

REVUE DE STATISTIQUE APPLIQUÉE

T. FOUCART

Analyse de la collinéarité classification de variables

Revue de statistique appliquée, tome 44, n° 4 (1996), p. 41-57

[<http://www.numdam.org/item?id=RSA_1996__44_4_41_0>](http://www.numdam.org/item?id=RSA_1996__44_4_41_0)

© Société française de statistique, 1996, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ANALYSE DE LA COLLINÉARITÉ CLASSIFICATION DE VARIABLES

T. Foucart

Département de Mathématiques

Université de Poitiers

Avenue du Recteur Pineau

86000 Poitiers

E-mail : foucart@mathrs.univ-poitiers.fr

RÉSUMÉ

Nous étudions dans cet article la collinéarité statistique entre des variables statistiques, et distinguons deux notions de collinéarité lorsque les variables sont réparties dans plusieurs groupes : la collinéarité inter d'une part et la collinéarité intra d'autre part. Nous en déduisons des algorithmes de classification caractérisant la collinéarité entre ces variables, et nous les appliquons à des données simulées pour en montrer l'intérêt par rapport aux méthodes classiques d'analyse de données.

Un programme permettant d'appliquer les méthodes présentées est disponible auprès de l'auteur.

Mots-clés : *Régression, modèle linéaire, collinéarité, facteur d'inflation, multicollinéarité, corrélation canonique, analyse en composantes principales.*

ABSTRACT

We present in this paper a new approach to the statistical collinearity between variables and distinguish two notions when the variables are dealed in several groups : first collinearity between and second collinearity within variables. Then we propose two algorithms in order to classify these variables in according to these collinearities, and perform them on simulated data to exhibit their interest versus principal classical data analysis methods.

A program to perform the presented methods is available from the author.

Keywords : *Regression, linear model, collinearity, inflation factor, multicollinearity, canonical correlation, principal component analysis.*

1. Analyse de la multicollinéarité

Nous analysons dans ce paragraphe la multicollinéarité statistique entre des variables statistiques et la décomposons en multicollinéarité inter et multicollinéarité intra de façon analogue à l'analyse de variance.

1.1. Mesure de la collinéarité

La collinéarité statistique entre des variables statistiques est caractérisée par l'existence d'une relation linéaire approximative entre ces variables. Pour mesurer cette collinéarité, il existe différents indices tels que les facteurs d'inflation, l'indice de multicollinéarité ou l'indice de conditionnement généralisé. Nous utilisons les deux premiers.

Soit $x_1, \dots, x_p$ les variables statistiques considérées. La relation linéaire n'étant qu'approximative, la matrice de corrélation $\mathbf{R}$ des variables est symétrique définie positive : ses valeurs propres sont toutes non nulles, et nous supposons qu'elle est inversible sans difficulté numérique.

On peut alors effectuer la régression d'une des variables considérées, notée x_j , par les autres : on obtient un coefficient de détermination R_j^2 d'autant plus proche de 1 que la variable x_j est proche d'une combinaison linéaire des $x_k, k \neq j$. Le terme de la forme $1/(1 - R_j^2)$ est appelé facteur d'inflation de la variable x_j et noté f_j (Chatterjee et Price, 1977). Ces facteurs d'inflation constituent un ensemble de mesures de la collinéarité et l'on peut les résumer par leur moyenne, appelée indice de multicollinéarité I (Tomassone *et alii.*, 1992). On montre en outre que l'indice de multicollinéarité est égal à la moyenne des inverses des valeurs propres λ_j de la matrice de corrélation, toutes non nulles puisque la matrice $\mathbf{R}$ est supposée inversible (Hoerl et Kennard, 1970).

$$f_j = \frac{1}{1 - R_j^2} \quad I = (1/p) \sum_{j=1}^p f_j = (1/p) \sum_{j=1}^p 1/\lambda_j$$

Exemple de collinéarité statistique

On dit souvent que les variables explicatives x_j sont statistiquement collinéaires – plus simplement ici collinéaires – lorsque le coefficient de corrélation de deux d'entre elles au moins est élevé, de l'ordre de 0,99 par exemple. En réalité, il est connu que cette appréciation est très restrictive (Belsley *et alii.*, 1980), et il existe des algorithmes permettant de construire des variables collinéaires statistiquement dont les coefficients de corrélation sont faibles (Foucart, 1991 et 1992).

Considérons par exemple la matrice de corrélation ci-dessous :

	x_1	x_2	x_3	x_4
x_1	1			
x_2	0,6	1		
x_3	-0,279	0,6	1	
x_4	0,0446	0	0	1

La régression de x_4 par x_1, x_2, x_3 donne un coefficient de détermination égal à 0,995358. Il existe donc une collinéarité statistique entre les variables x_1, x_2, x_3 et

x_4 bien que les coefficients de corrélation entre x_4 et les x_j soient nuls ou très faibles et que les autres ne soient pas très élevés.

Nous obtenons des facteurs d'inflation et un indice de multicollinéarité très élevés :

Facteurs d'inflation :	
x_1	107787,6583
x_2	154973,3630
x_3	107452,6475
x_4	215,4069
Indice de multicollinéarité :	123476,3586

Cette collinéarité apparaît aussi dans la plus petite valeur propre de la matrice égale à $2,7 \times 10^{-6}$ (le calcul a été effectué en double précision). On détecte donc bien la collinéarité statistique entre les variables x_1, x_2, x_3 et x_4 , qui n'apparaît pas de façon évidente dans la matrice initiale malgré sa petite taille et qui rend la régression très instable : en fixant $r_{1,2}$ à 0,59 au lieu de 0,6, la régression de x_4 par x_1, x_2 et x_3 donne un coefficient de détermination de 0,077.

