

NOTES DE STATISTIQUE ET D'INFORMATIQUE

2000/1

L'ANALYSE DE LA VARIANCE MULTIVARIÉE
ET L'ANALYSE CANONIQUE DISCRIMINANTE :
PRINCIPES ET APPLICATIONS
EN ANALYSE DE LA VARIANCE

R. PALM

Faculté universitaire des
Sciences agronomiques

Centre de Recherches
agronomiques

GEMBLOUX

(Belgique)

L'ANALYSE DE LA VARIANCE MULTIVARIÉE ET L'ANALYSE CANONIQUE DISCRIMINANTE : PRINCIPES ET APPLICATIONS

R. PALM*

RÉSUMÉ

Cette note décrit les principes de l'analyse de la variance multivariée et de l'analyse canonique discriminante. Les méthodes sont illustrées par deux applications numériques.

SUMMARY

This note describes the principles of multivariate analysis of variance and of canonical discriminant analysis. The methods are illustrated by two examples.

1. INTRODUCTION

L'*analyse de la variance multivariée*¹ est une extension naturelle de l'analyse de la variance univariée au cas où plusieurs variables quantitatives ont été observées simultanément sur les mêmes objets (parcelles, individus, unités expérimentales). Par rapport à une série d'analyses univariées indépendantes, l'analyse multivariée prend en compte les corrélations qui existent très souvent entre les variables étudiées.

L'*analyse canonique discriminante*, appelée aussi *analyse factorielle discriminante*², constitue un complément logique de l'analyse de la variance multivariée, dans la mesure où elle a comme objectif de décrire les différences liées aux facteurs étudiés, du moins lorsque de telles différences existent. Cette analyse ne doit pas être confondue avec l'*analyse discriminante décisionnelle*³ dont l'objectif est de définir une règle d'affectation permettant de classer un individu donné

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

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

2. En anglais: *canonical discriminant analysis, descriptive discriminant analysis*.

3. En anglais: *predictive discriminant analysis*.

dans un groupe particulier, parmi plusieurs groupes préalablement définis. Une note technique a été consacrée à ce sujet [PALM, 1999] et nous verrons les liens qui peuvent exister, dans certaines situations, entre ces deux problèmes.

Nous examinerons d'abord le cas d'un seul facteur étudié. Nous présenterons l'analyse de la variance (paragraphe 2) et l'analyse canonique discriminante (paragraphe 3). Nous envisagerons alors le problème de la sélection et de la hiérarchisation des variables (paragraphe 4). Le paragraphe 5 sera ensuite consacré à une généralisation des analyses aux cas où plusieurs facteurs sont étudiés. Enfin, nous terminerons par quelques informations complémentaires (paragraphe 6).

Les diverses notions présentées seront illustrées par des exemples numériques traités par le logiciel SAS principalement. Différentes figures proviennent des documents générés par les procédures ANOVA, CANDISC, DISCRIM et STEPDISC de ce logiciel.

Des présentations de l'analyse de la variance multivariée et de l'analyse canonique discriminante peuvent être trouvées dans la plupart des livres généraux relatifs à l'analyse multivariée. L'ouvrage de HUBERTY [1994], consacré aux deux aspects de l'analyse discriminante évoqués ci-dessus (analyse canonique discriminante et analyse décisionnelle), peut être particulièrement recommandé, de même que la note de TOMASSONE [1988].

2. ANALYSE DE LA VARIANCE À UN CRITÈRE

2.1. Décomposition des sommes de carrés et de produits d'écart

Le principe de l'analyse de la variance univariée à un critère est de diviser la somme des carrés des écarts totale en deux parties qui sont la somme des carrés des écarts factorielle et la somme des carrés des écarts résiduelle :

$$SCE_t = SCE_f + SCE_e .$$

Pour tester l'hypothèse d'égalité des moyennes des g groupes :

$$H_0 : m_1 = m_2 = \dots = m_g ,$$

on détermine le rapport suivant :

$$F_{obs} = \frac{SCE_f / (g - 1)}{SCE_e / (n - g)} ,$$

qui, à une constante près, est le rapport de la somme des carrés des écarts factorielle à la somme des carrés des écarts résiduelle. Dans cette relation, g est le nombre de moyennes comparées, et n est la somme des effectifs des g échantillons :

$$n = \sum_{i=1}^g n_i .$$

Un autre critère, appelé rapport de corrélation ou coefficient de corrélation non linéaire et souvent désigné par R^2 , est défini de la manière suivante [DAGNELIE 1998]:

$$R^2 = \text{SCE}_f / (\text{SCE}_f + \text{SCE}_e) .$$

Ce paramètre donne la proportion de la somme des carrés des écarts totale qui est attribuée au groupement en g classes. Il est directement lié à F_{obs} , puisque:

$$R^2 = \frac{(g-1)F_{obs}}{(g-1)F_{obs} + (n-g)} ,$$

et il pourrait être utilisé pour tester H_0 , au même titre que F_{obs} . Sous l'hypothèse nulle, sa distribution est liée aux distributions bêta [HUBERTY, 1994].

Le principe de l'analyse univariée peut être étendu au cas de p variables. L'équation fondamentale de l'analyse de la variance s'écrit alors :

$$\mathbf{T} = \mathbf{H} + \mathbf{E} ,$$

\mathbf{T} , \mathbf{H} et \mathbf{E} étant les matrices de sommes de carrés d'écarts et de sommes de produits d'écarts. \mathbf{T} est la matrice totale, \mathbf{H} est la matrice factorielle et \mathbf{E} est la matrice résiduelle. En particulier, l'élément jj de chacune des matrices correspond à la somme des carrés des écarts totale, à la somme des carrés des écarts factorielle et à la somme des carrés des écarts résiduelle de l'analyse de la variance univariée, relative à la variable j ($j = 1, \dots, p$).

Différents critères permettant de tester l'égalité des vecteurs de moyennes sont définis à partir des matrices \mathbf{H} et \mathbf{E} , comme nous le verrons au paragraphe suivant.

Afin d'illustrer les principes qui viennent d'être présentés, nous reprenons l'exemple initialement étudié par FISHER [1936], concernant des observations relatives à trois espèces d'iris (*Iris setosa*, *Iris versicolor* et *Iris virginica*). Cet exemple avait déjà été retenu pour illustrer, dans une note antérieure, les principes de l'analyse discriminante décisionnelle [PALM, 1999]. Les données complètes se trouvent, notamment, dans HAND *et al.* [1994], KENDALL *et al.* [1983] et dans la documentation SAS [SAS, 1989]. Quatre variables ont été observées sur 50 fleurs de chacune des espèces :

- PLONG : longueur des pétales (en cm),
- PLARG : largeur des pétales (en cm),
- SLONG : longueur des sépales (en cm),
- SLARG : largeur des sépales (en cm).

Le tableau 1 donne la moyenne et l'écart-type des variables pour chaque espèce. Les données de départ sont présentées avec une décimale. Toutefois, pour assurer une meilleure concordance des valeurs numériques obtenues manuellement avec celles obtenues par les logiciels, nous avons volontairement laissé, dans

Tableau 1. Moyennes et, entre parenthèses, écarts-types des quatre variables pour les trois espèces d'iris.

Variabes	<i>setosa</i>	<i>versicolor</i>	<i>viginica</i>
PLONG	1,462 (0,174)	4,260 (0,470)	5,552 (0,552)
PLARG	0,246 (0,105)	1,326 (0,198)	2,026 (0,275)
SLONG	5,006 (0,352)	5,936 (0,516)	6,588 (0,634)
SLARG	3,428 (0,379)	2,770 (0,314)	2,974 (0,322)

le tableau 1, des décimales non significatives. La figure 1 donne les tableaux des analyses de la variance univariées. On constate que les probabilités associées aux valeurs F_{obs} sont toutes très faibles, ce qui nous permet de conclure à l'existence de différences importantes entre les espèces pour chacune des variables.

D'une manière générale, on se rappellera qu'avant de tirer les conclusions, il faut vérifier les conditions d'application de l'analyse de la variance: les échantillons doivent être aléatoires, simples et indépendants et les populations doivent être normales et de même variance. Pour l'exemple considéré, on peut supposer qu'effectivement les échantillons ont été prélevés de manière aléatoire et simple et indépendamment d'une population à l'autre. Par contre, le caractère normal des populations n'est pas vérifié, pour toutes les combinaisons variables-espèces, sur la base du test de SHAPIRO et WILK proposé par SAS [SAS, 1989]. Toutefois, on peut considérer que cette condition d'application est assez secondaire dans cet exemple, compte tenu de l'effectif élevé des échantillons. De même, l'hypothèse d'égalité des variances est rejetée sur la base du test de BARTLETT pour trois variables (PLONG, PLARG, SLONG). Mais, ici aussi, l'analyse de la variance est robuste vis-à-vis de l'hétéroscédasticité, car les effectifs sont constants.

Enfin, on notera aussi que les différences de moyennes, du moins pour les caractéristiques des pétales, sont à ce point importantes qu'un test statistique est pratiquement superflu.

La figure 2 donne les matrices factorielle et résiduelle des sommes des carrés et des produits des écarts. On peut vérifier que les éléments diagonaux de ces matrices correspondent bien aux sommes des carrés des écarts factorielles et résiduelles des tableaux d'analyse de la variance repris dans la figure 1.

2.2. Différents tests d'égalité des moyennes

Pour tester de façon rigoureuse l'égalité de plusieurs vecteurs de moyennes, il faut que certaines conditions soient remplies. Ces conditions sont des extensions naturelles des conditions d'application de l'analyse de la variance univariée: il faut que les échantillons soient aléatoires, simples et indépendants et il faut que les g populations aient des distributions multinormales à p dimensions, de même matrice de variances et covariances:

$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma.$$

General Linear Models Procedure

Dependent Variable: PLONG LONGUEUR DES PETALES (CM)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	437.10280000	218.55140000	1180.16	0.0001
Error	147	27.22260000	0.18518776		
Corrected Total	149	464.32540000			

Dependent Variable: PLARG LARGEUR DES PETALES (CM)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	80.41333333	40.20666667	960.01	0.0001
Error	147	6.15660000	0.04188163		
Corrected Total	149	86.56993333			

Dependent Variable: SLONG LONGUEUR DES SEPALES (CM)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	63.21213333	31.60606667	119.26	0.0001
Error	147	38.95620000	0.26500816		
Corrected Total	149	102.16833333			

Dependent Variable: SLARG LARGEUR DES SEPALES (CM)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.34493333	5.67246667	49.16	0.0001
Error	147	16.96200000	0.11538776		
Corrected Total	149	28.30693333			

Figure 1. Analyses de la variance univariées.

E = Error SS&CP Matrix

	PLONG	PLARG	SLONG	SLARG
PLONG	27.2226	6.2718	24.6246	8.1208
PLARG	6.2718	6.1566	5.645	4.8084
SLONG	24.6246	5.645	38.9562	13.63
SLARG	8.1208	4.8084	13.63	16.962

H = Type III SS&CP Matrix for ESPECE

	PLONG	PLARG	SLONG	SLARG
PLONG	437.1028	186.774	165.2484	-57.2396
PLARG	186.774	80.41333333	71.27933333	-22.93266667
SLONG	165.2484	71.27933333	63.21213333	-19.95266667
SLARG	-57.2396	-22.93266667	-19.95266667	11.34493333

Figure 2. Matrice résiduelle et matrice factorielle des sommes de carrés et de produits d'écart.

L'importance de ces conditions et la manière de les vérifier seront discutées au paragraphe 6.1.

