLE (LOG)	0.0004	0.0176	0.0177
PLLARGPE LARGEUR DES PETIOLES (0/1)	0.0015	0.0021	0.0025
PLPUBESC PUBESCENCE (0/1)	0.0000	0.0015	0.0015
PLLOBES LOBES (0/1)	0.0283	0.0286	0.0319
PLGAUFRU GAUFRURE (0/1)	0.0580	0.0586	0.0623

Squared Multiple Correlations Between the PLANTE
and the First M Canonical Variables of the CHICON

M	4	5
PLPOIDSF POIDS DES FEUILLES (LOG)	0.2261	0.2262
PLPOIDSR POIDS DE LA RACINE (LOG)	0.4200	0.4200
PLDIAMRC DIAM. RACINE AU COLLET (LOG)	0.3708	0.3708
PLDIAMFC DIAM. BASE DU FEUILLAGE (LOG)	0.3312	0.3313
PLHAUTGF HAUT. PLUS GRANDE FEUILLE (LOG)	0.0201	0.0201
PLLARGPE LARGEUR DES PETIOLES (0/1)	0.0025	0.0029
PLPUBESC PUBESCENCE (0/1)	0.0031	0.0039
PLLOBES LOBES (0/1)	0.0326	0.0326
PLGAUFRU GAUFRURE (0/1)	0.0630	0.0633

Squared Multiple Correlations Between the CHICON
and the First M Canonical Variables of the PLANTE

M	1	2	3
CHICLONG LONGUEUR DU CHICON (LOG)	0.0044	0.0380	0.0392
CHICLARG LARGEUR DU CHICON (LOG)	0.2869	0.2869	0.2871
CHICPOID POIDS DU CHICON (LOG)	0.3153	0.3234	0.3235
CHICLAXE LONGUEUR AXE DU CHICON (LOG)	0.0272	0.0399	0.0597
CHICQUAL QUALITE DU CHICON (0/1)	0.0144	0.0313	0.0351

Squared Multiple Correlations Between the CHICON
and the First M Canonical Variables of the PLANTE

M	4	5
CHICLONG LONGUEUR DU CHICON (LOG)	0.0392	0.0397
CHICLARG LARGEUR DU CHICON (LOG)	0.2919	0.2919
CHICPOID POIDS DU CHICON (LOG)	0.3238	0.3240
CHICLAXE LONGUEUR AXE DU CHICON (LOG)	0.0597	0.0597
CHICQUAL QUALITE DU CHICON (0/1)	0.0351	0.0363

Figure 9. Analyse de la redondance (suite).

On constate donc que la redondance des variables du groupe "plante", étant donnée la disponibilité des variables du groupe "chicon", égale à 16 %, est différente de la redondance des variables du groupe "chicon", étant donnée la disponibilité des variables du groupe "plante", qui est égale à 13 %. Alors que le carré du premier coefficient de corrélation canonique exprime la part de la variance commune à la première variable canonique de chaque groupe, c'est-à-dire aussi leur recouvrement, la redondance d'un groupe exprime le véritable recouvrement des variables du groupe par les variables de l'autre groupe, par l'intermédiaire du premier couple de variables canoniques [COOLEY et LOHNES, 1971].

Quant aux coefficients de redondance relatifs aux autres couples de variables canoniques, ils sont tous très faibles, la valeur la plus élevée étant de l'ordre de 1 % (figure 8). La figure 9 confirme également que la prise en considération de plusieurs variables canoniques n'améliore guère la part de la variance expliquée de chacune des variables initiales: les coefficients de détermination multiples sont, en effet, du même ordre de grandeur que les coefficients de détermination simple, quel que soit le nombre de variables canoniques prises en considération.

Enfin, l'examen de la figure 8 permet encore de faire deux remarques générales. Tout d'abord, on constate que les proportions de variance des variables d'un groupe donné expliquées par les différentes variables canoniques du même groupe ne sont pas nécessairement décroissantes. Ainsi, par exemple, CHICON4 explique 7,2 % de la variance des variables du groupe "chicon", alors que CHICON5 explique 14,9 % de cette même variance. De telles situations peuvent se présenter tout à fait normalement, car les variables canoniques sont déterminées de manière à maximiser les corrélations canoniques et non pas les corrélations avec les variables de leur groupe. D'autre part, on constate aussi que, pour le groupe "chicon", la proportion cumulée de la variance des variables du groupe expliquée par l'ensemble des variables canoniques du même groupe est égale à l'unité. Pour le groupe "plante", par contre, la proportion cumulée n'atteint pas l'unité, mais est égale à 0,64. L'explication en est la suivante: pour le groupe qui contient le moins de variables (ici, le groupe "chicon"), on remplace les cinq variables initiales par cinq combinaisons linéaires, non corrélées, de ces variables initiales, ce qui conserve toute l'information, alors que pour l'autre groupe (ici, le groupe "plante"), on remplace les neuf variables initiales par cinq combinaisons linéaires non corrélées, ce qui se traduit inévitablement par une perte d'information.