Dans le cas de données collinéaires, il est fréquent que tous les coefficients R_j^2 soient très proches de 1 et que tous les termes de la matrice R^{-1} prennent des valeurs très élevées (Belsley *et alii.*, 1980, p. 186, Tomassone *et alii.*, 1992, p. 154) même lorsque seules deux variables sont liées; il est alors difficile de déterminer ces dernières (Erkel-Rousse, 1995).

1.2. Multicollinéarités inter et intra

Supposons maintenant que les variables x_j soient classées en deux groupes G_1 et G_2 d'effectifs respectifs p_1 et p_2 ; la matrice de corrélation $\mathbf{R}$ se met sous la forme :

$$R = \begin{pmatrix} R_{1,1} & R_{1,2} \\ R_{2,1} & R_{2,2} \end{pmatrix}$$

où $R_{1,1}$ et $R_{2,2}$ sont les matrices de corrélation des groupes G_1 et G_2 , et où $R_{1,2}$ est la matrice constituée des coefficients de corrélation des variables du groupe G_1 avec les variables du groupe G_2 .

Etudions tout d'abord les indices de multicollinéarité I_1 et I_2 des groupes G_1 et G_2 :

$$I_1 = \frac{1}{p_1} \sum_{j_1 \in G_1} \frac{1}{1 - R_{j_1}(1)^2} \quad I_2 = \frac{1}{p_2} \sum_{j_2 \in G_2} \frac{1}{1 - R_{j_2}(2)^2}$$

où $R_{j_1}(1)^2$ et $R_{j_2}(2)^2$ sont les coefficients de détermination dans la régression de la variable x_{j_1} par les autres variables du groupe G_1 et de la variable x_{j_2} par les autres variables du groupe G_2 .

On a :

$$p_1 I_1 + p_2 I_2 = \sum_{j_1 \in G_1} \frac{1}{1 - R_{j_1}(1)^2} + \sum_{j_2 \in G_2} \frac{1}{1 - R_{j_2}(2)^2}$$

Considérons une variable x_{j_1} du groupe G_1 : le coefficient de détermination $R_{j_1}(1)^2$ est inférieur ou égal au coefficient de détermination $R_{j_1}^2$ calculé par la régression de la variable x_{j_1} en considérant comme variables explicatives non seulement les variables du groupe G_1 , mais aussi celles du groupe G_2 . On a donc :

$$\forall j_1 \in G_1 \quad \frac{1}{1 - R_{j_1}(1)^2} \leq \frac{1}{1 - R_{j_1}^2}$$

Le même raisonnement peut être tenu pour les variables du groupe G_2 , et l'on obtient finalement :

$$p_1 I_1 + p_2 I_2 \leq \sum_{j=1}^p \frac{1}{1 - R_j^2}$$

Soit

$$\frac{p_1 I_1 + p_2 I_2}{p} \leq I \quad (\text{propriété 1})$$

D'où les définitions :

– nous appelons indice de multicollinéarité totale I l'indice de multicollinéarité de la totalité des variables.

– nous appelons indice de multicollinéarité intra I_W la moyenne des indices de multicollinéarité des groupes G_1 et G_2 pondérés par leurs effectifs.

– nous appelons indice de multicollinéarité inter I_B la différence entre l'indice de multicollinéarité total et l'indice de multicollinéarité intra.

L'indice de multicollinéarité inter est nul si et seulement si les coefficients de corrélation linéaire des variables du groupe G_1 avec les variables du groupe G_2 sont tous nuls; la condition suffisante étant évidente, supposons les égalités suivantes :

$$\begin{aligned} \forall j_1 \in G_1 \quad R_{j_1}^2(1) &= R_{j_1}^2 \\ \forall j_2 \in G_2 \quad R_{j_2}^2(2) &= R_{j_2}^2 \end{aligned}$$

Tous les coefficients de corrélation partielle entre les variables de G_1 et de G_2 sont nuls, sinon les relations ci-dessus ne seraient pas toutes vérifiées; la matrice inverse de la matrice de corrélation $\mathbf{R}$, dont les termes sont les covariances partielles de deux variables conditionnellement aux autres, s'écrit :

$$R^{-1} = \begin{pmatrix} R_{1,1}^{-1} & 0 \\ 0 & R_{2,2}^{-1} \end{pmatrix}$$

En recalculant l'inverse de R^{-1} , on obtient évidemment la matrice de corrélation R et la propriété cherchée.

L'indice de multicollinéarité inter est nul si les variables du groupe G_1 sont non corrélées aux variables du groupe G_2 , ce qui est équivalent à la nullité des coefficients de corrélation canonique; il est lié, d'une façon que nous n'avons pas déterminée dans le cas général, aux coefficients de corrélation canonique et caractérise la liaison entre les deux groupes de variables.

On peut définir, par analogie au rapport de corrélation, le rapport de multicollinéarité, noté μ^2 :

$$\mu^2 = I_B/I$$

Sachant que l'indice de multicollinéarité intra I_W est supérieur ou égal à 1, on a :

$$0 \leq \mu^2 \leq (I - 1)/I$$

Toutes ces propriétés et définitions se généralisent sans difficulté au cas d'un nombre quelconque de groupes de variables, de façon analogue à l'analyse de variance. Cette généralisation consiste à comparer plusieurs ensembles de variables observées sur un même ensemble d'individus.

2. Algorithmes de classification

Les notions de multicollinéarités inter et intra introduites dans le paragraphe précédent peuvent être utilisées pour classer des variables statistiques; nous proposons deux algorithmes de classification.