Lorsque ces conditions d'application sont remplies, différents critères ont été proposés pour tester l'hypothèse nulle d'égalité des vecteurs de moyennes :

$$H_0 : \mathbf{m}_1 = \mathbf{m}_2 = \dots = \mathbf{m}_g .$$

Ces critères font tous intervenir, d'une façon ou d'une autre, les matrices \mathbf{H} et \mathbf{E} , qui ont été définies au paragraphe précédent. Avant d'examiner ces critères et afin de comprendre les liens entre ceux-ci, nous précisons tout d'abord quelques propriétés des matrices \mathbf{H} et \mathbf{E} . Il s'agit de matrices carrées symétriques, de dimensions $p \times p$, p étant le nombre de variables prises en considération. Il en résulte que :

$$\mathbf{E}^{-1} \mathbf{H} = \mathbf{H} \mathbf{E}^{-1} \quad \text{et} \quad \mathbf{H}(\mathbf{H} + \mathbf{E})^{-1} = (\mathbf{H} + \mathbf{E})^{-1} \mathbf{H} .$$

La matrice $\mathbf{E}^{-1} \mathbf{H}$, et donc aussi la matrice $\mathbf{H} \mathbf{E}^{-1}$, possède s valeurs propres positives, notées l_1, l_2, \dots, l_s , avec :

$$s \leq \min (p, g - 1) ,$$

l'égalité étant vérifiée notamment lorsqu'il n'y a pas de colinéarité exacte entre les variables, ce qui est le cas le plus fréquent.

La matrice $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$, et donc aussi la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$, possède également s valeurs propres positives, notées l'_1, l'_2, \dots, l'_s , liées aux valeurs propres de $\mathbf{E}^{-1}\mathbf{H}$ par la relation :

$$l'_i = l_i / (1 + l_i).$$

Les différents critères que nous allons examiner sont fonction des valeurs propres l_i ou l'_i .

Le critère le plus ancien, probablement aussi le plus utilisé, est *le rapport de vraisemblance de WILKS*⁴ :

$$\Lambda_{obs} = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|},$$

qui peut aussi s'écrire :

$$\Lambda_{obs} = \frac{1}{|\mathbf{E}^{-1}\mathbf{H} + \mathbf{I}|} = \prod_{i=1}^s (1/(l_i + 1)) = \prod_{i=1}^s (1 - l'_i),$$

car un déterminant est égal au produit de ses valeurs propres et les valeurs propres de $(\mathbf{E}^{-1}\mathbf{H} + \mathbf{I})$ sont égales à $l_i + 1$. Cette expression justifie l'appellation *test du produit des valeurs propres*, qui est également donnée au test de WILKS.

La première expression de Λ_{obs} montre qu'il s'agit du rapport du déterminant de la matrice résiduelle et du déterminant de la matrice totale. Ce critère est lié à la généralisation du critère R^2 défini dans le cas univarié. En effet, pour une seule variable, on a :

$$\Lambda_{obs} = \text{SCE}_e / (\text{SCE}_f + \text{SCE}_e) = 1 - R^2.$$

On constate que ce critère Λ_{obs} est d'autant plus petit que les différences entre les moyennes des groupes sont importantes, puisque dans ce cas, $|\mathbf{E}|$ est petit par rapport à $|\mathbf{H} + \mathbf{E}|$ ou, dans le cas univarié, la somme des carrés des écarts résiduelle est petite par rapport à la somme des carrés des écarts totale.

Lorsque les conditions d'application sont vérifiées et si l'hypothèse nulle est vraie, Λ_{obs} est une valeur observée d'une variable Λ de WILKS, de paramètres p , $k_1 = g - 1$ et $k_2 = n - g$. On rejettera donc l'hypothèse nulle lorsque $\Lambda_{obs} < \Lambda_\alpha$, α étant le niveau de signification du test.

Les variables de WILKS jouent, dans le cas multivarié, un rôle comparable aux variables F de FISHER-SNEDECOR. Elles sont liées, de façon exacte, aux variables F lorsque $p = 1$ ou $p = 2$ (une ou deux variables prises en considération) ou lorsque $k = 1$ ou $k = 2$ (deux ou trois groupes). Dans ces cas, on peut donc transformer la valeur Λ_{obs} en une valeur F_{obs} et comparer cette valeur F_{obs} à $F_{1-\alpha}$. Les formules de passage sont données, notamment, par DAGNELIE [1975].

Lorsqu'on ne se trouve pas dans les différents cas évoqués ci-dessus ($p = 1$ ou $p = 2$ ou $k = 1$ ou $k = 2$), les probabilités relatives aux variables de WILKS peuvent être calculées, de façon approchée, à l'aide des distributions

4. En anglais : *WILKS (lambda) criterion, likelihood ratio test.*

χ^2 [DAGNELIE 1975; HUBERTY 1994]. Une autre approximation, basée sur les distributions F , a été proposée par RAO [1952]. Cette dernière approximation est celle retenue dans les logiciels Minitab et SAS, par exemple.

En pratique donc, le test de WILKS conduit à une valeur Λ_{obs} , qui est transformée en une valeur F_{obs} , par une relation exacte ou approximative, et cette dernière valeur est comparée à la valeur $F_{1-\alpha}$, ou, ce qui est équivalent, la probabilité :

$$prob = P(F > F_{obs}),$$

est comparée au niveau de signification α . On rejette l'hypothèse d'égalité des vecteurs de moyennes si cette probabilité est inférieure à α .

Le test de PILLAI, ou de BARTLETT-PILLAI, utilise comme critère la trace de la matrice $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ ou de la matrice $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$, c'est-à-dire aussi la somme des valeurs propres de cette matrice. Une transformation de cette trace possède approximativement une distribution F , lorsque l'hypothèse nulle est vraie.

Le test de HOTELLING-LAWLEY est basé sur la somme des valeurs propres de $\mathbf{E}^{-1}\mathbf{H}$ ou de $\mathbf{H}\mathbf{E}^{-1}$, dont la distribution, lorsque l'hypothèse nulle est vraie, est aussi liée, de manière approchée, à une distribution F .

Le critère de ROY est la plus grande valeur propre, l'_1 , de $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ ou, ce qui revient au même, la plus grande valeur propre, l_1 , de $\mathbf{E}^{-1}\mathbf{H}$, puisque ces deux valeurs sont liées par une relation exacte :

$$l'_1 = l_1 / (1 + l_1).$$

Des tables spéciales sont nécessaires pour l'utilisation de ce critère. Des informations à ce sujet sont données par DAGNELIE [1975]. Minitab et SAS utilisent la valeur propre l_1 et SAS donne une valeur F_{obs} , qui est une transformation de l_1 .

Le test de ROY, qui ne prend en compte que la première valeur propre, est analogue à une analyse de la variance univariée, réalisée sur la première variable canonique [HUBERTY, 1994]. La définition et l'étude des variables canoniques feront l'objet du paragraphe suivant. A cette occasion, on comprendra que l'utilisation du test de ROY ne se justifie que si la première valeur propre de $\mathbf{E}^{-1}\mathbf{H}$ est largement supérieure à toutes les autres valeurs propres.

L'utilisateur dispose donc de plusieurs critères permettant de tester l'hypothèse nulle et ces critères sont transformés en une valeur F_{obs} . Les formules de conversion des critères vers les variables F sont données notamment par HUBERTY [1994], SAS [1989] et X [1994]. Sauf dans le cas particulier où $s = 1$, c'est-à-dire où $p = 1$ et/ou $g = 2$, les variables F prises en considération n'ont pas des degrés de liberté identiques et les probabilités associées à ces tests sont différentes.

Aucun de ces tests ne peut être considéré comme uniformément le plus puissant et aucun test ne peut être recommandé, de manière systématique, de préférence aux autres [DAGNELIE, 1975]. Rappelons simplement que le test de WILKS est le plus populaire [HUBERTY, 1994].

Manova Test Criteria and F Approximations for the Hypothesis
of no Overall ESPECE Effect

H = Type III SS&CP Matrix for ESPECE E = Error SS&CP Matrix

S=2 M=0.5 N=71

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.0234386	199.145	8	288	0.0001
Pillai's Trace	1.1918988	53.466	8	290	0.0001
Hotelling-Lawley Trace	32.4773202	580.532	8	286	0.0001
Roy's Greatest Root	32.1919292	1166.957	4	145	0.0001

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

NOTE: F Statistic for Wilks' Lambda is exact.

Figure 3. Critères multivariés utilisés pour les tests d'égalité des moyennes.

Pour l'exemple des trois espèces d'iris caractérisés par quatre variables, la matrice $\mathbf{E}^{-1} \mathbf{H}$ possède deux valeurs propres non nulles :

$$l_1 = 32,1919 \quad \text{et} \quad l_2 = 0,2854,$$

et, par conséquent, les valeurs propres de $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ sont égales à :

$$l'_1 = \frac{32,1919}{33,1919} = 0,96987 \quad \text{et} \quad l'_2 = \frac{0,2854}{1,2854} = 0,22203.$$

A partir de ces valeurs propres, on peut retrouver, aux erreurs d'arrondis près, les différents critères définis ci-dessus et repris dans la figure 3 :

$$\Lambda_{obs} = (1 - 0,96987)(1 - 0,22203) = 0,02344 \quad (\text{lambda de WILKS}),$$

$$tr \mathbf{H}(\mathbf{H} + \mathbf{E})^{-1} = 0,96987 + 0,22203 = 1,1919 \quad (\text{trace de BARTLETT-PILLAI}),$$

$$tr \mathbf{E}^{-1} \mathbf{H} = 32,1919 + 0,2854 = 32,4773 \quad (\text{trace de HOTELLING-LAWLEY}),$$

$$l_1 = 32,1919 \quad (\text{plus grande valeur propre de ROY}).$$

La figure 3 donne également les valeurs F_{obs} avec les degrés de liberté correspondants et la probabilité associée, pour chacun des quatre tests. Pour le test de WILKS, la transformation de Λ_{obs} en F_{obs} est une transformation exacte, car on se trouve dans le cas de la comparaison de trois populations ($k_1 = 2$). Les quantités S, M et N reprises dans la figure 3 interviennent dans le calcul des valeurs F_{obs} pour les différents tests. Leur définition est donnée dans SAS [1989].

Pour cet exemple, les quatre tests conduisent indiscutablement au rejet de l'hypothèse nulle d'égalité des vecteurs de moyennes.

On notera que, comme dans le cas des analyses de la variance univariées, les conditions d'application de l'analyse de la variance multivariée ne sont pas remplies. Le caractère non normal de certaines distributions marginales signalé au paragraphe 2.1 exclut le caractère normal à quatre dimensions. De même, l'inégalité des variances des variables prises individuellement (paragraphe 2.1) exclut l'égalité des matrices de variances et covariances. Le test multivarié d'égalité des matrices de variances et covariances, qui sera mentionné au paragraphe 6.1, conduit d'ailleurs au rejet de l'hypothèse nulle. Le non-respect des conditions d'application nous paraît cependant d'une importance secondaire, d'une part, en raison de la taille élevée et identique des échantillons et, d'autre part, en raison de l'importance des différences observées entre les moyennes des échantillons.

2.3. Importance du facteur étudié

Pour l'analyse de la variance univariée, nous avons déjà signalé que le rapport :

$$R^2 = \text{SCE}_f / (\text{SCE}_f + \text{SCE}_e),$$

est une mesure de l'intensité de la liaison qui existe entre la variable qui fait l'objet de l'étude et la variable qui définit le groupement en classes (paragraphe 2.1).

Nous avons vu également que le critère Λ_{obs} de WILKS est lié à une généralisation au cas multivarié du coefficient R^2 (paragraphe 2.2) :

$$R^2 = 1 - \Lambda_{obs}.$$

D'autres coefficients, appelés *indices d'association*⁵, sont proposés dans la littérature pour quantifier l'importance du facteur étudié sur les variables prises en considération. Ainsi, en relation avec le critère de WILKS, on définit le critère τ^2 :

$$\tau^2 = 1 - \Lambda_{obs}^{1/s},$$

s étant le minimum de p et $g - 1$. Le coefficient ζ^2 est dérivé du critère de HOTELLING-LAWLEY :

$$\zeta^2 = \frac{\text{tr} \mathbf{E}^{-1} \mathbf{H}}{s + \text{tr} \mathbf{E}^{-1} \mathbf{H}},$$

tandis que le coefficient ξ^2 est dérivé du critère de BARTLETT-PILLAI :

$$\xi^2 = \frac{\text{tr} \mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}}{s}.$$

Lorsque $s = 1$, c'est-à-dire lorsque $p = 1$ ou $g = 2$, les quatre indices d'association sont égaux. Par contre, dans les autres situations, ils ne le sont pas.