3. RELATION ENTRE LA CORRÉLATION CANONIQUE ET D'AUTRES MÉTHODES STATISTIQUES MULTIVARIÉES

3.1. Calcul des coefficients de corrélation canonique et des variables canoniques

Des informations relatives à la détermination des coefficients de corrélation canonique et des variables canoniques sont données dans la plupart des livres d'analyse statistique multivariée et, notamment, par CHATFIELD et COLLINS [1980], COOLEY et LOHNES [1971], DAGNELIE [1982], KSHIRSAGAR [1972],

MORRISON [1967] et TATSUKOA [1971]. Nous nous limiterons ici à la présentation rapide des diverses formules de calcul qui peuvent être employées, afin de permettre la mise en évidence, dans les paragraphes ultérieurs, des relations existant entre l'analyse des corrélations canoniques et plusieurs autres méthodes statistiques multivariées.

Supposons que les $q + p$ variables se répartissent en deux groupes, le premier groupe étant constitué des q variables y_1, \dots, y_q et le second groupe comprenant les p variables x_1, \dots, x_p . Supposons, en outre, que $q \leq p$. La matrice de corrélation des $q + p$ variables peut être subdivisée en quatre sous-matrices:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}.$$

Les matrices \mathbf{R}_{11} et \mathbf{R}_{22} , de dimensions $q \times q$ et $p \times p$ donnent, respectivement, les corrélations entre les variables du premier et du second groupe. La matrice \mathbf{R}_{12} , de dimensions $q \times p$, et sa transposée \mathbf{R}_{21} donnent les corrélations entre les couples de variables, lorsqu'une variable appartient au premier groupe et l'autre variable au second groupe.

On peut montrer que les coefficients de corrélation canonique sont les solutions positives, r_k ($k = 1, \dots, q$), de l'équation matricielle:

$$| \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21} - r_k^2 \mathbf{R}_{11} | = 0.$$

Ces coefficients sont aussi les solutions positives de l'équation matricielle:

$$| \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{12} - r_k^2 \mathbf{R}_{22} | = 0.$$

Quant aux vecteurs, \mathbf{b}_k et \mathbf{c}_k , contenant les coefficients canoniques, c'est-à-dire les coefficients qui interviennent dans les combinaisons linéaires constituant les variables canoniques, ils s'obtiennent en résolvant les équations suivantes:

$$(\mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21} - r_k^2 \mathbf{R}_{11}) \mathbf{b}_k = \mathbf{0}$$

et

$$(\mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{12} - r_k^2 \mathbf{R}_{22}) \mathbf{c}_k = \mathbf{0}.$$

En prémultipliant par \mathbf{R}_{11}^{-1} ou par \mathbf{R}_{22}^{-1} , on vérifie que les vecteurs \mathbf{b}_k et \mathbf{c}_k sont les vecteurs propres des matrices:

$$\mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21} \quad \text{et} \quad \mathbf{R}_{22}^{-1} \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{12},$$

et que les coefficients de corrélation canonique sont les valeurs propres de ces mêmes matrices.

3.2. Corrélation canonique, régression multiple et régression multiple multivariée

La corrélation canonique est directement liée à la régression multiple classique et à sa généralisation, la *régression multiple multivariée*⁶.

6. En anglais: *multivariate regression*.

Considérons d'abord le cas de la régression multiple classique. On dispose d'un premier vecteur de données, \mathbf{y} , de dimensions $n \times 1$ et d'un ensemble de p vecteurs de variables explicatives, $\mathbf{x}_1, \dots, \mathbf{x}_p$, constituant la matrice \mathbf{X} , de dimensions $n \times p$. L'objectif de la régression multiple est de trouver la combinaison linéaire des p variables:

$$\hat{\mathbf{y}} = \mathbf{x}_1 b_1 + \mathbf{x}_2 b_2 + \dots + \mathbf{x}_p b_p = \mathbf{X} \mathbf{b},$$

qui maximise la corrélation entre \mathbf{y} et $\hat{\mathbf{y}}$. Ce coefficient de corrélation, qui est précisément le coefficient de corrélation multiple classique, est aussi le coefficient de corrélation canonique entre la variable y et les p variables explicatives, et les coefficients de régression partielle sont directement liés aux coefficients canoniques des variables x_1, \dots, x_p .