2.1. Minimisation de la multicollinéarité intra

Le premier algorithme permet d'obtenir rapidement un minimum local de la multicollinéarité intra.

Précisons tout de suite qu'il est mathématiquement possible de déterminer, pour un nombre de groupes fixé, le minimum absolu puisque le nombre de variables explicatives est fini. Supposons par exemple que l'on classe les variables d'un ensemble G dans 2 ensembles non vides G_1 et G_2 : le nombre de cas à considérer est égal au nombre de sous-ensembles de la forme (G_1, G_2) , dont on extrait les deux couples (Φ, G) et (G, Φ) et que l'on divise par 2 puisque le couple (G_1, G_2) est équivalent au couple (G_2, G_1) . On obtient ainsi $2^{p-1} - 1$ cas possibles. Pour $p = 10$, cela donne 511 couples à étudier.

Pour trois groupes, le calcul est un peu plus complexe : il y a C_p^k façons de choisir un groupe d'effectif k , et, parmi les $p - k$ variables restantes, il y a $2^{p-k} - 2$ façons de choisir deux autres groupes (avec répétition). Le nombre k varie de 1 à $p - 2$, et il faut diviser par 3! pour tenir compte des répétitions; on obtient alors le

nombre de façons de choisir trois groupes parmi p variables :

$$1/6 \sum_{k=1}^{p-2} c_p^k (2^{p-k} - 2)$$

Pour $p = 10$, on trouve 9 330, pour $p = 15$, 2 375 101, plus de 5×10^8 pour $p = 20$. D'où l'intérêt d'obtenir un minimum local, lorsque le résultat est établi rapidement. Le problème est analogue à celui de la recherche du système de prédicteurs minimisant l'estimation sans biais de la variance résiduelle (Furnival et Wilson, 1974).

Le nombre de groupes étant fixé, on cherche, à chaque pas de l'algorithme, à transférer une variable d'un groupe à un autre de façon à minimiser l'indice de multicollinéarité intra.

On dispose d'un groupe initial : l'ensemble des variables explicatives. On crée un deuxième groupe en séparant celle dont l'éviction du groupe initial en diminue le plus l'indice de multicollinéarité.

Les étapes suivantes consistent à étudier, pour chaque variable, l'indice de multicollinéarité intra lorsqu'on la transfère d'un groupe à un autre, déjà existant ou nouvellement créé, et à effectuer le transfert de celle qui correspond à la diminution maximum de l'indice de multicollinéarité intra.

On crée éventuellement de nouveaux groupes tant que leur nombre n'est pas égal au nombre choisi; la propriété 1 du paragraphe 1.2 montre que le nombre de groupes fixé au départ est atteint, la multicollinéarité intra étant minimum et égale à 1 pour un nombre de groupes égal au nombre de variables.

2.2. Maximisation de la multicollinéarité intra

Dans le paragraphe 2.1, notre objectif était de construire des groupes dont les variables sont relativement peu liées entre elles, la multicollinéarité apparaissant dans les corrélations entre les groupes.

Nous proposons maintenant la démarche inverse : nous recherchons des groupes dont les variables sont fortement liées, et dont la multicollinéarité inter est faible.

La propriété 1 du paragraphe 1.3 met en évidence un inconvénient de l'algorithme : on maximise la multicollinéarité intra en rassemblant les variables, et, si l'on n'impose pas de contrainte, on obtient un groupe de taille élevé et des groupes réduits à une seule variable.

D'où l'idée d'imposer une contrainte, en fixant la taille maximale des groupes.

L'algorithme est alors le suivant :

- on dispose initialement de p groupes réduits chacun à une variable.
- on fixe le nombre de groupes voulu.
- on impose un nombre maximal de variables par groupes.

- on commence par rassembler deux variables de façon à maximiser l'indice de multicollinéarité intra; cela revient à réunir les variables dont le coefficient de corrélation est maximum.
- on réunit ensuite deux groupes choisis parmi les $p - 1$ groupes obtenus précédemment, de façon à maximiser l'indice de multicollinéarité intra, en respectant la contrainte de taille maximale..
- et ainsi de suite, jusqu'à l'obtention du nombre de groupes fixé au départ.

3. Application

Nous considérons un vecteur aléatoire gaussien $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$ dont la structure est définie *a priori* par les trois groupes de variables suivants : $G_1 = \{x_1, x_8, x_9\}$, $G_2 = \{x_2, x_3, x_5\}$, $G_3 = \{x_4, x_6, x_7\}$ orthogonaux deux à deux (les coefficients de corrélation de deux variables appartenant à des groupes différents sont nuls) et dont les matrices de corrélation sont égales à la matrice ci-dessous :

TABLEAU 1
Matrice de corrélation théorique de chaque groupe

1,0000		
0,5000	1,0000	
0,5000	-0,4990	1,0000

Les valeurs propres, toutes d'ordre de multiplicité égal à 3, sont égales à 1,5003; 1,4990; 0,0007. Les vecteurs principaux de la matrice de corrélation théorique complète du vecteur $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$ ne sont pas uniques et sont définis à une rotation près dans les sous-espace principaux. Pour comparer les classifications obtenues par les algorithmes précédents aux résultats donnés par l'analyse en composantes principales et par une CAH, nous avons simulé un échantillon gaussien de taille 100 du vecteur précédent. La matrice des corrélations observées est donnée ci-dessous :