Comme pour le coefficient de détermination multiple en régression, on peut ajuster les coefficients d'association, de manière à obtenir des estimateurs moins biaisés. Pour R^2 , on a par exemple [HUBERTY, 1994] :

$$R_a^2 = R^2 - \frac{p^2 + (g - 1)^2}{n} (1 - R^2),$$

5. En anglais : *indices of association, effect-size indices.*

et, en remplaçant successivement R^2 par les autres indices, on obtient les valeurs ajustées de ces autres indices.

En pratique, le choix d'un indice d'association dépend, dans une certaine mesure, du choix du critère utilisé pour tester l'égalité des vecteurs des moyennes. L'utilisateur qui donne la préférence au test de PILLAI mesurera l'association par le paramètre ξ^2 ; celui qui donne la préférence au test de HOTELLING-LAWLEY mesurera l'association par le paramètre ζ^2 . Par contre, un certain choix subsiste pour l'utilisateur du test de WILKS (R^2 ou τ^2).

A titre d'illustration, les valeurs des quatre indices d'association sont les suivantes, pour l'exemple des iris :

$$\begin{aligned} R^2 &= 1 - 0,0234 = 0,9766, \\ \tau^2 &= 1 - 0,0234^{1/2} = 0,8470, \\ \zeta^2 &= 32,4773/(2 + 32,4773) = 0,9420, \\ \xi^2 &= 1,1919/2 = 0,5960. \end{aligned}$$

2.4. Comparaison de moyennes deux à deux et autres contrastes

Les comparaisons de moyennes deux à deux peuvent être faites par l'intermédiaire des carrés des distances de MAHALANOBIS :

$$\hat{\Delta}_{hl}^2 = (\bar{x}_h - \bar{x}_l)' \hat{\Sigma}^{-1} (\bar{x}_h - \bar{x}_l),$$

avec :

$$\hat{\Sigma} = \frac{1}{(n-g)} \mathbf{E}.$$

$\hat{\Sigma}$ est la matrice des variances et covariances résiduelle estimée et n est le nombre total d'observations. La distance de MAHALANOBIS est une mesure de la distance entre les moyennes des deux échantillons, qui tient compte à la fois des dispersions et des corrélations des différentes variables. Une justification de la formule de calcul de cette distance est donnée, notamment, dans PALM [1999].

Si l'hypothèse nulle :

$$H_0 : \mathbf{m}_h = \mathbf{m}_l,$$

est vraie, et si les conditions d'application de l'analyse de la variance sont remplies, la transformation :

$$F_{obs} = \frac{n_h n_l (n - g - p + 1)}{p(n_h + n_l)(n - g)} \hat{\Delta}_{hl}^2,$$

est distribuée selon une variable F à p et $n - g - p + 1$ degrés de liberté. On peut, par conséquent, déterminer la probabilité associée à cette valeur F_{obs} .

Indépendamment des tests proprement dits, le calcul des distances généralisées de MAHALANOBIS permet de se faire une idée de l'éloignement des différents centres de gravité des groupes dans l'espace à p dimensions.

Il faut noter que le test d'égalité de deux moyennes peut aussi être réalisé, de manière tout à fait équivalente, par l'intermédiaire des tests classiques de

l'analyse de la variance multivariée, à partir des matrices \mathbf{E} et \mathbf{H} , la matrice \mathbf{E} étant la matrice résiduelle basée sur les g groupes et la matrice \mathbf{H} étant la matrice factorielle basée sur les groupes h et l uniquement.

D'autres contrastes, plus complexes, peuvent aussi être testés. Leur expression générale est la suivante :

$$\boldsymbol{\theta} = \sum_{i=1}^g \gamma_i \mathbf{m}_i = \mathbf{0},$$

les γ_i étant des constantes telles que :

$$\sum_{i=1}^g \gamma_i = 0.$$

Les contrastes sont testés par les tests classiques de l'analyse de la variance, la matrice \mathbf{H} étant la matrice des sommes des carrés des écarts et des sommes des produits des écarts associée au contraste.

Lors de la réalisation de comparaisons multiples de moyennes, il ne faut pas perdre de vue que le risque global de première espèce (rejet d'une hypothèse nulle vraie) peut être largement supérieur au risque nominal fixé à α . Il peut, dès lors, se justifier d'utiliser, pour chacune des c comparaisons réalisées, un risque de première espèce α' égal à :

$$\alpha' = 1 - (1 - \alpha)^{1/c} \simeq \alpha/c,$$

de manière à maintenir le risque global au niveau α .

Quant aux coefficients d'association R^2 , τ^2 , ζ^2 et ξ^2 , ils peuvent être calculés à partir des différents critères utilisés pour les tests de signification des contrastes en question. On a, par exemple :

$$R^2 = 1 - \Lambda_{obs},$$

Λ_{obs} étant la variable de WILKS associée au contraste testé.

A titre d'illustration, la figure 4 donne les carrés des distances de MAHALANOBIS, les valeurs F_{obs} correspondantes et les probabilités associées. Toutes ces probabilités sont très faibles (inférieures ou égales à 0,0001), et on peut conclure que chaque espèce a un vecteur de moyennes différent de celui des deux autres espèces. Les carrés des distances montrent aussi que l'espèce *setosa* est la plus différente et que les espèces *versicolor* et *virginica* sont les moins différentes.

Bien que cela ne présente sans doute aucun intérêt pratique dans le cadre de ces données, nous allons illustrer la notion de contraste en considérant l'hypothèse nulle :

$$H_0 : \boldsymbol{\theta} = 2 \mathbf{m}_1 - \mathbf{m}_2 - \mathbf{m}_3 = \mathbf{0},$$

qui consiste à vérifier si l'espèce *setosa* est différente des deux autres espèces.

Canonical Discriminant Analysis
 Pairwise Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \text{COV}^{-1} (\bar{X}_i - \bar{X}_j)$$

Squared Distance to ESPECE

From ESPECE	SETOSA	VERSIC	VIRGIN
SETOSA	0	89.86419	179.38471
VERSIC	89.86419	0	17.20107
VIRGIN	179.38471	17.20107	0

F Statistics, NDF=4, DDF=144 for
 Squared Distance to ESPECE

From ESPECE	SETOSA	VERSIC	VIRGIN
SETOSA	0	550.18889	1098
VERSIC	550.18889	0	105.31265
VIRGIN	1098	105.31265	0

Prob > Mahalanobis Distance for
 Squared Distance to ESPECE

From ESPECE	SETOSA	VERSIC	VIRGIN
SETOSA	1.0000	0.0001	0.0001
VERSIC	0.0001	1.0000	0.0001
VIRGIN	0.0001	0.0001	1.0000

Figure 4. Carrés des distances de MAHALONOBIS entre les trois espèces, valeurs F_{obs} et probabilités correspondantes.

H = Contrast SS&CP Matrix for setosa / versi+virgi

	PLONG	PLARG	SLONG	SLARG
PLONG	395.3712	164.164	144.1888	-63.8288
PLARG	164.164	68.163333333	59.869333333	-26.50266667
SLONG	144.1888	59.869333333	52.584533333	-23.27786667
SLARG	-63.8288	-26.50266667	-23.27786667	10.304533333

Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall setosa / versi+virgi Effect

H=Contrast SS&CP Matrix for setosa/versi+virgi

E=Error SS&CP Matrix

	S=1	M=1	N=71			
Statistic	Value	F	Num DF	Den DF	Pr > F	
Wilks' Lambda	0.0327311	1063.871	4	144	0.0001	
Pillai's Trace	0.9672689	1063.871	4	144	0.0001	
Hotelling-Lawley Trace	29.5519688	1063.871	4	144	0.0001	
Roy's Greatest Root	29.5519688	1063.871	4	144	0.0001	

Figure 5. Matrice factorielle \mathbf{H} et critères pour le test du contraste $\boldsymbol{\theta} = 2\mathbf{m}_1 - \mathbf{m}_2 - \mathbf{m}_3 = \mathbf{0}$.

La figure 5 donne la matrice \mathbf{H} relative à ce contraste. Les éléments de la diagonale de cette matrice s'obtiennent comme dans le cas univarié :

$$\text{SCE}_{\boldsymbol{\theta}} = \left[\sum_{i=1}^g \gamma_i n_i \bar{x}_i \right]^2 / \sum_{i=1}^g n_i \gamma_i^2,$$

\bar{x}_i étant la moyenne de l'échantillon i , relative à la variable en question et n_i étant l'effectif de l'échantillon i .

Pour la longueur des pétales, par exemple, on a, en reprenant les moyennes données dans le tableau 1 :

$$\text{SCE}_{\boldsymbol{\theta}} = 50[(2)(1,462) - 4,260 - 5,552]^2 / (2^2 + 1 + 1) = 395,3712.$$

Les éléments hors diagonale s'obtiennent d'une manière similaire :

$$\text{SPE}_{\boldsymbol{\theta}} = \left[\sum_{i=1}^g \gamma_i n_i \bar{x}_i \right] \left[\sum_{i=1}^g \gamma_i n_i \bar{x}'_i \right] / \sum_{i=1}^g n_i \gamma_i^2,$$

\bar{x}_i et \bar{x}'_i désignant les moyennes de l'échantillon i relatives aux deux variables considérées.

Ainsi pour PLONG et PLARG, par exemple, on a :

$$\begin{aligned} \text{SPE}_\theta &= 50[(2)(1,462) - 4,260 - 5,552][(2)(0,246) - 1,326 - 2,026]/(2^2 + 1 + 1) \\ &= 164,164. \end{aligned}$$

A partir de la matrice \mathbf{H} de la figure 5 et de la matrice \mathbf{E} de la figure 2, on peut calculer la matrice $\mathbf{E}^{-1} \mathbf{H}$, dont l'unique valeur propre non nulle vaut 2,9254. On peut alors retrouver les critères donnés dans la figure 5, qui conduisent tous à la même valeur F_{obs} . On rejette évidemment l'hypothèse de nullité du contraste.

3. ANALYSE CANONIQUE DISCRIMINANTE

3.1. Calcul des variables canoniques

D'une manière générale, lorsqu'on dispose de deux ensembles de variables, on peut déterminer, pour chacun des deux ensembles, une combinaison linéaire des variables de l'ensemble, de telle sorte que le coefficient de corrélation de ces deux combinaisons linéaires soit maximum. Ces deux combinaisons linéaires constituent le premier couple de variables canoniques et le coefficient de corrélation qui a été maximisé est le premier coefficient de corrélation canonique.

On détermine ensuite, pour chaque ensemble, une deuxième variable canonique, non corrélée à la première, en calculant une autre combinaison linéaire des variables initiales. Ces combinaisons linéaires sont calculées de manière à assurer également le maximum de la valeur 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 rang de la matrice de corrélation du groupe de variables initiales qui est le plus petit.

Une description plus détaillée de l'analyse des corrélations canoniques peut être trouvée dans une autre note [PALM, 1990].

Le principe de calcul des variables canoniques et des coefficients de corrélation canonique peut s'appliquer aux données utilisées pour une analyse de la variance multivariée. Les variables soumises à l'analyse de la variance constituent un premier ensemble et le deuxième ensemble est représenté par g variables indicatrices, de type 0/1, permettant de décrire l'appartenance d'un individu particulier à l'un des g groupes.

Les carrés des coefficients de corrélation canonique sont les valeurs propres, l'_i , des matrices $(\mathbf{H} + \mathbf{E})^{-1} \mathbf{H}$ ou $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$ et les coefficients des combinaisons linéaires relatives à l'ensemble des variables soumises à l'analyse de la variance sont les vecteurs propres associés aux valeurs propres des matrices $\mathbf{E}^{-1} \mathbf{H}$ ou $(\mathbf{H} + \mathbf{E})^{-1} \mathbf{H}$, ces deux matrices ayant les mêmes vecteurs propres (paragraphe 2.2).