En effet, si les variables sont centrées, et si on divise la matrice \mathbf{A} des sommes des carrés et des produits des écarts des $p + 1$ variables y, x_1, \dots, x_p , de la façon suivante:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

on a, pour la somme des carrés des écarts totale, pour la somme des carrés des écarts liée à la régression et pour la somme des carrés des écarts résiduelle:

$$SCE_t = \mathbf{y}' \mathbf{y} = a_{11},$$

$$SCE_{rég} = \mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21},$$

$$SCE_{rés} = \mathbf{y}' \mathbf{y} - \mathbf{y}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = a_{11} - \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21}.$$

Par conséquent, le coefficient de détermination multiple, qui est le rapport entre la somme des carrés des écarts liée à la régression et la somme des carrés des écarts totale est égal à:

$$r^2 = a_{11}^{-1} \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21}.$$

Si on remplace les variables de départ par des variables centrées et réduites, on ne modifie évidemment pas la valeur du coefficient de détermination multiple, qui est insensible à tout changement d'origine et d'unité. Si, de plus, on divise les sommes des carrés des écarts par n , on peut remplacer les sommes de carrés et de produits des écarts par les corrélations:

$$r^2 = r_{11}^{-1} \mathbf{r}_{12} \mathbf{R}_{22}^{-1} \mathbf{r}_{21},$$

$r_{11}, \mathbf{r}_{12}, \mathbf{r}_{21}$ et \mathbf{R}_{22} désignant les éléments de la matrice de corrélation des $p + 1$ variables y, x_1, \dots, x_p , après division de cette matrice de façon analogue à celle présentée au paragraphe 3.1. On notera également que l'élément r_{11} , qui est égal à l'unité, a été maintenu dans l'expression pour mieux faire apparaître l'analogie entre cette expression et la matrice:

$$\mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21},$$

utilisée pour le calcul des coefficients de corrélation canonique et des variables canoniques. Dans le cas où le premier groupe de variables est constitué d'une seule variable, cette dernière matrice est égale à l'expression précédente, et se ramène donc à un scalaire. La valeur propre de la matrice se confond donc avec l'élément de la matrice et la valeur r^2 est bien le carré du coefficient de corrélation canonique entre y et l'ensemble des variables x_1, \dots, x_p .

D'autre part, le rapport θ , entre la somme des carrés des écarts liée à la régression et la somme des carrés des écarts résiduelle est directement fonction du coefficient de corrélation canonique. En effet:

$$\theta = \frac{\mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21}}{a_{11} - \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21}} = \frac{r^2}{1 - r^2},$$

et θ est donc aussi la valeur propre de la matrice:

$$(a_{11} - \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21})^{-1} \mathbf{a}_{12} \mathbf{A}_{22}^{-1} \mathbf{a}_{21},$$

dont l'importance apparaîtra par la suite.

La régression multiple classique peut se généraliser au cas où on dispose de q variables à expliquer en fonction de p variables explicatives. Les données se présentent alors sous la forme d'une première matrice, \mathbf{Y} , de dimensions $n \times q$ et d'une seconde matrice, \mathbf{X} , de dimensions $n \times p$. Pour plus de simplicité, considérons que les $q+p$ variables sont centrées. L'objectif de la régression multivariée est de trouver les q combinaisons linéaires des p variables explicatives:

$$\mathbf{Y} = \mathbf{X}\mathbf{B},$$

\mathbf{B} étant une matrice de dimensions $p \times q$, qui rend minimum la somme des sommes des carrés des résidus relatives à chacune des variables à expliquer. Cette matrice est en fait constituée des q vecteurs des coefficients de régression obtenus en calculant q équations de régression multiples classiques indépendantes:

$$\mathbf{B} = (\mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_q).$$

La matrice des sommes des carrés et des produits $\mathbf{Y}'\mathbf{Y}$, de dimensions $q \times q$, peut se décomposer en une somme de deux matrices: la première matrice, \mathbf{H} , est liée à la régression et la seconde matrice, \mathbf{E} , est liée aux résidus. Si on subdivise la matrice des sommes des carrés et des produits des $q+p$ variables de la façon suivante:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

les matrices \mathbf{H} et \mathbf{E} sont respectivement égales à:

$$\mathbf{H} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

et
$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}.$$

En relation avec la décomposition de la matrice $\mathbf{Y}'\mathbf{Y}$, on peut montrer que les valeurs propres de la matrice:

$$(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21},$$

sont les carrés des coefficients de corrélation canonique, r_k^2 , et que les valeurs propres, θ_k , de la matrice $\mathbf{E}^{-1}\mathbf{H}$ sont égales à:

$$\theta_k = \frac{r_k^2}{1 - r_k^2}.$$

De même, si on effectue la régression de \mathbf{X} sur \mathbf{Y} , on obtient:

$$\mathbf{H} = \mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$$

et
$$\mathbf{E} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.$$

Les valeurs propres de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$ sont identiques aux carrés des coefficients de corrélation canonique r_k et les valeurs propres de la matrice $\mathbf{E}^{-1}\mathbf{H}$ sont égales aux valeurs propres θ_k obtenues ci-dessus.