TABLEAU 2
Matrice des corrélations observées des variables

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1,0000								
x_2	0,0913	1,0000							
x_3	0,1149	0,5839	1,0000						
x_4	-0,0034	-0,0046	0,0067	1,0000					
x_5	-0,0227	0,4392	-0,4720	-0,0159	1,0000				
x_6	-0,0945	-0,0222	0,0658	0,4618	-0,0983	1,0000			
x_7	0,0803	0,0206	-0,0531	0,5913	0,0787	-0,4411	1,0000		
x_8	0,5148	0,0715	0,0535	0,0249	0,0204	0,0340	-0,0093	1,0000	
x_9	0,6308	0,0355	0,0785	-0,0246	-0,0445	-0,1306	0,0946	-0,3394	1,0000

3.1. Analyse en composantes principales des variables centrées réduites

Nous nous limitons aux résultats de l'analyse en composantes principales concernant les variables. Les résultats sont les suivants :

TABLEAU 3
Valeurs propres de la matrice de corrélation

Val.pro	% exp	% cum			
λ_1	1,853	21	21	/	*****
λ_2	1,641	18	39	/	*****
λ_3	1,566	17	56	/	*****
λ_4	1,458	16	72	/	*****
λ_5	1,342	15	87	/	*****
λ_6	1,139	13	100	/	*****
λ_7	0,001	0	100	/	
λ_8	0,001	0	100	/	
λ_9	0,001	0	100	/	

Les trois dernières valeurs propres (tableau 3), proches de 0, montrent qu'il existe une forte collinéarité entre les variables, caractérisée par les trois vecteurs principaux associés.

TABLEAU 4
Vecteurs principaux de la matrice de corrélation

$u_1/$	0,6144	0,3573	0,3573	0,0343	-0,0047	-0,2101	0,2258	0,2269	0,4684
$u_2/$	0,0186	-0,2116	-0,5125	0,2786	0,3364	-0,3407	0,5925	-0,1277	0,1343
$u_3/$	0,0767	-0,0972	-0,2232	-0,7312	0,1446	-0,4679	-0,3140	-0,1376	0,2069
$u_4/$	0,1837	-0,6162	-0,0314	0,0742	-0,6333	0,0831	-0,0046	-0,1860	0,3707
$u_5/$	-0,3957	0,2856	0,3607	0,0200	-0,0913	-0,1620	0,1709	-0,7212	0,2180
$u_6/$	-0,1658	-0,1031	0,2621	-0,1459	-0,4035	-0,5459	0,3452	0,2961	-0,4519
$u_7/$	-0,0315	0,0270	-0,0253	0,5981	-0,0204	-0,5372	-0,5913	0,0237	0,0286
$u_8/$	0,3605	-0,4806	0,4903	0,0388	0,4426	-0,0325	-0,0367	-0,2992	-0,3307
$u_9/$	-0,5172	-0,3385	0,3450	-0,0092	0,3119	0,0073	0,0104	0,4253	0,4695

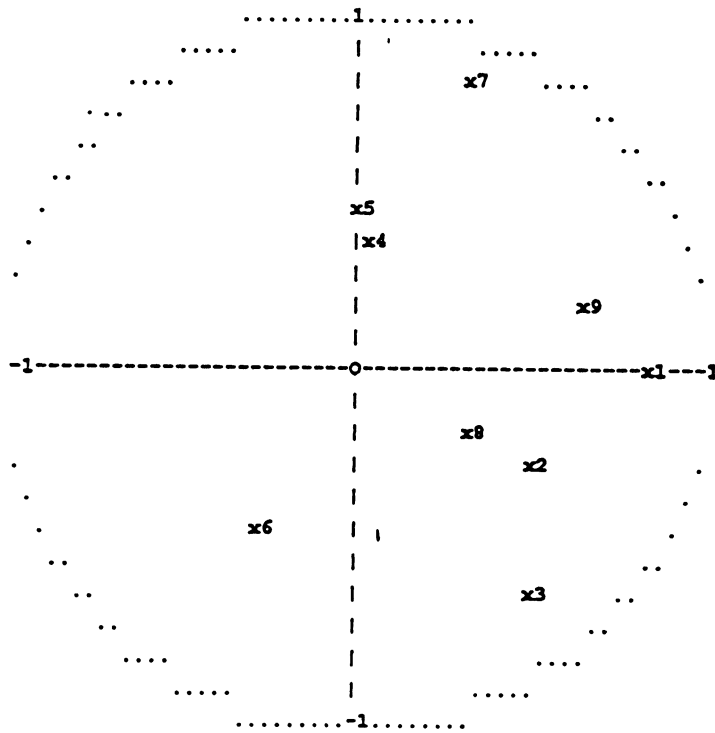
Le vecteur propre u_7 (tableau 4) possède trois coordonnées élevées par rapport aux autres : u_7^4 , u_7^6 , u_7^7 , ce qui indique une liaison entre les variables x_4 , x_6 et x_7 . Il est difficile de déduire de l'étude des vecteurs u_8 et u_9 des conclusions analogues.

Former des groupes de variables à partir des coordonnées des variables sur les composantes principales (tableau 5 ci-dessous) n'est guère facile : la première composante principale est fortement liée aux variables x_1 et x_9 , la deuxième à x_7 et x_3 , la troisième à x_4 , la quatrième à x_2 et x_5 , la cinquième à x_8 , la sixième (que nous ne donnons pas faute de place) à x_6 . Les autres composantes principales sont de variance très faible et ne représentent pas bien les variables.

On distingue bien sur le cercle de corrélation 1×2 ci-dessus la forte corrélation entre x_1 et x_9 (0,6308), entre x_3 et x_5 (-0,472), mais on ne distingue facilement pas les liaisons entre trois variables à la fois, celle par exemple entre x_4 , x_6 et x_7 que nous avons détectée en étudiant le vecteur u_7 .