Les coefficients des combinaisons linéaires des variables indicatrices ne présentent pas d'intérêt particulier et, par la suite, lorsque nous parlerons de variables canoniques, il s'agira toujours des combinaisons linéaires des variables soumises à l'analyse de la variance.

Les coefficients de ces combinaisons linéaires représentent les poids des variables initiales dans la variable canonique qui est construite. A ce sujet, il faut remarquer que plusieurs présentations de ces coefficients peuvent être adoptées. Tout d'abord, les coefficients peuvent être définis de manière à être appliqués aux variables initiales centrées, mais non réduites ou, au contraire, aux variables centrées et réduites. Dans ce dernier cas, la réduction des variables initiales peut être faite de manière à ce que l'écart-type total de la variable soit unitaire, ou, au contraire, de manière à ce que l'écart-type résiduel de la variable soit unitaire.

D'autre part, les combinaisons linéaires sont toujours définies à une constante près et deux modalités de standardisation sont utilisées : les coefficients peuvent être définis de manière à ce que la variable canonique soit de variance totale unitaire, ou, au contraire, de manière à ce que cette variable canonique soit de variance résiduelle unitaire.

A titre d'illustration, la figure 6, obtenue par la procédure CANDISC de SAS, donne trois ensembles de coefficients canoniques. La première série correspond aux coefficients qu'il faut appliquer aux variables PLONG, PLARG, SLONG et SLARG, après que celles-ci aient été centrées et réduites, la réduction se faisant en divisant les écarts par rapport à la moyenne par l'écart-type total de la variable. Ainsi, pour la première variable canonique, notée CAN1 :

$$\text{CAN1} = 3,8858[(\text{PLONG} - 3,758)/1,7653] + \dots$$

les valeurs 3,758 et 1,7653 représentent respectivement la moyenne et l'écart-type de PLONG pour les 150 iris.

La deuxième série de coefficients correspond aux coefficients qu'il faut appliquer aux variables, lorsque celles-ci sont centrées et réduites, la réduction se faisant en divisant les écarts par rapport à la moyenne par l'écart-type résiduel de la variable. On a, par exemple :

$$\text{CAN1} = 0,9472[(\text{PLONG} - 3,758)/0,4303] + \dots,$$

la valeur 0,4303 étant l'écart-type résiduel de la variable PLONG, qu'on peut déduire de l'analyse de la variance univariée (figure 1).

Enfin, la troisième série de coefficients correspond aux coefficients qu'il faut appliquer aux variables lorsque celles-ci sont centrées mais non réduites :

$$\text{CAN1} = 2,2012(\text{PLONG} - 3,758) + \dots$$

Les trois ensembles de coefficients conduisent aux mêmes variables canoniques, celles-ci étant standardisées de manière à ce que leur moyenne soit nulle et leur écart-type résiduel unitaire. On peut vérifier que la moyenne générale est bien nulle, à partir des moyennes générales par groupe données dans la figure 6. Pour la première variable canonique, par exemple, on a :

$$(-7,6076 + 1,8250 + 5,7826)/3 = 0.$$

De manière plus générale, en présence de groupes d'effectifs non constants, les moyennes des groupes doivent être pondérées par les effectifs des groupes.

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2	
PLONG	3.885795047	-1.645118866	LONGUEUR DES PETALES (CM)
PLARG	2.142238715	2.164135931	LARGEUR DES PETALES (CM)
SLONG	-0.686779533	0.019958173	LONGUEUR DES SEPALES (CM)
SLARG	-0.668825075	0.943441829	LARGEUR DES SEPALES (CM)

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2	
PLONG	0.9472572487	-.4010378190	LONGUEUR DES PETALES (CM)
PLARG	0.5751607719	0.5810398645	LARGEUR DES PETALES (CM)
SLONG	-.4269548486	0.0124075316	LONGUEUR DES SEPALES (CM)
SLARG	-.5212416758	0.7352613085	LARGEUR DES SEPALES (CM)

Raw Canonical Coefficients

	CAN1	CAN2	
PLONG	2.201211656	-0.931921210	LONGUEUR DES PETALES (CM)
PLARG	2.810460309	2.839187853	LARGEUR DES PETALES (CM)
SLONG	-0.829377642	0.024102149	LONGUEUR DES SEPALES (CM)
SLARG	-1.534473068	2.164521235	LARGEUR DES SEPALES (CM)

Class Means on Canonical Variables

ESPECE	CAN1	CAN2
SETOSA	-7.607599927	0.215133017
VERSIC	1.825049490	-0.727899622
VIRGIN	5.782550437	0.512766605

Figure 6. Coefficients canoniques et moyennes des variables canoniques par espèce.

Dependent Variable: CAN1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4732.213592	2366.106796	2366.11	0.0001
Error	147	147.000000	1.000000		
Corrected Total	149	4879.21359215			

Dependent Variable: CAN2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41.952483	20.976242	20.98	0.0001
Error	147	147.000000	1.000000		
Corrected Total	149	188.95248327			

Figure 7. Analyses de la variance réalisées sur les valeurs des deux variables canoniques.

Les analyses de la variance, réalisées à titre d'illustration (figure 7), montrent également que les variances résiduelles sont bien égales à l'unité.

La première variable canonique discriminante est telle que le rapport de la variance entre les groupes à la variance dans les groupes est maximum. Le rapport :

$$R^2 = \text{SCE}_f / (\text{SCE}_f + \text{SCE}_e),$$

obtenu lorsqu'on réalise l'analyse de la variance sur les valeurs de la première variable canonique est égal au carré du premier coefficient de corrélation canonique, c'est-à-dire aussi égal à l'_1 .

La seconde variable canonique vise aussi à maximiser le rapport de la somme des carrés des écarts factorielle et de la somme des carrés des écarts totale, avec, en plus, la contrainte de non-corrélation à la première variable déjà construite. Ce rapport est égal au carré du deuxième coefficient de corrélation canonique et donc aussi à l'_2 . Et ainsi de suite pour les autres variables canoniques, lorsque $s > 2$.

Pour l'exemple considéré, on vérifie bien que les coefficients R^2 des analyses de la variance réalisées sur les deux variables canoniques sont égaux à (figure 7 et paragraphe 2.2) :

$$l'_1 = 0,96987 \quad \text{et} \quad l'_2 = 0,22203.$$

Géométriquement, les valeurs des variables canoniques correspondent aux projections des points individus sur une série d'axes qui représentent les directions dans lesquelles les différences entre les groupes sont les plus marquées.

Exprimées en pour cent de leur somme, les valeurs propres de $\mathbf{E}^{-1} \mathbf{H}$ donnent une idée de la qualité de la représentation sur les différents axes. Ainsi, si la première valeur propre correspond à un pourcentage important, cela signifie que, dans l'espace initial à p dimensions, les g moyennes se trouvent à proximité d'une droite. Si la somme des deux premières valeurs propres correspond à un pourcentage important, les g moyennes se situent à proximité d'un plan. Et ainsi de suite, pour trois valeurs propres, quatre valeurs propres, etc.

Compte tenu de ce qui vient d'être dit, on comprend aisément que le nombre de valeurs propres non nulles est, au maximum, égal à p et à $g - 1$. En effet, si on a par exemple deux variables et plus de trois groupes, les vecteurs de moyennes sont nécessairement dans un espace à une ou deux variables (sur un axe ou dans un plan), puisque la dimension maximum de l'espace est égale à 2. Si par contre, on a par exemple trois groupes et quatre variables, la plus grande dimension de l'espace dans lequel se trouvent les trois vecteurs de moyennes est un plan, puisque par trois points d'un espace on peut toujours faire passer un plan.

Dans le cas particulier de deux groupes ($g = 2$), et quel que soit le nombre p de variables, l'unique variable canonique peut être déterminée très simplement par régression linéaire multiple. Il suffit de créer une variable indicatrice, y , à deux modalités, en donnant à y par exemple la valeur 0 si l'individu appartient au premier groupe et la valeur 1, s'il appartient au deuxième groupe. Ensuite, on calcule la régression multiple :

$$y = a + b_1 x_1 + \dots + b_p x_p,$$

et, à une transformation linéaire près, les valeurs de y données par l'équation sont les valeurs de la variable canonique.

Afin d'illustrer davantage la signification géométrique des variables canoniques, considérons l'exemple des iris, en nous limitant à deux variables, afin de permettre la représentation graphique dans un plan.

Si on ne prend en compte que les variables PLONG et PLARG, par exemple, les deux variables canoniques s'écrivent :

$$\text{CAN1} = 0,665 \text{ PLONGR} + 0,492 \text{ PLARGR}$$

et
$$\text{CAN2} = -0,930 \text{ PLONGR} + 1,032 \text{ PLARGR},$$

PLONGR et PLARGR représentant les variables PLONG et PLARG après standardisation par rapport à l'écart-type résiduel.

La figure 8 donne le diagramme de dispersion des longueurs et largeurs des pétales (en variables centrées réduites par rapport à l'écart-type résiduel). Elle reprend aussi la position des axes canoniques. On voit clairement que les iris possèdent une variabilité bien plus grande lorsqu'on les projette sur le premier axe canonique que lorsqu'on les projette sur le deuxième axe canonique.

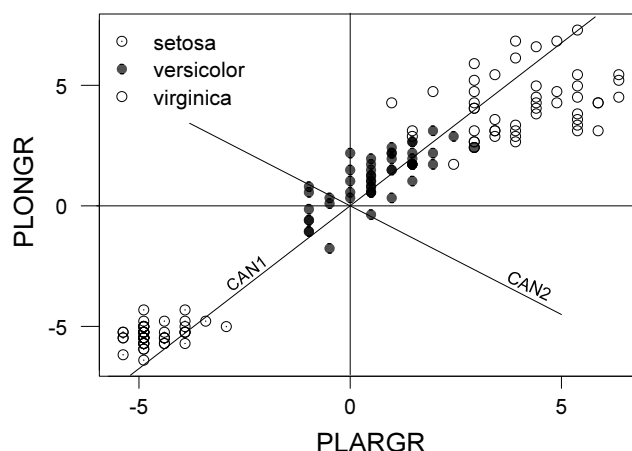


Figure 8. Diagramme de dispersion de PLONG et PLARG (données centrées réduites) et position des deux axes canoniques.

En pratique, on porte rarement les deux systèmes d'axes sur un même graphique, mais on représente couramment les observations dans le plan formé par les axes CAN1 et CAN2, comme dans la figure 9, obtenue à la suite de l'analyse canonique discriminante sur les quatre variables. On notera que des échelles différentes sont utilisées en abscisse et en ordonnée, afin d'éviter un allongement excessif de la figure. Cette distorsion a cependant comme conséquence de faire apparaître moins clairement les différences importantes de variances qui existent entre les deux variables canoniques.

Dans les cas où on a déterminé plus de deux variables canoniques, on réalise une série de représentations graphiques à deux dimensions, en combinant les variables canoniques deux à deux, comme on le fait lors de l'analyse en composantes principales. La prise en compte des variables canoniques d'ordre plus élevé ne se justifie cependant que si celles-ci présentent un intérêt pratique. Ce problème fait l'objet du paragraphe suivant.

3.2. Test de signification et importance relative des variables canoniques

Lorsque les conditions d'application de l'analyse de la variance sont remplies (g populations normales à p dimensions, de même matrice de variances et covariances et échantillons aléatoires, simples et indépendants), on peut tester la signification des coefficients de corrélation canonique.