Quant aux coefficients des variables canoniques successives, ce sont les éléments des vecteurs propres de la matrice:

$$\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21},$$

pour les variables y_1, \dots, y_q , et de la matrice:

$$\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12},$$

pour les variables x_1, \dots, x_p . Ces deux matrices sont précisément les matrices $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$, relatives, d'une part, à la régression de \mathbf{Y} sur \mathbf{X} et, d'autre part, de \mathbf{X} sur \mathbf{Y} . Enfin, ces vecteurs propres sont aussi les vecteurs propres des deux matrices $\mathbf{E}^{-1}\mathbf{H}$ relatives également à la régression de \mathbf{Y} sur \mathbf{X} et de \mathbf{X} sur \mathbf{Y} . Il faut noter, cependant, que les coefficients ainsi obtenus sont les coefficients qui permettent le calcul des valeurs des variables canoniques à partir des variables initiales non réduites. Pour obtenir les coefficients relatifs aux variables réduites, analogues à ceux donnés dans la figure 5, il faut multiplier les éléments des vecteurs propres obtenus ci-dessus par l'écart-type de la variable correspondante.

A titre d'illustration, reprenons l'exemple relatif à l'étude de la chicorée witloof, mais en nous limitant à deux caractéristiques relatives à la plante (le poids des feuilles et le poids des racines) et à une seule caractéristique relative au chicon (le poids du chicon).

La matrice des sommes des carrés et des produits d'écart pour les trois variables vaut:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21} & \mathbf{a}_{22} \end{pmatrix},$$

avec:

$$\mathbf{A}_{11} = \begin{pmatrix} 131,945 & 87,009 \\ 87,009 & 122,574 \end{pmatrix},$$

$$\mathbf{a}_{21} = \mathbf{a}'_{12} = (54,700 \quad 66,733)$$

et

$$a_{22} = 118,399.$$

La régression des variables du groupe "plante" sur la variable du groupe "chicon", conduit aux résultats suivants:

$$\mathbf{b} = (0,46200 \quad 0,56363),$$

$$(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H} = \begin{pmatrix} 0,048252 & 0,058866 \\ 0,217274 & 0,265070 \end{pmatrix}$$

et

$$\mathbf{E}^{-1}\mathbf{H} = \begin{pmatrix} 0,070269 & 0,085726 \\ 0,316413 & 0,386018 \end{pmatrix}.$$

Le vecteur \mathbf{b} contient les deux coefficients de régression simples obtenus lors du calcul de la droite de régression de PLPOIDSF en fonction de CHICPOID et de PLPOIDSR en fonction de CHICPOID.

La première valeur propre de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$ vaut 0,3133 et est égale au carré du coefficient du premier coefficient de corrélation canonique ($r_1 = 0,5597$). La seconde valeur propre est nulle, la matrice étant singulière.

Si on fixe arbitrairement à l'unité la première composante du vecteur propre associé à la première valeur propre de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$, on trouve, pour la seconde composante, la valeur 4,5020. Pour obtenir les coefficients canoniques relatifs aux variables standardisées, il faut multiplier chaque composante du vecteur propre par l'écart-type de la variable à laquelle elle est associée. La première composante doit donc être multipliée par l'écart-type de PLPOIDSF, soit 0,3632, et la seconde par l'écart-type de PLPOIDSR, soit 0,3501.

A une constante près, la variable canonique, u_1 , relative au groupe "plante" est donc égale à:

$$u_1 = 0,3632\text{PLPOIDSF}^* + 1,5762\text{PLPOIDSR}^*,$$

où les variables PLPOIDSF* et PLPOIDSR* correspondent aux variables centrées réduites relatives à PLPOIDSF et PLPOIDSR. Elle n'a cependant pas une variance unitaire. Compte tenu de la corrélation entre le poids des feuilles et le poids des racines, la variance de u_1 vaut, en effet:

$$0,3632^2 + 1,5762^2 + (2)(0,6842)(0,3632)(1,5762) = 3,3997.$$

Pour que u_1 soit d'écart-type unitaire, il faut diviser les coefficients de la variable canonique par l'écart-type de u_1 et on obtient:

$$u_1 = 0,1970\text{PLPOIDSF}^* + 0,8549\text{PLPOIDSR}^*,$$

ces deux derniers coefficients étant identiques aux coefficients obtenus par la procédure CANCELL du logiciel SAS.

Quant à la matrice $\mathbf{E}^{-1}\mathbf{H}$, on vérifie qu'elle est également singulière. La valeur propre est égale à:

$$\frac{0,3133}{1-0,3133} = 0,4562,$$

et le vecteur propre relatif à cette valeur propre est identique au vecteur propre de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$.