TABLEAU 5
Corrélations des variables et des composantes principales

	Axe 1		Axe 2		Axe 3		Axe 4		Axe 5	
	coord.	r^2	coord.	r^2	coord.	r^2	coord.	r^2	coord.	r^2
x_1	0,836	0,699	0,024	0,001	0,096	0,009	0,222	0,049	-0,458	0,210
x_2	0,486	0,236	-0,271	0,073	-0,122	0,015	-0,744	0,554	0,331	0,109
x_3	0,486	0,237	-0,656	0,431	-0,279	0,078	-0,038	0,001	0,418	0,175
x_4	0,047	0,002	0,357	0,127	-0,915	0,837	0,090	0,008	0,023	0,001
x_5	-0,006	0,000	0,431	0,186	0,181	0,033	-0,765	0,585	-0,106	0,011
x_6	-0,286	0,082	-0,436	0,190	-0,586	0,343	0,100	0,010	-0,188	0,035
x_7	0,307	0,094	0,759	0,576	-0,393	0,154	-0,006	0,000	0,198	0,039
x_8	0,309	0,095	-0,164	0,027	-0,172	0,030	-0,225	0,050	-0,835	0,698
x_9	0,638	0,407	0,172	0,030	0,259	0,067	0,448	0,200	0,253	0,064



Cercle de corrélation 1×2 ($\lambda_1 = 1,853$, $\lambda_2 = 1,641$)

Ces difficultés dans l'interprétation des résultats sont dues en partie à l'ordre de multiplicité des valeurs propres théoriques; les sous-espaces propres sont en effet de dimension 3, et les vecteurs principaux associés à une même valeur propre sont définis à une rotation près dans le sous-espace propre correspondant. En outre, les deux premières valeurs propres théoriques sont très proches l'une de l'autre : un

échantillon de taille 100 ne permet pas de les distinguer et nous avons en réalité une valeur propre égale 1,5 environ dont l'ordre de multiplicité est égal à 6.

3.2. Indices de similarité et de dissimilarité

Nous avons recherché la structure des données en étudiant des indices de dissimilarité ou de similarité entre les variables.

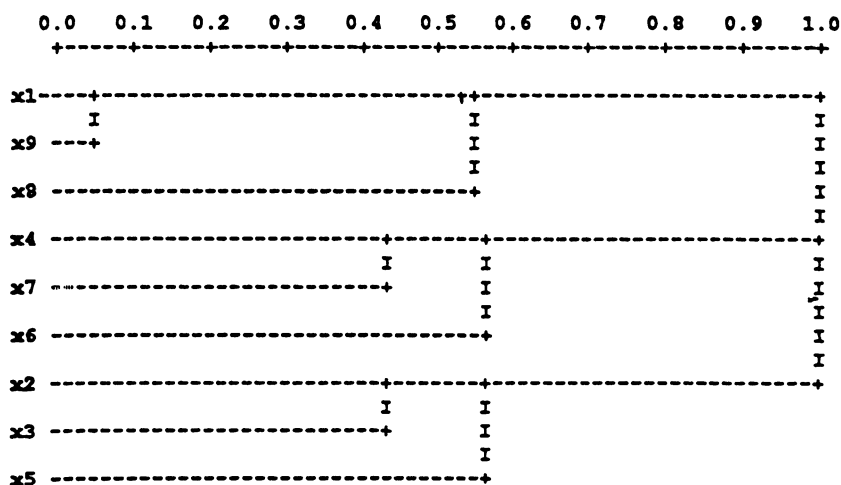
Dans le cas d'un indice de dissimilarité, nous avons tout d'abord calculé la matrice de la forme bilinéaire symétrique que l'on peut en déduire (matrice de Torgerson, F. Cailliez et J.-P. Pagès, 1976, p. 251). Par l'analyse factorielle de cette matrice, nous avons ensuite calculé les coordonnées des variables sur les axes principaux associés aux valeurs propres positives, et enfin procédé à la représentation graphique des plans principaux ou à la classification ascendante hiérarchique par la méthode de Ward (minimisation de l'inertie intra par rapport à l'inertie totale).

Dans le cas d'un indice de similarité, la matrice des similarités a été directement traitée comme une matrice de produit scalaire, ce qu'elle est d'ailleurs dans le cas de l'indice de similarité défini par le carré du coefficient de corrélation (produits scalaires entre les opérateurs de projection orthogonale associés à chaque variable, Foucart, 1984, p. 25).

Les indices de similarité $s_{i,j}$ ou de dissimilarité $d_{i,j}$ ci-dessous, qui font intervenir la valeur absolue ou le carré des coefficients de corrélation, permettent de classer les variables conformément à leur structure.

$$d_{i,j} = 1 - |r_{i,j}| \quad d_{i,j} = [1 - r_{i,j}^2] \quad s_{i,j} = r_{i,j}^2$$

Nous donnons ci-dessous l'arbre obtenu en utilisant comme indice de similarité entre deux variables le carré de leur coefficient de corrélation linéaire :