Pour le premier coefficient, on émet l'hypothèse nulle suivante:

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

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

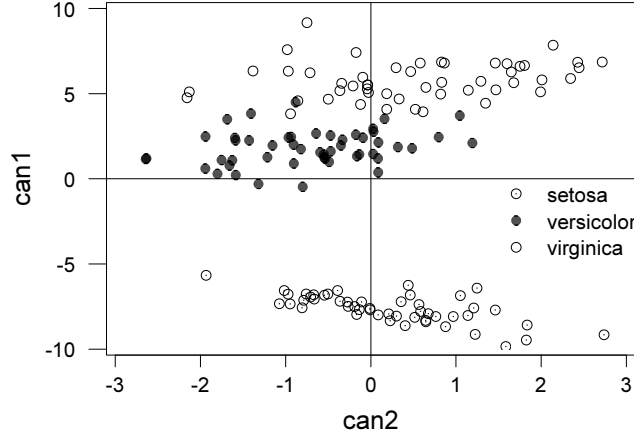


Figure 9. Diagramme de dispersion des 150 iris dans le plan canonique défini par les quatre variables initiales.

On calcule alors la valeur :

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

qu'on compare, après transformation, à la valeur $F_{1-\alpha}$. Ce test est en fait le test de WILKS décrit au paragraphe 2.2 car $r_k^2 = l'_k$ (paragraphe 3.1).

Pour le deuxième coefficient, on émet l'hypothèse nulle suivante :

$$H_0 : \rho_2 = \rho_3 = \dots = \rho_s = 0,$$

et on calcule :

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

qu'on compare, après transformation, à la valeur $F_{1-\alpha}$.

La procédure séquentielle se poursuit jusqu'à ce qu'on accepte l'hypothèse de nullité des $s - (l + 1)$ derniers coefficients de corrélation. On considère alors que les g moyennes se situent dans un espace à l dimensions. Des informations concernant les transformations des valeurs Λ_{obs} en valeurs F_{obs} peuvent être trouvées notamment dans HUBERTY [1994] et PALM [1990].

L'importance relative d'une variable canonique est mesurée par la valeur propre de $\mathbf{E}^{-1} \mathbf{H}$, exprimée en proportion de la somme des valeurs propres. Elle correspond à la proportion de variation factorielle qui est prise en compte par la variable canonique. Les proportions peuvent s'additionner, car les variables canoniques sont non corrélées. Ainsi, la proportion de variation factorielle prise

en compte par un plan canonique est égale à la somme des proportions prises en compte par les deux variables canoniques déterminant ce plan, et ainsi de suite pour les sous-espaces de dimension de plus en plus grande.

La figure 10 donne les résultats de l'analyse canonique discriminante pour l'exemple des iris. Dans la partie supérieure de la figure, on retrouve les deux coefficients de corrélation canonique, les coefficients de corrélation canonique ajustés et l'erreur-standard des coefficients de corrélation canonique.

Les coefficients de corrélation donnés dans la première colonne sont, en effet, des estimations biaisées des coefficients de corrélation canonique théoriques et LAWLEY [1959] a proposé des estimations moins biaisées, du moins lorsque les coefficients de corrélation canonique théoriques ne présentent, ni des valeurs trop proches les unes des autres, ni des 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 [SAS, 1989].

L'erreur-standard approximative des coefficients de corrélation canonique est égale à :

$$(1 - r_k^2)/\sqrt{n},$$

r_k^2 étant le $k^{ième}$ coefficient de corrélation canonique observé et n l'effectif total.

La figure 10 donne aussi les informations relatives aux valeurs propres de $\mathbf{E}^{-1} \mathbf{H}$ et les tests de signification des coefficients de corrélation canonique. Bien que les deux coefficients de corrélation canonique doivent être considérés comme significativement différents de zéro, on note que le premier coefficient est nettement plus grand que le second et surtout que la première valeur propre de $\mathbf{E}^{-1} \mathbf{H}$ est nettement plus grande que la deuxième valeur propre. Cette première valeur propre correspond à 99,12 % de la somme des valeurs propres. Cela signifie donc que les différences entre les trois vecteurs de moyennes se marquent essentiellement selon une direction.

Le pourcentage ci-dessus est aussi égal à la somme des carrés des écarts factorielle de la première variable canonique, exprimée en proportion de la somme des deux sommes des carrés des écarts factorielles (figure 7) :

$$4.732,21/(4.732,21 + 41,95) = 0,9912.$$

Ce pourcentage est encore égal à la somme des carrés des coordonnées des trois moyennes sur la première variable canonique, exprimée en pour cent de la somme des carrés des coordonnées des trois moyennes sur les deux axes (figure 6) :

$$\begin{aligned} & [(-7,6076)^2 + 1,8250^2 + 5,7826^2] / \\ & [(-7,6076)^2 + \dots + 5,7826^2 + 0,2151^2 + \dots + 0,5128^2] \\ & = 94,6447/(94,6447 + 0,8391) = 0,9912. \end{aligned}$$

Cette dernière relation n'est exacte que lorsque les effectifs des échantillons sont constants. D'une manière plus générale, il faut multiplier les carrés des

Eigenvalues of $INV(E)*H$ = $CanRsq/(1-CanRsq)$								
	Canonical Correl.	Adjusted Canonical Correl.	Approx Standard Error	Squared Canonical Correl.	Eigenvalue	Differ.	Pro-portion	Cumu-lative
1	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912
2	0.471197	0.461445	0.063734	0.222027	0.2854	.	0.0088	1.0000

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

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.02343863	199.1453	8	288	0.0001
2	0.77797337	13.7939	3	145	0.0001

Total Canonical Structure

	CAN1	CAN2	
PLONG	0.984951	0.046037	LONGUEUR DES PETALES (CM)
PLARG	0.972812	0.222902	LARGEUR DES PETALES (CM)
SLONG	0.791888	0.217593	LONGUEUR DES SEPALES (CM)
SLARG	-0.530759	0.757989	LARGEUR DES SEPALES (CM)

Between Canonical Structure

	CAN1	CAN2	
PLONG	0.999750	0.022358	LONGUEUR DES PETALES (CM)
PLARG	0.994044	0.108977	LARGEUR DES PETALES (CM)
SLONG	0.991468	0.130348	LONGUEUR DES SEPALES (CM)
SLARG	-0.82565	0.564171	LARGEUR DES SEPALES (CM)

Pooled Within Canonical Structure

	CAN1	CAN2	
PLONG	0.706065	0.167701	LONGUEUR DES PETALES (CM)
PLARG	0.633178	0.737242	LARGEUR DES PETALES (CM)
SLONG	0.222596	0.310812	LONGUEUR DES SEPALES (CM)
SLARG	-0.119012	0.863681	LARGEUR DES SEPALES (CM)

Figure 10. Analyse canonique discriminante et corrélations des variables initiales avec les variables canoniques.

coordonnées de chaque moyenne par l'effectif de l'échantillon. On retrouve alors les sommes des carrés des écarts factorielle (figure 7):

$$(50)(94, 6447) = 4732, 20 \quad \text{et} \quad (50)(0, 8391) = 41, 96 .$$

Enfin, le pourcentage mentionné ci-dessus est encore égal au rapport de la somme des carrés des distances euclidiennes entre les moyennes mesurées sur le premier axe et la somme des carrés des distances euclidiennes mesurées dans l'espace des deux variables canoniques. Ainsi, pour *setosa* et *versicolor*, on a, sur le premier axe (figure 6) :

$$\left(\widehat{\Delta}_{12}^2\right)_1 = (-7, 6076 - 1, 8250)^2 = 88, 97 ,$$

et dans l'espace complet :

$$\left(\widehat{\Delta}_{12}^2\right)_{12} = (-7, 6076 - 1, 8250)^2 + (0, 2151 - (-0, 7279))^2 = 89, 86 .$$

De même, pour les autres couples d'espèces, on a :

$$\left(\widehat{\Delta}_{13}^2\right)_1 = 179, 30 \quad \text{et} \quad \left(\widehat{\Delta}_{23}^2\right)_1 = 15, 66$$

$$\left(\widehat{\Delta}_{13}^2\right)_{12} = 179, 38 \quad \text{et} \quad \left(\widehat{\Delta}_{23}^2\right)_{12} = 17, 20$$

$$\text{et} \quad (88, 97 + 179, 30 + 15, 66) / (89, 86 + 179, 38 + 17, 20) = 0, 9912 .$$

On remarque aussi que les distances euclidiennes entre les moyennes dans le plan canonique sont identiques aux distances de MAHALANOBIS entre les moyennes dans l'espace des variables initiales (figure 4).

Ces différents modes de calcul de la proportion expliquée par le premier axe canonique nous éclairent sur l'interprétation de ce pourcentage. D'une manière générale, une valeur propre de $\mathbf{E}^{-1} \mathbf{H}$, lorsqu'elle est exprimée en proportion de la somme des valeurs propres, correspond à la proportion de variation factorielle qui est prise en considération par l'axe canonique associé à la valeur propre.

3.3. Interprétation des variables canoniques

L'analyse discriminante revient donc à remplacer p variables initiales, le plus souvent corrélées, par un nombre généralement plus réduit de variables canoniques, qui sont non corrélées. Le problème se pose alors de tenter de trouver à quoi correspondent ces nouvelles variables.

L'interprétation des variables canoniques se base sur l'examen des corrélations entre les variables initiales et les variables discriminantes et, plus précisément, sur l'examen des corrélations intragroupes qui éliminent les différences entre les moyennes des groupes : on recherche, d'une part, les variables initiales

qui présentent les corrélations positives les plus importantes et, d'autre part, celles qui présentent les corrélations négatives les plus importantes avec une variable canonique donnée et on s'efforce, à partir de ces corrélations, de trouver une interprétation de la variable canonique.

Comme pour toutes les méthodes statistiques multivariées basées sur le calcul de combinaisons linéaires, il faut remarquer que les variables artificielles qui sont construites n'ont pas toujours une signification physique réelle, ce qui rend alors leur interprétation moins aisée.

Pour les trois espèces d'iris, l'examen des corrélations intragroupes des variables canoniques et des variables initiales (figure 10) montre que la première variable canonique est surtout corrélée à la longueur et à la largeur des pétales. Il s'agit donc d'un axe qui représente la taille des pétales. Sur cet axe, les moyennes des trois espèces s'ordonnent de la manière suivante: *setosa*, *versicolor*, *virginica*. L'espèce *setosa* est donc caractérisée par des pétales de petite taille, l'espèce *versicolor* par des pétales de taille intermédiaire et *virginica* par des pétales de taille plus grande. Cette interprétation est confirmée par l'examen des moyennes données dans le tableau 1.

La deuxième variable canonique est corrélée principalement à la largeur des pétales et des sépales et les moyennes des trois espèces s'ordonnent de la manière suivante: *versicolor*, *setosa*, *virginica*. L'interprétation de cet axe est difficile car les différences entre espèces sont peu marquées, comme nous l'avons constaté ci-dessus. De plus, cet axe doit être vu comme une forme de correction de l'interprétation des vecteurs de moyennes données par le premier axe. Le deuxième axe permet de nuancer cette interprétation: la différence entre les vecteurs des moyennes est effectivement liée à la taille globale des pétales (interprétation liée aux différences selon le premier axe), mais à taille globale des pétales identique, le deuxième axe permet de prendre en compte l'importance de la largeur des pièces du périanthe (largeur des pétales et largeur des sépales). Ainsi, il ne serait pas correct d'affirmer que le deuxième axe montre que *versicolor* est caractérisé par des pièces du périanthe étroites, que *virginica* est caractérisé par des pièces larges et que *setosa* a des pièces de largeur intermédiaire. Une telle interprétation ne serait d'ailleurs pas confirmée par l'examen des moyennes du tableau 1.

4. SÉLECTION ET HIÉRARCHISATION DES VARIABLES

4.1. Sélection des variables

Il arrive fréquemment que l'utilisateur de l'analyse de la variance multivariée et de l'analyse canonique discriminante dispose d'un nombre relativement élevé de variables pour décrire les groupes et qu'il souhaite identifier, ou même éliminer, les variables qui apportent peu ou pas d'information pour la différenciation de ces groupes.