Si on exprime, par contre, le poids des chicons en fonction du poids des feuilles et du poids des racines, on obtient:

$$\mathbf{b} = \begin{pmatrix} 0,10444 \\ 0,47029 \end{pmatrix},$$

$$(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H} = 0,3133$$

et

$$\mathbf{E}^{-1}\mathbf{H} = 0,4563.$$

Le vecteur \mathbf{b} contient les deux coefficients de régression partielle. A une constante près, ce vecteur est bien égal au vecteur des coefficients de la variable canonique du groupe "plante", lorsque les variables ne sont pas standardisées. En effet, si on divise le vecteur \mathbf{b} par 0,10444, la première composante devient égale à l'unité et la seconde composante devient égale à 4,5029. Cette valeur diffère légèrement de la valeur obtenue précédemment (4,5020) à cause de l'imprécision des calculs. Quant aux matrices $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$ et $\mathbf{E}^{-1}\mathbf{H}$, elles sont de dimensions 1×1 et sont, par conséquent, égales à leur valeur propre. On vérifie aussi que la valeur propre de chacune des deux matrices est égale à la valeur propre de la matrice correspondante établie lors de la régression des variables du groupe "plante" sur la variable CHICPOID.

Enfin, le vecteur propre de ces matrices n'a qu'une seule composante, qui peut donc prendre n'importe quelle valeur. Toutefois, en donnant à cette composante la valeur unitaire et en considérant la variable centrée réduite CHICPOID*, on obtient la variable canonique v_1 de variance unitaire:

$$v_1 = \text{CHICPOID}^*.$$

3.3. Corrélation canonique, analyse de la variance multivariée et analyse factorielle discriminante

L'analyse de la variance multivariée⁷ est également directement liée à la corrélation canonique. En effet, les tests de signification classiques utilisés en analyse de la variance multivariée sont basés sur l'examen des valeurs propres de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$ (test de WILKS, test de PILLAI) ou de la matrice $\mathbf{E}^{-1}\mathbf{H}$ (test de HOTELLING et LAWLEY, test de ROY), la matrice \mathbf{H} étant la matrice

7. En anglais: *multivariate analysis of variance, MANOVA*.

factorielle et la matrice E la matrice résiduelle. Des informations concernant ces différents tests sont donnés par DAGNELIE [1982].

Le lien entre l'analyse de la variance et la corrélation canonique s'explique particulièrement bien si on considère que les q variables observées constituent un premier groupe de variables et que l'autre groupe de variables est formé par une série de p variables indicatrices. Ainsi supposons, par exemple, que les cinq variables décrivant les chicons aient été observées sur des chicons provenant de trois origines différentes et que le problème soit de vérifier s'il existe des différences significatives entre ces origines. Le problème pourrait être résolu en utilisant un programme de corrélation canonique. Pour cela, il suffirait en effet de générer, par exemple, trois variables indicatrices x_k ($k=1, 2$ ou 3), x_k étant une variable dont la valeur vaut 0 si l'individu n'appartient pas à la $k^{\text{ième}}$ origine et 1 si l'individu appartient à la $k^{\text{ième}}$ origine. Il serait alors possible de déterminer deux coefficients de corrélation canonique et par conséquent aussi deux couples de variables canoniques, puisque les trois variables indicatrices présentent une relation linéaire. Pour tout individu i , on a, en effet:

$$x_{i1} + x_{i2} + x_{i3} = 1.$$

Le test de signification du premier coefficient de corrélation canonique (paragraphe 3) serait, dans ces conditions, strictement équivalent au test de WILKS.

D'autre part, les deux variables canoniques relatives aux caractéristiques des chicons sont précisément les variables canoniques définies dans le cadre de l'*analyse factorielle discriminante*⁸, appelée aussi analyse des variables canoniques ou encore analyse canonique discriminante [DAGNELIE, 1982]. On notera simplement que les logiciels standardisent parfois différemment les variables canoniques dans le cas de la corrélation canonique et dans le cas de l'analyse discriminante: dans le premier cas, les variables canoniques sont de variance unitaire, alors que, dans le second cas, elles sont telles que la variance résiduelle (ou intra-groupes) est unitaire [X, 1985].

On notera également que, pour les problèmes de comparaisons de populations auxquels nous venons de faire allusion, les conditions d'application sont différentes de celles que nous avons énoncées au paragraphe 2.2. En particulier, on ne suppose pas la normalité à $q + p$ dimensions des variables des deux groupes, mais on suppose simplement la normalité à q dimensions des variables de départ.

3.4. Corrélation canonique et analyse factorielle des correspondances

Enfin, l'étude des corrélations et des variables canoniques est également étroitement liée à l'*analyse factorielles des correspondances*⁹. Considérons, en effet, un échantillon de n individus répartis en $p \times q$ classes, en fonction de deux critères. Ces données peuvent être présentées sous la forme d'un tableau de contingence à p lignes et q colonnes. Elles peuvent aussi être présentées sous

8. En anglais: *factorial discriminant analysis, canonical variate analysis*.

9. En anglais: *correspondence analysis*.

la forme d'une matrice de n lignes et de $(p - 1) + (q - 1)$ colonnes, ces colonnes correspondant à des variables indicatrices x_k ($k = 1, \dots, p - 1$) et y_l ($l = 1, \dots, q - 1$) telles que:

$x_{ik} = 1$ si l'individu i possède la modalité k pour le premier caractère et $x_{ik} = 0$ sinon;

$y_{il} = 1$ si l'individu i possède la modalité l pour le second caractère et $y_{il} = 0$ sinon.