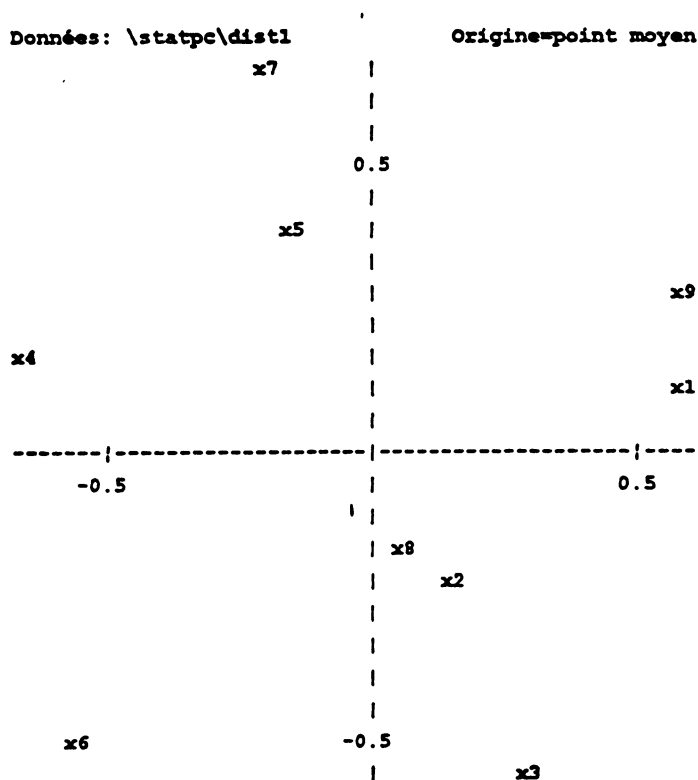


Les trois groupes obtenus par troncature au niveau 0,6 sont bien ceux de la structure définie *a priori*.

Par contre, les indices de dissimilarité ci-dessous donnent de mauvais résultats :

$$d_{i,j} = [2(1 - r_{i,j})]^{1/2} \quad (1) \quad d_{i,j} = [2(1 - r_{i,j})] \quad (2)$$

Le premier indice est une distance euclidienne dont les produits scalaires sont les coefficients de corrélation. On note que les distances les plus faibles, $d_{1,8}$ (0,985) et $d_{1,9}$ (0,859) par exemple, concernent des variables d'un même groupe, corrélées positivement. Mais les distances les plus grandes sont les distances $d_{8,9}$ (1,637), $d_{3,5}$ (1,716), $d_{6,7}$ (1,698) calculées à partir de coefficients de corrélation négatifs : elles concernent des variables faisant partie d'un même groupe, et il est vain d'espérer reconstruire la structure à l'aide de cet indice. Il n'est donc pas étonnant que ces indices donnent des résultats non conformes à la structure. Le graphique ci-dessous donne le plan 1×2 obtenu par l'analyse factorielle du tableau de dissimilarités définies par la formule (1) :



Plan 1 × 2 ($\lambda_1 = 0,1872$) ($\lambda_2 = 0,1822$)

3.3. Classification des variables par maximisation de la multicollinéarité intra

Nous appliquons tout d'abord l'algorithme de maximisation de la multicollinéarité intra, en choisissant un nombre de groupes égal au nombre de faibles valeurs propres, donc à 3, et en limitant le nombre de variables par groupes à 3. On obtient les résultats suivants :

TABLEAU 6
Classification par maximisation de la multicollinéarité intra

Groupe 1 : x_1, x_9, x_8	Groupe 2 : x_2, x_3, x_5	Groupe 3 : x_4, x_7, x_6
Matrice de corrélation :	Matrice de corrélation :	Matrice de corrélation :
1,000	1,000	1,000
0,631 1,000	0,584 1,000	0,591 1,000
0,515-0,339 1,000	0,439-0,472 1,000	0,462-0,441 1,000
Valeurs propres :	Valeurs propres :	Valeurs propres :
1,670285, 1,329103,	1,586231, 1,413198,	1,592132, 1,407101,
0,000585	0,000627	0,000744
multicollinéarité	multicollinéarité	multicollinéarité
570,5	532,4	448,2

En écrivant la matrice de corrélation en réordonnant les variables suivant les groupes obtenus (tableau 7 ci-dessous), on voit apparaître trois groupes de trois variables dont les matrices de corrélation ne possèdent qu'une seule valeur propre presque nulle, présentant chacun une très forte multicollinéarité (l'indice de multicollinéarité est de l'ordre de 500) due à une liaison entre les trois variables. Nous avons donc une combinaison linéaire de x_1, x_9, x_8 , une deuxième combinaison linéaire de x_2, x_3, x_5 et une troisième de x_4, x_7 et x_6 – celle que nous avons déjà détectée en étudiant u_7 – toutes de très faibles variances, comparables aux trois dernières valeurs propres de la matrice de corrélation ($\lambda_7 = 0,0007255$, $\lambda_8 = 0,0006487$, $\lambda_9 = 0,0005448$). Ces groupes correspondent exactement à la structure théorique des données.

La classification des variables obtenue par maximisation de la multicollinéarité intra est visiblement très satisfaisante : les coefficients de corrélation entre les variables de deux groupes différents sont presque tous inférieurs à 0,1 en valeur absolue et beaucoup plus faibles que les autres. La multicollinéarité apparaît au sein de chaque groupe, et le rapport de multicollinéarité est très faible :

TABLEAU 7
Matrice de corrélation après classification par maximisation

	x_1	x_9	x_8	x_2	x_3	x_5	x_4	x_7	x_6
x_1	1,0000								
x_9	0,6308	1,0000							
x_8	0,5148	-0,3394	1,0000						
x_2	0,0913	0,0355	0,0715	1,0000					
x_3	0,1149	0,0785	0,0535	0,5839	1,0000				
x_5	-0,0227	-0,0445	0,0204	0,4392	-0,4720	1,0000			
x_4	-0,0034	-0,0246	0,0249	-0,0046	0,0067	-0,0159	1,0000		
x_7	0,0803	0,0946	-0,0093	0,0206	-0,0531	0,0787	0,5913	1,0000	
x_6	-0,0945	-0,1306	0,0340	-0,0222	0,0658	-0,0983	0,4618	-0,4411	1,0000

TABLEAU 8
Analyse de la multicollinéarité

Multicollinéarité			
totale	inter	intra	rapport
528,8632	11,8167	517,0465	2,234358E-02