Le problème est tout à fait similaire au problème de sélection de variables en régression multiple et les logiciels statistiques offrent, comme en régression

multiple, des procédures d'analyse discriminante pas à pas. Dans le cas particulier de deux groupes, le choix des variables peut d'ailleurs être réalisé par un programme de régression multiple pas à pas, puisque, dans ce cas, la variable canonique peut être déterminée par régression, comme nous l'avons signalé au paragraphe 3.1.

Si la discrimination doit se faire entre plus de deux populations normales de même matrice de variances et covariances, une extension des procédures de sélection des variables pas à pas peut être envisagée. Comme en régression multiple, le principe consiste à mesurer, à une étape donnée du processus, l'apport d'une variable, compte tenu des variables déjà sélectionnées.

Cet apport de la variable peut se mesurer par l'intermédiaire des variables de WILKS, qui interviennent dans l'analyse de la variance multivariée. Le principe consiste à comparer les valeurs Λ de WILKS obtenues, d'une part, lors de l'analyse de la variance multivariée sur les variables déjà sélectionnées et, d'autre part, lors de l'analyse de la variance après adjonction de la variable dont on veut tester l'utilité. En effet, si p' variables parmi les p variables disponibles sont déjà sélectionnées et si $\Lambda_{p'}$ et $\Lambda_{p'+1}$ sont les valeurs du rapport de WILKS pour p' et $p' + 1$ variables, on a la relation suivante :

$$F_{obs} = \frac{n - g - p'}{g - 1} \left(\frac{\Lambda_{p'}}{\Lambda_{p'+1}} - 1 \right),$$

la valeur F_{obs} étant une valeur observée d'une variable F de FISHER-SNEDECOR à $k_1 = g - 1$ et $k_2 = n - g - p'$ degrés de liberté [HUBERTY, 1994; SEBER, 1984].

Tout comme en régression multiple, l'utilisation d'algorithmes de sélection des variables pas à pas ne garantit pas que l'ensemble des p' variables sélectionnées à une étape donnée du processus soit la meilleure combinaison de p' variables. Des algorithmes permettant l'étude de toutes les combinaisons de p' variables sont, pour cette raison, proposés dans la littérature [HUBERTY, 1994].

A titre d'illustration, la figure 11 reprend les deux premières étapes de la sélection progressive des variables dans le cas des trois espèces d'iris décrites par quatre variables.

Cette figure a été obtenue par la procédure STEPDISC du logiciel SAS. Des informations concernant cette procédure sont données dans le manuel d'utilisation [SAS, 1989]. Nous nous limiterons à l'examen des seules valeurs F_{obs} .

A la première étape, le logiciel calcule les valeurs relatives aux quatre analyses de la variance univariées. On constate que la variable la plus discriminante est la longueur des pétales, PLONG ($F_{obs} = 1.180$). La variable PLONG est donc introduite en premier lieu. La valeur observée de la variable de WILKS est alors égale à 0,0586.

A la deuxième étape, on constate que la variable SLARG est ajoutée à la variable PLONG, car la valeur F_{obs} relative à l'adjonction de cette variable est la plus grande et est significative au niveau de signification fixé à 0,05. La valeur de la variable de WILKS, pour le modèle à deux variables, est égale à 0,0369.

Stepwise Selection: Step 1

Statistics for Entry, DF = 2, 147

Variable	R**2	F	Prob>F	Tolerance	Label
PLONG	0.9414	1180.161	0.0001	1.0000	LONGUEUR DES PETALES (CM)
PLARG	0.9289	960.007	0.0001	1.0000	LARGEUR DES PETALES (CM)
SLONG	0.6187	119.265	0.0001	1.0000	LONGUEUR DES SEPALES (CM)
SLARG	0.400	49.160	0.0001	1.0000	LARGEUR DES SEPALES (CM)

Variable PLONG will be entered
The following variable(s) have been entered: PLONG

Multivariate Statistics

Wilks' Lambda = 0.05862828 F(2, 147) = 1180.161 Prob>F = 0.0001
Pillai's Trace = 0.941372 F(2, 147) = 1180.161 Prob>F = 0.0001
Average Squared Canonical Correlation = 0.47068586

Stepwise Selection: Step 2

Statistics for Removal, DF = 2, 147

Variable	R**2	F	Prob>F	Label
PLONG	0.9414	1180.161	0.0001	LONGUEUR DES PETALES (CM)

No variables can be removed

Statistics for Entry, DF = 2, 146

Variable	Partial R**2	F	Prob>F	Tolerance	Label
PLARG	0.2533	24.766	0.0001	0.0729	LARGEUR DES PETALES (CM)
SLONG	0.3198	34.323	0.0001	0.2400	LONGUEUR DES SEPALES (CM)
SLARG	0.3709	43.035	0.0001	0.8164	LARGEUR DES SEPALES (CM)

Variable SLARG will be entered
The following variable(s) have been entered: PLONG SLARG

Multivariate Statistics

Wilks' Lambda = 0.03688411 F(4, 292) = 307.105 Prob>F = 0.0001
Pillai's Trace = 1.119908 F(4, 294) = 93.528 Prob>F = 0.0001
Average Squared Canonical Correlation = 0.55995394

Figure 11. Procédure de sélection des variables pas à pas : résultats relatifs aux deux premières étapes.

On peut ainsi vérifier que la valeur F_{obs} pour l'adjonction de SLARG à PLONG est bien égale à :

$$\frac{150 - 3 - 1}{2} \left(\frac{0,058628}{0,036884} - 1 \right) = 43,035.$$

Aux deux étapes suivantes, qui ne sont pas reprises dans la figure 11, les variables PLARG et SLONG sont successivement introduites car les valeurs F_{obs} liées à l'addition de ces variables, respectivement égales à 34,57 et à 4,72, sont, toutes deux, significatives.

4.2. Hiérarchisation des variables

Pour quantifier le pouvoir discriminant de p variables, HUBERTY [1994] recommande de réaliser p analyses à $p - 1$ variables. Soit $\Lambda_{(-j)}$ la valeur observée de la variable de WILKS, lorsque la variable j est exclue. Les valeurs $\Lambda_{(-j)}$ sont alors classées par ordre décroissant, et, à côté de chacune de ces valeurs, on indique le nom de la variable j qui ne figure pas dans l'analyse de la variance en question. On obtient ainsi un classement des variables, de la plus discriminante à la moins discriminante. On pourra toutefois considérer comme *ex aequo* des variables dont l'élimination conduit à des valeurs de $\Lambda_{(-j)}$ qui sont proches.

Une solution tout à fait équivalente consiste à calculer les valeurs F_{obs} relatives à l'apport de la $p^{ième}$ variable, alors qu'il y a déjà $p - 1$ variables retenues, par la formule donnée au paragraphe 4.1, en considérant que $p' = p - 1$.

Une troisième solution, équivalente aux deux autres, se base sur le calcul des valeurs $R^2_{partiel}$, telles que définies dans la procédure STEPDISC de SAS :

$$R^2_{partiel} = \frac{\Lambda_{(-j)} - \Lambda_{obs}}{\Lambda_{(-j)}},$$

Λ_{obs} étant la valeur de la variable de WILKS pour les p variables. Ce $R^2_{partiel}$ mesure l'augmentation de R^2 due à la variable j , en proportion de ce qui n'est pas expliqué par les $p - 1$ autres variables. En effet, nous avons vu, au paragraphe 2.3., que pour p variables :

$$R^2 = 1 - \Lambda_{obs}.$$

De même, si on élimine la variable j , on a :

$$R^2_{(-j)} = 1 - \Lambda_{(-j)},$$

et il en résulte que :

$$R^2_{partiel} = \frac{R^2 - R^2_{(-j)}}{1 - R^2_{(-j)}}.$$

La procédure de classement des variables qui vient d'être décrite repose implicitement sur l'idée que le modèle complet à p variables sert de référence.

Stepwise Selection: Step 5

Statistics for Removal, DF = 2, 144

Variable	Partial	F	Prob > F	Label
	R**2			
PLONG	0.3308	35.590	0.0001	LONGUEUR DES PETALES (CM)
PLARG	0.2570	24.904	0.0001	LARGEUR DES PETALES (CM)
SLONG	0.0615	4.721	0.0103	LONGUEUR DES SEPALES (CM)
SLARG	0.2335	21.936	0.0001	LARGEUR DES SEPALES (CM)

Figure 12. Procédure de sélection des variables : résultats de la dernière étape.

La figure 12 reprend la dernière étape de la procédure STEPDISC, qui donne les valeurs $R^2_{partiel}$ et les valeurs F_{obs} pour chacune des quatre variables. Le classement des variables, de la plus discriminante à la moins discriminante, est donc le suivant :

PLONG, PLARG, SLARG et SLONG .

On peut remarquer que cet ordre est légèrement différent de l'ordre basé sur les valeurs F_{obs} des analyses de la variance univariées (figure 1) et qui est le suivant :

PLONG, PLARG, SLONG et SLARG .

Le classement est différent aussi de l'ordre d'introduction des variables dans la procédure STEPDISC, qui est (paragraphe 4.1) :

PLONG, SLARG, PLARG et SLONG .

5. AUTRES MODÈLES D'ANALYSE DE LA VARIANCE

5.1. Principes généraux

Dans les paragraphes précédents, nous nous sommes concentrés sur l'analyse de la variance multivariée à un critère. L'approche multivariée peut cependant être étendue à tous les autres modèles d'analyse de la variance qu'on rencontre dans le cas univarié.

Dans le cas univarié, le test relatif à un facteur donné se fait toujours en considérant le rapport du carré moyen relatif à la source de variation qu'on souhaite tester au carré moyen qui sert de comparaison. De même, en analyse de la variance multivariée, on retrouve les différents tests qui ont été présentés au

paragraphe 2.2. La matrice \mathbf{H} représente la matrice des sommes des carrés des écarts et des produits des écarts pour la source de variation qu'on souhaite tester et la matrice \mathbf{E} correspond à la matrice des sommes des carrés des écarts et des produits des écarts pour la source de variation qui sert de base de comparaison.

Les coefficients d'association, définis au paragraphe 2.3, peuvent également être calculés à partir des critères de WILKS, PILLAI, etc. Ces coefficients sont parfois appelés coefficients d'association partiels. Ainsi, pour un modèle d'analyse de la variance à deux critères croisés fixes, le coefficient R^2 pour le facteur A reflète le degré d'association entre les p variables soumises à l'analyse et le groupement des individus selon le facteur A , lorsque les effets du facteur B et de l'interaction AB ont été enlevés.

A partir des vecteurs propres de $\mathbf{E}^{-1}\mathbf{H}$, on peut aussi obtenir les variables canoniques relatives à chacune des sources de variation et la signification de ces variables peut être testée par l'intermédiaire des variables de WILKS, comme pour l'analyse à un critère.

D'autre part, les problèmes liés à l'interprétation des variables canoniques et à la sélection ou au classement des variables initiales se posent de la même manière que pour l'analyse de la variance à un critère.

5.2. Exemple : comparaison de la croissance de pommiers

Pour illustrer l'analyse de la variance multivariée à plusieurs critères, nous reprenons les données traitées par DEBOUCHE [1977] qui proviennent de l'étude de la croissance d'une variété de pommiers dans des conditions contrôlées de température de l'air et du sol. Le dispositif expérimental utilisé est le suivant : 48 pommiers en pots sont répartis dans deux chambres de culture. Dans l'une, la température de l'air est de 15°C et dans l'autre elle est de 17°C. Chaque chambre de culture est divisée en deux parties : dans une des parties, la température du sol est maintenue à 16°C et, dans l'autre partie, cette température est de 18°C. Douze pommiers sont installés dans chacune des deux parties, à raison de trois groupes de quatre pommiers, chaque groupe occupant une cellule.

Le modèle d'analyse de la variance correspondant à ce dispositif est un modèle à trois critères (facteur "température de l'air", facteur "température du sol" et facteur "cellule"). Les deux premiers facteurs sont fixes et croisés. Le facteur "cellule" est considéré comme aléatoire et est hiérarchisé par rapport aux deux autres facteurs. Les deux facteurs fixes et leur interaction seront testés par rapport au facteur "cellule".