Dans ces conditions, on peut montrer que l'analyse canonique effectuée sur les deux groupes de variables x et y donne des axes qui coïncident avec les facteurs de l'analyse des correspondances du tableau de contingence à p lignes et q colonnes [LEBART *et al.*, 1979].

4. CONCLUSIONS

Dans le paragraphe précédent, le rôle théorique central de l'analyse canonique a été mis en évidence, plusieurs méthodes statistiques multivariées pouvant être considérées comme des cas particuliers de l'analyse canonique.

D'un point de vue pratique, l'analyse canonique peut être utilisée dans un but descriptif et exploratoire [KSHIRSAGAR, 1972]. Elle résume les relations complexes et constitue une méthode de réduction de la dimension du problème qui, au départ, peut faire intervenir beaucoup de variables, en ne considérant que les quelques premiers couples de variables canoniques.

Vue sous l'angle essentiellement descriptif et de réduction de dimension, l'analyse canonique fait inévitablement penser également à l'analyse en composantes principales et certains auteurs proposent d'ailleurs des représentations graphiques tout à fait analogues à celles utilisées en analyse en composantes principales. Ainsi BOUROCHE et SAPORTA [1980] représentent l'ensemble des variables initiales dans les plans formés par les variables canoniques d'un des deux groupes, à la manière des cercles de corrélations de l'analyse en composantes principales, tandis que BERTIER et BOUROCHE [1975] établissent des diagrammes de dispersion des variables canoniques. On notera également que, comme en analyse des composantes, les variables canoniques, construites à partir de combinaisons linéaires des variables initiales, n'ont pas nécessairement une signification physique [KSHIRSAGAR, 1972].

Enfin, on peut encore signaler que l'analyse canonique peut être étendue au cas où les variables de départ se répartissent en plus de deux groupes. Des informations complémentaires, ainsi que des références bibliographiques spécifiques, sont données par HARRIS [1975], MORRISSON [1967] et PRESS [1972].

5. BIBLIOGRAPHIE

- BERTIER P. et BOUROCHE J.M. [1975]. *Analyse des données multidimensionnelles*. Paris, Presses Universitaires de France, 270 p.
- BOUROCHE J.M. et SAPORTA G. [1980]. *L'analyse des données*. Paris, Presses Universitaires de France, 127 p.
- CHATFIELD C. et COLLINS A. [1980]. *Introduction to multivariate analysis*. London, Chapman and Hall, 246 p.
- COOLEY W.W. et LOHNES P.R. [1971]. *Multivariate data analysis*. New York, Wiley, 364 p.
- DAGNELIE P. [1979-1980]. *Théorie et méthodes statistiques: applications agronomiques* (2 vol.). Gembloux, Presses Agronomiques, 378 + 463 p.
- DAGNELIE P. [1982]. *Analyse statistique à plusieurs variables*. Gembloux, Presses Agronomiques, 362 p.
- HARRIS R. [1975]. *A primer of multivariate statistics*. New York, Academic Press, 332 p.
- KSHIRSAGAR A. [1972]. *Multivariate analysis*. New York, Dekker, 534 p.
- LAWLEY D.N. [1959]. Test of significance in canonical analysis. *Biometrika* 46, 59-66.
- LEBART L., MORINEAU A. et FENELON J.P. [1979]. *Traitement des données statistiques - méthodes et programmes*. Paris, Dunod, 510 p.
- MORRISSON D. [1967]. *Multivariate statistical methods*. New York, McGraw Hill, 338 p.
- PRESS J. [1972]. *Applied multivariate analysis*. New York, Holt, Rinehart and Winston, 521 p.
- RAO C.R. [1952]. *Advanced statistical methods in biometric research*. New York, Wiley, 390 p.
- RONCHAIINE J. [1962]. *Contribution à l'étude de la croissance et du développement de Cichorium intybus L. en culture* (2 vol.). Gembloux, Institut Agronomique, 588 p.
- TATSUKOA M. [1971]. *Multivariate analysis: techniques for educational and psychological research*. New York, Wiley, 310 p.
- X [1985]. *SAS user's guide: statistics*. Cary, SAS Institute, 956 p.