Compte tenu de la remarque illustrée par l'exemple donné dans le paragraphe 1.1, l'absence de collinéarité inter doit être vérifiée par le calcul des coefficients de corrélation canonique :

TABLEAU 9
Corrélations canoniques entre les trois groupes de variables

Coefficients de corrélation canonique		
entre G_1 et G_2	entre G_1 et G_3	entre G_2 et G_3
0,172; 0,071; 0,0427	0,167; 0,080; 0,0196	0,164; 0,076; 0,0585

On peut alors effectuer l'ACP de chacun de ces sous-espaces : en se limitant aux deux premières composantes principales de chaque groupe, presque orthogonales deux à deux et qui expliquent la quasi-totalité de l'inertie (les valeurs propres sont données dans le tableau 6), on dispose d'un ensemble de représentations graphiques beaucoup plus simples à interpréter que celles qui sont fournies par l'ACP de l'ensemble des variables.

3.4. Classification des variables par minimisation de la multicollinéarité intra

Nous choisissons encore comme nombre de groupes le nombre de valeurs propres très faibles, mais il n'est pas nécessaire ici de fixer une limite au nombre de variables dans chaque groupe. L'algorithme donne la classification suivante :

TABLEAU 10
Classification par minimisation de la multicollinéarité intra

Groupe 1 : x_2, x_7, x_9	Groupe 2 : x_1, x_4, x_5	Groupe 3 : x_3, x_6, x_8
Matrice de corrélation :	Matrice de corrélation :	Matrice de corrélation :
1,000	1,000	1,000
0,021 1,000	- 0,003 1,000	0,066 1,000
0,036 0,095 1,000	- 0,023-0,016 1,000	0,054 0,034 1,000
Valeurs propres :	Valeurs propres :	Valeurs propres :
1,109115; 0,986778;	1,026169; 1,003223;	1,103310; 0,967030;
0,904106	0,970607	0,929660
multicollinéarité	multicollinéarité	multicollinéarité
1,0070	1,0005	1,0054

Examinons le groupe 1 par exemple, constitué des variables x_2, x_7 et x_9 : la variable x_2 est extraite du groupe 2 obtenue par maximisation de la multicollinéarité intra (cf. paragraphe précédent), la variable x_7 du groupe 3, et la variable x_9 du groupe 1. Il en est de même pour les autres groupes. Les groupes obtenus sont constitués de variables dont la collinéarité est très faible; la multicollinéarité totale apparaît, à la suite de cette classification, dans les liaisons entre les groupes, comme l'indique la décomposition de l'indice :

TABLEAU 11
Analyse de la multicollinéarité

Multicollinéarité			
totale	inter	intra	rapport
528,8632	527,8588	1,004308	0,998101

Les coefficients de corrélation canonique sont donnés dans le tableau 12 ci-dessous :

TABLEAU 12
Corrélations canoniques entre les trois groupes de variables

Coefficients de corrélation canonique		
entre G_1 et G_2	entre G_1 et G_3	entre G_2 et G_3
0,636; 0,609; 0,434	0,592; 0,459; 0,335	0,552; 0,501; 0,416

Les trois sous-espaces ne sont donc pas orthogonaux; l'ACP de chaque groupe de variables est bien sûr possible, mais les représentations graphiques obtenues ne sont pas indépendantes l'une de l'autre et d'ailleurs ne présentent guère d'intérêt compte tenu des matrices de corrélation intra; on peut étudier de façon beaucoup plus précise les trois tableaux de données définis par les trois groupes de variables à

l'aide des méthodes appropriées telles que l'analyse canonique, l'analyse canonique généralisée (Carroll, 1968) ou STATIS (Escoufier, 1979, Foucart, 1984).

TABLEAU 13
matrice de corrélation après classification
par minimisation de la multicollinéarité intra

	x_2	x_7	x_9	x_1	x_4	x_5	x_3	x_6	x_8
x_2	1,0000								
x_7	0,0206	1,0000							
x_9	0,0355	0,0946	1,0000						
x_1	0,0913	0,0803	0,6308	1,0000					
x_4	-0,0046	0,5913	-0,0246	-0,0034	1,0000				
x_5	0,4392	0,0787	-0,0445	-0,0227	-0,0159	1,0000			
x_3	0,5839	-0,0531	0,0785	0,1149	0,0067	-0,4720	1,0000		
x_6	-0,0222	-0,4411	-0,1306	-0,0945	0,4618	-0,0983	0,0658	1,0000	
x_8	0,0715	-0,0093	-0,3394	0,5148	0,0249	0,0204	0,0535	0,0340	1,0000

On notera que les corrélations entre les variables de groupes différents (tableau 13 ci-dessus) ne sont pas très élevées de l'ordre de 0,5 en valeur absolue, de même que les coefficients de corrélation canonique.

La multicollinéarité inter apparaît très nettement dans les corrélations partielles (tableau 14) dont certaines, très proches de 1 en valeur absolue, caractérisent une forte collinéarité (Farrar et Glauber, 1967).