Sur les 48 pommiers, la longueur de la pousse terminale a été observée chaque semaine durant la période de croissance, soit durant 17 semaines. Des ajustements non linéaires ont permis de condenser l'information relative à chaque arbre. Le modèle de GOMPertz, qui correspond à une courbe sigmoïde non symétrique :

$$y = M \exp \left[- \exp \left(- \frac{x - a}{b} \right) \right],$$

Tableau 2. Moyennes et, entre parenthèses, écarts-types des paramètres pour les différents niveaux de température de l'air et du sol .

Niveaux	M	b	a
AIR = 15	191,5 (39,8)	16,8 (3,3)	68,4 (4,1)
AIR = 17	177,0 (31,4)	15,7 (2,7)	60,0 (4,8)
SOL = 16	176,0 (43,3)	17,4 (3,3)	64,1 (6,3)
SOL = 18	192,5 (25,8)	15,1 (2,2)	64,3 (6,1)
AIR = 15 et SOL = 16	184,8 (51,2)	18,3 (3,7)	68,5 (4,6)
AIR = 15 et SOL = 18	198,1 (24,3)	15,3 (1,8)	68,3 (3,6)
AIR = 17 et SOL = 16	167,1 (33,4)	16,5 (2,7)	59,7 (4,2)
AIR = 17 et SOL = 18	186,9 (27,0)	15,0 (2,6)	60,2 (5,4)

a été retenu. Dans cette relation, y est la taille de la pousse, en cm, x est le temps en jours. M , a et b sont les trois paramètres de la courbe. M est l'asymptote horizontale, c'est-à-dire la valeur maximale vers laquelle tend y lorsque x augmente indéfiniment. Le paramètre a est une valeur particulière de x qui situe la courbe sur l'axe des x par son point d'inflexion : c'est pour $x = a$ que la vitesse de croissance est maximum. Le paramètre b est exprimé dans les unités de x et est sensible à l'étalement du phénomène de croissance, car le temps de croissance permettant de passer d'une taille de $0,05 M$ à $0,95 M$ est égal à $4b$ [DEBOUCHE, 1977]. Les valeurs obtenues pour les 48 courbes sont reprises en annexe et les analyses de la variance vont porter sur ces valeurs.

En ce qui concerne les conditions d'application, le caractère aléatoire et simple des échantillons est assuré par le dispositif expérimental (répartition au hasard des 48 arbres dans les 12 cellules). Une vérification partielle de la condition de multinormalité consiste à examiner les distributions marginales des résidus de l'analyse de la variance multivariée. Le test de normalité de SHAPIRO-WILK a conduit aux probabilités suivantes : 0,833 pour M , 0,069 pour a et 0,072 pour b . On accepte donc la normalité des résidus pour chacune des variables. Bien que la normalité des distributions marginales n'implique pas la normalité à trois dimensions, nous supposons que cette condition est néanmoins remplie. La non-normalité de la distribution ne serait d'ailleurs pas un obstacle majeur à l'utilisation de l'analyse de la variance étant donné la robustesse de l'analyse de la variance, comme nous le verrons au paragraphe 6.1. Quant à l'égalité des matrices de variances et covariances, le test qui sera présenté au paragraphe 6.1 conduit à une probabilité de 0,35. On peut donc accepter l'hypothèse d'égalité des matrices de variances et covariances des populations au sein des 12 cellules

Les moyennes et les écarts-types pour les deux niveaux du facteur "température de l'air" et du facteur "température du sol" et pour les combinaisons des niveaux de ces deux facteurs sont donnés dans le tableau 2, et le tableau 3 reprend les probabilités associées aux valeurs F_{obs} , obtenues pour les trois analyses de

Tableau 3. Probabilités associées aux valeurs F_{obs} pour les trois analyses de la variance univariées.

Sources de variation	M	b	a
air	0,198	0,390	0,000
sol	0,147	0,093	0,893
air*sol	0,758	0,540	0,775

Tableau 4. Corrélations des variables canoniques liées au facteur "température de l'air" et au facteur "température du sol" avec les paramètres M , b , a .

Paramètres	Air	Sol
M	0,19	-0,32
b	0,12	0,38
a	0,87	-0,03

la variance univariées. On constate qu'aucune interaction n'est significative, ce qui nous autorise à examiner les effets principaux. Seul l'effet "température de l'air" est significatif pour le paramètre a . Le tableau des moyennes montre que la température de l'air élevée conduit à une moyenne du paramètre a plus faible de 8,4 jours : les pommiers soumis à la température de l'air plus élevée ont donc une avance de 8 jours environ par rapport aux pommiers placés dans une température de l'air plus basse.

Les résultats de l'analyse de la variance multivariée sont donnés dans la figure 13. Dans la mesure où on n'a qu'un seul degré de liberté pour les effets principaux et pour l'interaction, les différents tests multivariés conduisent aux mêmes résultats. Pour le facteur "température de l'air", la probabilité associée aux tests multivariés est égale à 0,005; pour le facteur "température du sol", cette probabilité est de 0,028; pour l'interaction, elle est de 0,94. On accepte donc l'hypothèse d'absence d'interaction et on constate que les différences liées à la température de l'air sont hautement significatives et que les différences liées à la température du sol sont significatives.

Le tableau 4 donne les coefficients de corrélation intragroupes des variables soumises à l'analyse, M , b et a , d'une part, avec l'unique variable canonique liée au facteur "température de l'air" et, d'autre part, avec l'unique variable canonique liée au facteur "température du sol".

Pour le premier facteur, la variable canonique est essentiellement corrélée au paramètre a . Cet axe peut donc s'interpréter comme un axe lié au retard de la croissance, puisqu'une valeur élevée de a correspond à un retard. Le paramètre a est d'ailleurs la seule variable présentant des différences significatives dues à la température de l'air. La corrélation positive de la variable canonique avec M signifie que la variable canonique est aussi un axe de taille finale. Cette corré-

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall AIR Effect

H = Anova SS&CP Matrix for AIR
E = Anova SS&CP Matrix for CELLULE(AIR*SOL)

S=1 M=0.5 N=2

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.12966287	13.4246	3	6	0.0045
Pillai's Trace	0.87033713	13.4246	3	6	0.0045
Hotelling-Lawley Trace	6.71230825	13.4246	3	6	0.0045
Roy's Greatest Root	6.71230825	13.4246	3	6	0.0045

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall SOL Effect

H = Anova SS&CP Matrix for SOL
E = Anova SS&CP Matrix for CELLULE(AIR*SOL)

S=1 M=0.5 N=2

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.24086567	6.3034	3	6	0.0277
Pillai's Trace	0.75913433	6.3034	3	6	0.0277
Hotelling-Lawley Trace	3.15169162	6.3034	3	6	0.0277
Roy's Greatest Root	3.15169162	6.3034	3	6	0.0277

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall AIR*SOL Effect

H = Anova SS&CP Matrix for AIR*SOL
E = Anova SS&CP Matrix for CELLULE(AIR*SOL)

S=1 M=0.5 N=2

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94318092	0.1205	3	6	0.9447
Pillai's Trace	0.05681908	0.1205	3	6	0.9447
Hotelling-Lawley Trace	0.06024197	0.1205	3	6	0.9447
Roy's Greatest Root	0.06024197	0.1205	3	6	0.9447

Figure 13. Analyse de la variance multivariée des courbes de croissance.

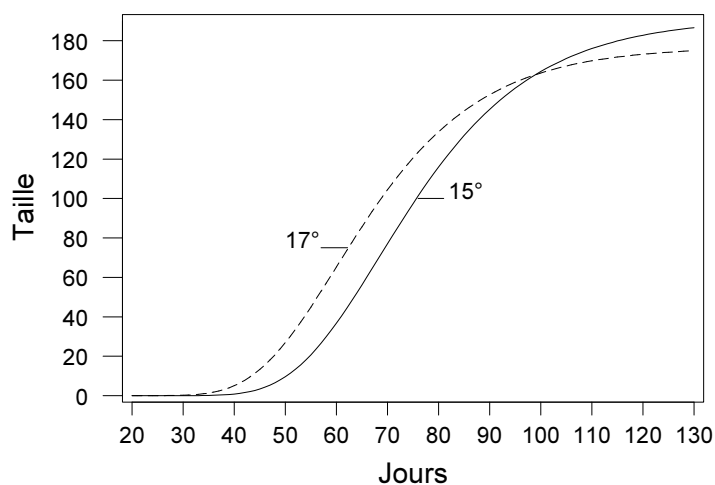


Figure 14. Courbes de croissance moyennes pour les deux niveaux de la température de l'air.

lation est cependant nettement plus faible que la corrélation avec le paramètre a . De même, la corrélation de la variable canonique avec b est assez faible et nous ne l'interpréterons pas. Sur l'axe canonique, la projection du vecteur des moyennes pour la température élevée de l'air se situe donc du côté négatif de l'axe (a faible et M plus petit), et la projection du vecteur des moyennes pour la température plus basse de l'air se situe du côté positif (a grand et M plus grand). Ces projections sont, en effet, respectivement égales à $-1,058$ pour la température de 17°C et à $1,058$ pour la température de 15°C . La figure 14 donne les courbes de croissance moyennes pour les deux niveaux de température de l'air. On constate bien que la différence des deux courbes se marque surtout par le décalage horizontal qui traduit la différence de précocité.

Pour le facteur "température du sol", l'interprétation est un peu plus difficile, car les corrélations sont moins marquées. La corrélation négative de la variable canonique avec M et la corrélation positive de cette variable avec b nous permet d'interpréter l'axe comme étant un axe lié à la lenteur de la croissance et à la faible taille finale. Le point moyen lié à la température du sol de 18°C se situe du côté négatif de l'axe (M grand et b petit), alors que le point moyen lié à la température du sol de 16°C se situe du côté positif de l'axe (M petit et b grand). Les moyennes sur cet axe sont, en effet, égales à $-0,725$ pour la température du sol de 18°C et à $0,725$ pour la température du sol de 16°C . La figure 15 donne les courbes de croissance moyennes pour les deux températures du sol. L'effet du traitement sur le paramètre b n'apparaît pas clairement sur cette figure; par contre l'effet du traitement sur la taille finale est bien marqué.

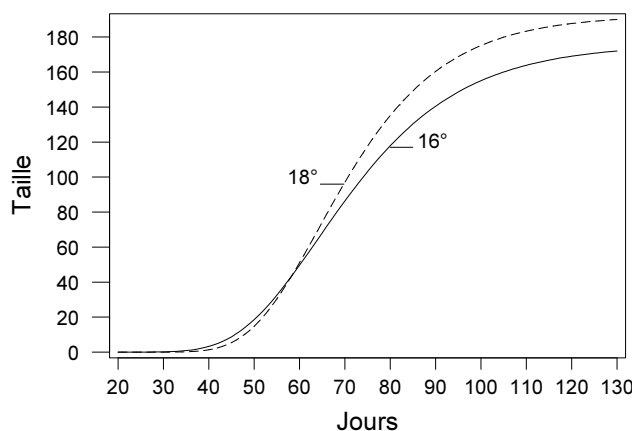


Figure 15. Courbes de croissance moyennes pour les deux niveaux de la température du sol.

Par rapport aux analyses univariées, l'analyse multivariée permet donc de mettre en évidence des différences liées à la température du sol. Ces différences portent simultanément sur deux paramètres des courbes de croissance (M et b) et n'ont pas pu être mises en évidence par les analyses univariées.

6. QUELQUES INFORMATIONS COMPLÉMENTAIRES

6.1. Conditions d'application

Nous avons vu, dans les paragraphes précédents, que la normalité à p dimensions des populations et l'égalité des matrices de variances et covariances de ces populations constituent, avec le caractère aléatoire, simple et indépendant des échantillons, les conditions d'application de l'analyse de la variance multivariée et de l'analyse canonique discriminante.