Obs	PLPOIDSF	PLPOIDSR	PLDIAMRC	PLDIAMFC	PLHAUTGF	PLLARGPE	PLPUBESC	PLLOBES	PLGAUFRU	CHICLONG	CHICLARG	CHICPOID
1	4.26268	4.02535	3.36730	3.04452	5.73657	0	1	0	0	5.19296	3.52636	4.27667
2	4.33073	3.93183	3.33220	3.04452	6.06379	0	1	0	0	5.25227	3.66356	4.33073
3	4.66344	3.68888	3.33220	3.13549	6.01616	0	0	0	0	4.82831	3.21888	3.46574
4	4.68213	5.15906	3.71357	3.33220	6.38012	0	0	0	0	5.27300	3.52636	4.49981
5	4.69135	4.12713	3.46574	3.17805	6.29157	0	0	0	0	5.34711	3.33220	4.49981
6	4.70953	4.07754	3.46574	3.17805	6.04025	0	0	0	0	5.13580	3.29584	3.87120
7	4.70953	4.43082	3.49651	3.17805	6.25383	0	0	0	0	5.41610	3.63759	4.85203
8	4.71850	4.54329	3.61092	3.33220	6.27288	0	1	0	0	5.13580	3.63759	4.51086
9	4.74493	4.04305	3.33220	3.17805	5.99146	0	0	0	0	5.22575	3.46574	4.14313
10	4.74493	4.07754	3.29584	3.13549	5.91350	0	1	0	0	5.19296	3.49651	4.14313
11	4.74493	4.12713	3.49651	3.25810	6.10925	0	0	0	0	5.22036	3.52636	4.24850
12	4.76217	4.15888	3.43399	3.09104	6.21461	1	0	0	0	5.27300	3.66356	4.56435
13	4.77068	4.24850	3.40120	3.13549	6.10925	0	1	0	0	5.27300	3.55535	4.33073
14	4.79579	4.60517	3.55535	3.21888	6.41346	0	1	0	0	5.19296	3.52636	4.59512
15	4.80402	4.09434	3.40120	3.09104	6.17379	0	0	0	0	5.07517	3.52636	4.31749
16	4.80402	4.58497	3.68888	3.33220	6.15273	0	0	0	0	5.42495	4.02535	5.01728
17	4.81218	4.58497	3.63759	3.33220	6.08677	0	0	0	0	4.90527	3.49651	4.23411
18	4.82831	3.98898	3.40120	3.09104	6.23441	0	0	0	0	5.07517	3.46574	4.00733
19	4.82831	4.11087	3.33220	3.25810	6.34564	0	0	0	0	5.56834	3.46574	4.38203
20	4.82831	4.56435	3.46574	3.29584	6.25383	0	1	0	0	5.04343	3.52636	4.26268
21	4.83628	4.30407	3.25810	3.17805	6.29157	0	0	1	0	5.22036	3.29584	4.14313
22	4.84419	4.43082	3.52636	3.13549	5.96615	0	0	0	0	5.13580	3.55535	4.21951
23	4.85203	4.17439	3.36730	3.21888	6.23441	0	1	1	0	5.39363	3.40120	4.55388
24	4.85203	4.41884	3.52636	3.29584	6.17379	0	1	0	0	5.24702	3.61092	4.23411
25	4.85203	4.45435	3.52636	3.29584	6.15273	1	0	1	0	5.19296	3.46574	4.49981
26	4.85203	4.51086	3.61092	3.17805	6.06379	0	0	0	0	5.39363	3.52636	4.38203

Annexe. Variables de départ et variables canoniques, pour les premiers individus de l'échantillon.