TABLEAU 14
Matrice de corrélation inverse normalisée
(coefficients de corrélation partielle de x_i et x_j
conditionnellement à l'ensemble des autres)

	x_2	x_7	x_9	x_1	x_4	x_5	x_3	x_6	x_8
x_2	1,0000								
x_7	-0,0025	1,0000							
x_9	-0,0800	0,0082	1,0000						
x_1	0,0848	-0,0080	-0,9990	1,0000					
x_4	-0,0015	-0,9987	-0,0077	0,0074	1,0000				
x_5	-0,9986	-0,0050	0,0805	-0,0853	0,0089	1,0000			
x_3	-0,9989	-0,0008	0,0795	-0,0845	0,0048	0,9987	1,0000		
x_6	-0,0008	0,9983	0,0035	-0,0029	-0,9984	-0,0065	-0,0026	1,0000	
x_8	-0,0809	0,0121	0,9985	-0,9988	-0,0115	0,0812	0,0804	0,0071	1,0000

On retrouve ici la propriété mise en évidence par l'exemple du paragraphe 1.1 : la taille de chaque coefficient de corrélation considéré indépendamment du reste de

la matrice ne suffit pas pour caractériser l'existence ou l'inexistence d'une liaison entre les variables. Par contre, les autres coefficients de corrélation partielle sont presque nuls; il est donc très facile d'en déduire les variables collinéaires entre elles (on retrouve bien entendu les groupes déjà formés dans le paragraphe 3.2).

4. Conclusion

La collinéarité entre les variables statistiques peut être expliquée de deux façons différentes : une multicollinéarité inter, existant entre des groupes de variables, et une multicollinéarité intra, présente dans chaque groupe de variables. On peut rechercher les groupes minimisant ou maximisant localement la multicollinéarité intra, et obtenir ainsi des groupes homogènes de variables, c'est-à-dire fortement liées entre elles ou très peu collinéaires suivant que l'on maximise ou minimise la multicollinéarité intra.

La première classification, obtenue par minimisation de la multicollinéarité intra, donne des groupes de variables qui peuvent être utilisées sans difficulté comme régresseurs dans un modèle de régression linéaire multiple. La seconde, obtenue par maximisation de la multicollinéarité intra, correspond plus à la façon de penser traditionnelle en Analyse des Données, puisque les groupes obtenus sont constitués de variables liées tout en étant relativement différenciés.

Ces groupes de variables forment des tableaux de données que l'on peut étudier par des méthodes d'analyse factorielle telles que l'analyse canonique, l'analyse canonique généralisée ou la méthode STATIS lorsque la multicollinéarité totale s'explique pour une grande partie par la multicollinéarité inter ou l'analyse en composantes principales lorsque les sous-espaces qu'ils engendrent sont plus ou moins orthogonaux. Dans certains cas, il est impossible de constituer des groupes de variables dont le rapport de multicollinéarité inter soit très faible ou très élevé. On pourra alors effectuer les deux démarches pour comprendre la collinéarité présente dans les données.

Dans la simulation que nous avons effectuée, nous avons retrouvé sans difficulté à l'aide des algorithmes proposés la structure donnée *a priori* aux variables analysées. Les résultats obtenus sont beaucoup plus faciles à interpréter que ceux d'une analyse en composantes principales habituelle. L'analyse des indices de similarité ou de dissimilarité dont le calcul fait intervenir le carré du coefficient de corrélation ou sa valeur absolue, par l'analyse factorielle et la classification ascendante hiérarchique donne aussi de bons résultats; on notera la similitude de ces indices avec ceux que les algorithmes proposés utilisent (utilisation du carré du coefficient de corrélation, minimisation de l'inertie intra par rapport à l'inertie totale). Par contre les autres indices de dissimilarité que nous avons étudiés ne nous ont pas permis de retrouver la structure des données : pour ces indices, deux variables dont le coefficient de corrélation est proche de -1 sont éloignées l'une de l'autre, ce qui est contradictoire avec la recherche de la multicollinéarité.

Références bibliographiques

Belsley D.A., Kuh E., Welsh R.E. (1980) : *Regression diagnostics : identifying influential data and sources of collinearity*, Wiley, New York.

- Cailliez F. , Pagès J.P. (1976) : *Introduction à l'analyse des données*, SMASH, 9 rue Duban, 75016 Paris, France.
- Carroll J.D. (1968) : A generalization of canonical correlation analysis to three or more sets of variables. *Proc. 76th. Conv. Amer. Psych. Assoc.*, p. 222-228.
- Chatterjee S., Price B. (1977) : *Regression Analysis by Example*, Wiley, New York.
- Erkel-Rousse H. (1995) : Détection de la multicollinéarité dans un modèle linéaire ordinaire : quelques éléments pour un usage ouvert des indicateurs de Belsley, Kuh et Welsh, *Revue de Statistique Appliquée*, vol. XLIII, n°4, 19-42.
- Escoufier Y. (1979) : Operators related to a data matrix : in Recent developments in Statistics, Barra ed., North Holland, Amsterdam.
- Farrar, R.E. , Glauber R.R. (1967) : Multicollinearity in Regression Analysis : The Problem Revisited, *Review of Economics and Statistics*, **49**, 92-107.
- Foucart T. (1984) : *Analyse factorielle de tableaux multiples*, Masson, Paris.
- Foucart T. (1984) : Une nouvelle approche de la méthode Statis, *Revue de Statistique Appliquée*, vol. XXXI, n°2, 61-75.
- Foucart T. (1991) : Transitivité du produit scalaire, *Revue de Statistique Appliquée*, vol. XXXIX, n°3, 57-68.
- Foucart T. (1992) : Collinéarité dans une matrice de produits scalaire, *Revue de Statistique Appliquée*, vol. XXXX, n° , 5-17.
- Furnival G., Wilson R. (1974) : Regression by leaps and bounds, *Technometrics*, 16, 499-511.
- Hoerl A.E., Kennard R.W. (1970) : Ridge Regression : biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Tomassone R., Audrain S, Lesquoy E. et Millier C. (1992) : *La régression. Nouveaux regards sur une ancienne méthode statistique*, Masson, Paris.
- Whittaker J. (1990) : *Graphical models in applied multivariate statistics*, Wiley New York.