L'utilisateur peut vérifier ces conditions d'application. Ainsi, lorsque les échantillons de chacune des populations sont de taille suffisante, le caractère multinormal de la population peut être apprécié par l'examen des distances de MAHALANOBIS des individus d'un échantillon au centre de gravité de cet échantillon. En effet, pour des échantillons suffisamment grands, la distribution des carrés des distances est approximativement une distribution χ^2 dont le nombre de degrés de liberté est égal au nombre de variables. Pour vérifier la multinormalité, un test d'ajustement à une distribution χ^2 peut donc être réalisé. Des informations pratiques pour la réalisation de cet ajustement à l'aide du logiciel SAS sont données par KHATTREE et NAIK [1995]. Si le nombre de variables est suffisamment grand, la distribution des distances est approximativement normale [DAGNELIE, 1975]. Cette normalité des distances peut alors être testée par les méthodes uni-

variées classiques, comme par exemple le test de SHAPIRO et WILK, considéré comme un des tests de normalité les plus performants [ROYSTON, 1988].

Pour des populations multinormales, l'égalité des matrices de variances et covariances peut être vérifiée par une généralisation du test de BARTLETT [DAGNELIE, 1975]. Ce test est proposé dans la procédure DISCRIM de SAS [SAS, 1989].

En fait, ces conditions n'interviennent que lorsqu'on réalise l'inférence statistique. La décomposition des matrices de sommes de carrés des écarts et de sommes de produits des écarts, le calcul des valeurs Λ_{obs} , la détermination des variables canoniques, etc., peuvent être réalisés, même lorsque les conditions ci-dessus ne sont pas remplies, les méthodes étant alors utilisées dans un but simplement descriptif. Toutefois, si le non-respect des conditions est très marqué, l'interprétation des résultats des calculs, même sous l'angle descriptif, peut être assez difficile.

Par contre, comme dans le cas univarié, on peut supposer que les méthodes d'inférence statistique sont assez robustes vis-à-vis d'un non-respect modéré des conditions d'application, surtout si les effectifs sont importants et identiques d'un échantillon à l'autre. Des éventuelles transformations de variables peuvent aussi être envisagées dans le but de se rapprocher des conditions d'application.

Enfin, il faut encore signaler que des tests particuliers permettent de comparer deux populations normales de matrices de variances et covariances inégales et qu'il existe des tests non paramétriques multivariés. Des informations à ce sujet sont données par HUBERTY [1994].

6.2. Analyse canonique discriminante et analyse discriminante décisionnelle

Nous avons signalé, dans l'introduction, que l'objectif de l'analyse discriminante décisionnelle est de définir une règle permettant de classer un individu dans un groupe particulier, parmi différents groupes possibles et préalablement définis. Cette règle d'affectation est déterminée à partir de l'analyse d'un échantillon d'individus appartenant à chacun de ces groupes. Une description de l'analyse discriminante décisionnelle est présentée dans une autre note [PALM, 1999].

Lorsqu'à la suite de l'analyse factorielle discriminante, on constate que la première variable canonique permet de prendre en compte l'essentiel des différences entre les groupes, la règle d'affectation d'un individu à un groupe peut être définie en fonction de la première variable canonique, ce qui simplifie la procédure de classement. Dans le cas particulier de deux groupes, cette solution est d'ailleurs identique à la règle de classement définie par l'analyse discriminante linéaire, puisque dans ce cas, les différences entre groupes sont entièrement prises en compte par l'unique variable canonique.

Si on dispose de g groupes, on définit g zones sur le premier axe factoriel, les limites des zones étant situées à mi-chemin entre deux moyennes consécutives de groupes sur la première variable canonique.

Pour l'exemple des iris, les moyennes des trois espèces sur la première variable canonique ont été données dans la figure 6. Sur ce premier axe, le point de séparation de *Iris setosa* et *Iris versicolor* se situe en :

$$(-7,6076 + 1,8250)/2 = -2,8913,$$

et le point de séparation de *Iris versicolor* et *Iris virginica* se situe en :

$$(1,8250 + 5,7825)/2 = 3,8038.$$

Pour classer un iris d'espèce inconnue en supposant que les probabilités d'appartenance *a priori* à l'une des trois espèces soient identiques, il suffit de calculer la valeur de la première variable canonique pour cet iris. Si la valeur obtenue est inférieure ou égale à $-2,8913$, l'iris est classé dans l'espèce *setosa*; si la valeur est comprise entre $-2,8913$ et $3,8038$, l'iris est affecté à l'espèce *versicolor* et si la valeur est supérieure à $3,8038$, l'iris est affecté à l'espèce *virginica*.

Pour cet exemple, l'utilisation de la règle de classement basée sur la première variable canonique conduirait à des résultats très proches de ceux obtenus par l'analyse discriminante linéaire, car la première variable canonique discriminante prend en compte 99,1 % des différences entre les moyennes des groupes (figure 10)

7. BIBLIOGRAPHIE

- DAGNELIE P. [1975]. *Analyse statistique à plusieurs variables*. Gembloux, Presses Agronomiques, 362 p.
- DAGNELIE P. [1998]. *Statistique théorique et appliquée. Tome 2 : inférence statistique à une et à deux dimensions*. Bruxelles, De Boeck et Larcier, 659 p.
- DEBOUCHE C. [1977]. *Application de la régression non linéaire à l'étude et à la comparaison de courbes de croissance longitudinales*. (thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 304 p.
- FISHER R.A. [1936]. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179-188.
- HAND D.J., DALY F., LUNN A.D., MCCONWAYK.J., OSTROWSKI E. [1994]. (eds). *A handbook of small data sets*. London, Chapman and Hall, 458 p.
- HUBERTY C.J. [1994]. *Applied discriminant analysis*. New York, Wiley, 466 p.
- KENDALL M.G., STUART A., ORD J.K. [1983]. *The advanced theory of statistics* (vol. 3). London, Griffin, 780 p.
- KHATTREE R., NAIK D.N. [1995]. *Applied multivariate statistics with SAS software*. Cary, NC, SAS Institute, 396 p.
- LAWLEY D.N. [1959]. Test of significance in canonical analysis. *Biometrika* 46, 59-66.
- PALM R. [1990]. La corrélation canonique : principes et application. *Notes stat. Inform.* (Gembloux) 90/1, 28 p.

- PALM R. [1999]. L'analyse discriminante décisionnelle: principes et application. *Notes stat. Inform.* (Gembloux) 99/4, 41 p.
- RAO C.R. [1952]. *Advanced statistical methods in biometric research*. New York, Wiley, 390 p.
- ROYSTON J.J. [1988]. SHAPIRO-WILK W statistics. In : KOTZ S., JOHNSON N.L. (edit.). *Encyclopedia of statistical sciences* (vol. 8). New York, Wiley, 430-431.
- SAS INSTITUTE INC [1989]. *SAS/STAT User's guide, version 6*, Fourth edition (2 volumes). Cary NC: SAS Institute Inc. 943 + 946 p.
- SEBER G.A.F. [1984]. *Multivariate observations*. New York, Wiley, 686 p.
- TOMASSONE R. [1988]. *Comment interpréter les résultats d'une analyse factorielle discriminante?* Paris, Institut Technique des Céréales et des Fourrages, 56 p.
- X [1994]. *Minitab reference manual, release 10 for Windows*. PA State College, Minitab, 984 p.

ANNEXE

Données relatives aux courbes de croissance des 48 pommiers.

OBS	CELLULE	ARBRE	M	B	A	AIR	SOL
1	1	1	125	15.7	68.1	15	16
2	1	2	173	13.2	70.3	15	16
3	1	3	123	13.3	63.2	15	16
4	1	4	194	16.4	65.5	15	16
5	2	1	168	16.5	75.7	15	16
6	2	2	248	18.4	65.7	15	16
7	2	3	119	22.7	65.2	15	16
8	2	4	195	19.9	62.7	15	16
9	3	1	236	19.6	77.8	15	16
10	3	2	144	23.2	68.0	15	16
11	3	3	229	24.4	70.7	15	16
12	3	4	264	16.2	69.3	15	16
13	1	1	213	12.2	73.1	15	18
14	1	2	158	13.1	64.5	15	18
15	1	3	167	14.4	64.4	15	18
16	1	4	220	15.2	72.3	15	18
17	2	1	200	15.5	66.3	15	18
18	2	2	202	14.9	63.3	15	18
19	2	3	167	18.8	70.1	15	18
20	2	4	210	15.3	66.0	15	18
21	3	1	210	17.9	69.3	15	18
22	3	2	219	16.5	68.2	15	18
23	3	3	179	14.9	68.2	15	18
24	3	4	232	14.8	74.0	15	18
25	1	1	135	17.5	61.9	17	16
26	1	2	166	17.0	57.6	17	16
27	1	3	177	11.6	53.7	17	16
28	1	4	175	13.3	59.2	17	16
29	2	1	146	21.3	68.4	17	16
30	2	2	155	18.9	59.1	17	16

Données relatives aux courbes de croissance des 48 pommiers (suite).

OBS	CELLULE	ARBRE	M	B	A	AIR	SOL
31	2	3	143	17.0	55.8	17	16
32	2	4	207	17.8	59.0	17	16
33	3	1	137	15.3	60.3	17	16
34	3	2	235	15.0	66.5	17	16
35	3	3	127	18.5	58.1	17	16
36	3	4	202	14.4	56.4	17	16
37	1	1	237	16.9	62.5	17	18
38	1	2	189	13.1	56.0	17	18
39	1	3	187	17.2	72.0	17	18
40	1	4	166	13.9	63.3	17	18
41	2	1	150	12.4	58.0	17	18
42	2	2	182	12.3	50.5	17	18
43	2	3	233	18.2	62.7	17	18
44	2	4	185	17.9	58.0	17	18
45	3	1	172	14.3	55.5	17	18
46	3	2	153	18.0	64.4	17	18
47	3	3	189	10.6	60.3	17	18
48	3	4	200	14.9	59.7	17	18

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, 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:

PALM R. [1996]. La classification numérique : principes et application. *Notes Stat. Inform.* (Gembloux) 96/1, 28 p.

PALM R. [1996]. Cartes de contrôle : les cartes de SHEWHART. *Notes Stat. Inform.* (Gembloux) 96/2, 41 p.

PALM [1996]. Cartes de contrôle : combinaison de résultats, données corrélées ou multivariées. *Notes Stat. Inform.* (Gembloux) 96/3, 37 p.

PRÉVOT H. [1997]. Introduction au système d'exploitation WINDOWS NT. *Notes Stat. Inform.* (Gembloux) 97/1, 35 p.

VAN BELLE L., CLAUSTRIAUX J.J. [1997]. Introduction à l'analyse des données par le logiciel Minitab sous Windows. *Notes Stat. Inform.* (Gembloux) 97/2, 22 p.

PRÉVOT H. [1998]. Les outils de base du réseau Internet. *Notes Stat. Inform.* (Gembloux) 98/1, 24 p.

PALM [1998]. L'analyse en composantes principales : principes et applications. *Notes Stat. Inform.* (Gembloux) 98/2, 31 p.

BROSTAUX Y. [1999]. Introduction au système d'exploitation Unix. *Notes Stat. Inform.* (Gembloux) 99/1, 15 p.

CLAUSTRIAUX J.J., IEMMA A.F. [1999]. A propos des qualificatifs complet, orthogonal et équilibré en analyse de la variance. *Notes Stat. Inform.* (Gembloux) 99/2, 14 p.

IEMMA A.F., CLAUSTRIAUX J.J. [1999]. Etude des hypothèses de l'analyse de la variance à deux critères de classification : approche par l'exemple. *Notes Stat. Inform.* (Gembloux) 99/3, 14 p.

PALM R. [1999]. L'analyse discriminante décisionnelle : principes et application. *Notes Stat. Inform.* (Gembloux) 99/4, 41 p.

PALM R. [1999]. Indices d'aptitude des procédés de production. *Notes Stat. Inform.* (Gembloux) 99/5, 26 p.

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

D/2000/2371/1