Obs	CHICLAXE	CHICQUAL	PLANTE1	PLANTE2	PLANTE3	PLANTE4	PLANTE5	CHICON1	CHICON2	CHICON3	CHICON4	CHICON5
1	4.48864	0	-1.94973	-4.11400	0.49190	-1.25689	0.18513	-1.08055	0.51161	-1.76130	-0.28750	1.06142
2	3.25810	1	-2.30097	-2.69834	0.53945	-0.33762	0.17440	-0.73661	-0.91842	-0.11018	-1.42013	-1.69425
3	3.09104	0	-2.67851	-0.85883	0.93392	-1.48497	0.20322	-2.61754	-1.58854	-0.56785	0.48258	1.94445
4	3.29584	0	0.95194	-2.44997	0.55780	0.27115	-2.94388	-0.35323	0.13143	0.73825	0.33826	0.32963
5	3.63759	1	-1.64438	-0.65194	0.65702	-0.94228	-1.01253	-1.06965	-0.37658	-0.65597	1.44928	-1.91875
6	3.13549	1	-1.67534	-1.31496	0.61445	-1.71828	-0.64738	-2.25490	-1.44747	-0.62995	0.18402	-1.31718
7	3.93183	0	-0.96289	-1.52157	0.03318	-0.42793	-1.65934	0.30665	1.01145	0.14010	0.22823	0.01932
8	3.04452	0	-0.52427	-1.75469	1.49072	-0.07745	-0.07742	0.19941	-0.44057	1.09266	-0.01440	0.79940
9	3.13549	0	-1.90026	-1.77700	0.44156	-0.11561	-0.22685	-1.27209	-0.17680	0.59435	-0.31913	0.41967
10	3.87120	0	-1.97989	-2.76398	-0.13879	0.67420	0.86509	-1.33635	0.13648	-0.77762	-0.49950	0.82050
11	2.89037	0	-1.47901	-0.99720	1.29557	-1.35448	-0.42965	-0.82340	-0.31184	1.16628	-0.37487	0.36548
12	4.24850	0	-1.56837	-1.32934	-0.28360	-1.27583	-0.22936	-0.20663	0.68470	-0.88498	-0.56551	0.68223
13	4.02535	0	-1.55952	-2.16565	-0.39110	0.25428	0.11422	-0.94836	0.51279	-0.75422	-0.55541	0.57257
14	2.99573	1	-0.63733	-1.53414	-0.04968	0.54202	-0.78576	-0.03177	-1.20613	0.51058	0.83772	-1.51166
15	3.58352	1	-1.86747	-0.96562	-0.62744	-0.99756	-1.08344	-0.67752	-1.29791	-0.98107	0.18078	-0.85223
16	4.04305	0	-0.23569	-1.33486	1.18990	-1.94053	-1.38737	1.28926	1.13879	0.26748	-2.24453	0.04024
17	2.94444	0	-0.27888	-1.72394	1.09933	-1.38996	-1.21030	-0.24402	-1.27950	0.67178	0.86323	1.67168
18	2.94444	0	-2.11154	-0.39345	-0.48003	-1.22306	-0.90740	-1.27087	-0.78617	0.63122	-0.29538	0.94848
19	3.63759	0	-1.79899	-0.54625	1.00742	1.39755	-0.23372	-1.48186	1.23461	0.31691	-0.63363	-0.76804
20	2.89037	1	-0.70827	-2.16578	0.60347	1.70364	0.07772	-0.58805	-1.80088	0.19475	0.13421	-0.98113
21	2.89037	0	-1.50955	-1.57722	0.53527	3.00255	-1.20104	-1.45444	-0.33485	0.96862	1.04700	0.35463
22	3.89182	1	-0.91003	-2.09601	-0.89173	-1.83218	-1.57192	-1.10622	-0.95006	-1.56067	-0.62893	-1.00097
23	3.09104	0	-1.68087	-1.41086	1.22381	1.74660	0.36919	-0.60221	0.40504	1.25850	1.12441	-0.22953
24	3.17805	0	-0.92063	-1.73503	0.83579	0.16533	0.35128	-0.86392	-0.06637	0.69060	-1.19920	0.35811
25	3.87120	1	-0.58151	-1.76628	1.91167	-0.00596	-0.16366	-0.59638	-0.72942	-1.21068	0.90583	-1.20527
26	4.09434	0	-0.62755	-1.65877	-0.52041	-2.25092	-1.77700	-1.12827	0.94214	-0.69920	-0.55155	0.11138

Annexe (suite). Variables de départ et variables canoniques, pour les premiers individus de l'échantillon.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté des Sciences Agronomiques et du Centre de Recherches Agronomiques de Gembloux (Belgique).

Quelques titres récents:

DAGNELIE P., GOUMARI A., KINDERMANS M., LORENT F. et RAMLOT P. [1986]. Douze années de coopération avec l'Institut Hassan II, à Rabat (Maroc), dans le domaine de la statistique et de l'informatique. *Notes Stat. Inform.* (Gembloux) 86/4, 12 p.

CLAUSTRIAUX J.J. [1986]. Introduction au logiciel statistique Minitab. *Notes Stat. Inform.* (Gembloux) 86/5, 22 p.

LEDEINE J.M. [1986]. Présentation d'un logiciel de simulation sur ordinateur d'exercices d'échantillonnage. *Notes Stat. Inform.* (Gembloux) 86/6, 15 p.

RAMLOT P. [1986]. Programmation structurée en Cobol, en Fortran et en Basic: aspects méthodologiques. *Notes Stat. Inform.* (Gembloux) 86/7, 33 p.

RAMLOT P., TOUSSAINT A. et VANDEVANDEL J.P. [1986]. Programmation structurée en Cobol, en Fortran et en Basic: étude de cas. *Notes Stat. Inform.* (Gembloux) 86/8, 36 p.

PALM R. [1987]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 87/1, 25 p.

PALM R. [1987]. Etude des séries chronologiques par la méthode de BOX et JENKINS. *Notes Stat. Inform.* (Gembloux) 87/2, 40 p.

PALM R. [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.

PALM R. et DE BAST [1988]. Construction d'un modèle agrométéorologique pour la prévision des productions agricoles dans la Communauté Economique Européenne. *Notes Stat. Inform.* (Gembloux) 87/2, 14 p.

PALM R. [1989]. Quelques éléments de programmation linéaire. *Notes Stat. Inform.* (Gembloux) 89/1, 37 p.

DAGNELIE P. [1989]. Le choix d'une méthode d'analyse statistique et l'examen préliminaire des données. *Notes Stat. Inform.* (Gembloux) 89/2, 17 p.

Faculté universitaire des Sciences agronomiques
Avenue de la Faculté d'Agronomie 8
5030 GEMBLoux (Belgique)

D/1990/2371/1