UNE NOUVELLE APPROCHE DANS LE CHOIX DES REGRESSEURS DE LA REGRESSION MULTIPLE EN PRESENCE D'INTERACTIONS ET DE COLINEARITES

Michel Lesty

Coryent Conseil, 28, rue Sainte Adélaïde 78000 Versailles France, e-mail coryent@freenet.fr

Résumé

A partir d'exemples simples, on présente une méthode originale de choix des variables et interactions d'ordre 2 dans un modèle de régression multiple, en présence de multicolinéarités ou d'interactions. La méthode "CORICO" (Iconographie des Corrélations) est fondée sur les corrélations totales et partielles. La détection des points aberrants est obtenue au moyen de variables indicatrices des observations. Nous montrons comment l'introduction de fonctions logiques non-linéaires d'ordre 2 améliore l'interprétation du plan d'expérience.

Abstract

On the basis of simples examples, we present an original metod of choice of variables and interactions in a multiple regression model, in the presence of multicolinearities or interactions. CORICO metod is based on total and partial correlations. We display how this practice improves the interpretation of design experiment.

Mots-clés : Régression multiple, interactions, correlations partielles, méthode CORICO, plan d'expérience, prédiction.

1 - Introduction

L'instabilité des coefficients de régressions issus des données d'une expérience non parfaitement planifiée est à l'origine de la majorité des difficultés d'interprétation. Le problème du choix des régresseurs se pose avec une acuité d'autant plus grande que les régresseurs ne sont pas orthogonaux (corrélation nulle).

Une première approche, qui est celle de l'Analyse en Composantes Principales consiste, par une transformation de variables, à retrouver des propriétés d'orthogonalité, absentes à l'origine. Une fois la régression orthogonalisée, on retourne aux régresseurs initiaux tout en maîtrisant mieux les sources d'instabilité. Cette technique ainsi que celle de la régression pseudo-orthogonalisée a fait l'objet d'une abondante littérature (cf. Tomassone, Lesquoy et Millier [1983]).

Une seconde approche, celle de la régression PLS (Partial Least Square), permet de relier un ensemble de variables dépendantes Y à un ensemble de variables indépendantes X. On effectue les analyses en composantes principales des ensembles X et Y sous la contrainte que les composantes principales des Xj soient aussi corrélées que possible à celles des Yk. Les corrélations entre variables prédites et les variables initiales sont plus faibles qu'avec le modèle de régression linéaire classique, mais l'on peut obtenir, si les Yk sont bien choisis, un modèle plus robuste sur un nouvel échantillon (voir Tenenhaus M., Gauchi J.P. et Ménardo C. [1995]).

Toutefois ces approches ne permettent pas le mélange de variables qualitatives et quantitatives, ni la détection d'interactions. Nous présentons ici une autre approche fondée sur la méthode CORICO, permettant la sélection des régresseurs et de fonctions logiques non linéaires de ces derniers, sans recourir aux techniques ascendantes, descendantes ou mixtes. Les corrélations partielles vont nous permettre d'éliminer les régresseurs non pertinents.

2 - La régression multiple

On trouve dans Tenenhaus M., Gauchi J.P. et Ménardo C. [1995] le tableau 1 tiré de Jackson [1991]. Le docteur Linnerud de l'Université de Caroline du Nord a mesuré sur 20 hommes d'âge moyen, s'entraînant dans un club de gymnastique, 3 caractéristiques physiques (variables explicatives) et leurs résultats à 3 types d'exercices (variables à expliquer) :

Tableau 1 Performances en gymnastique

	Poids	TourDeTaille	Pouls	TRACTIONS	FLEXIONS	SAUT
e1	191	36	50	5	162	60
e2	189	37	52	2	110	60
e3	193	38	58	12	101	101
e4	162	35	62	12	105	37
e5	189	35	46	13	155	58
e6	182	36	56	4	101	42
e7	211	38	56	8	101	38
e8	167	34	60	6	125	40
e9	176	31	74	15	200	40
e10	154	33	56	17	251	250
e11	169	34	50	17	120	38
e12	166	33	52	13	210	115
e13	154	34	64	14	215	105
e14	247	46	50	1	50	50
e15	193	36	46	6	70	31
e16	202	37	62	12	210	120
e17	176	37	54	4	60	25
e18	157	32	52	11	230	80
e19	156	33	54	15	225	73
e20	138	33	68	2	110	43

Si nous pensons que la relation entre une variable à expliquer Y et les p variables explicatives X (ou *régresseurs*) est linéaire, nous écrivons pour chacune des n observations de Y:

$$Y_i = b_0 + b_1 X_{1i} + \ldots b_p X_{pi} + u_i$$

Les constantes $b_0, b_1, \ldots b_p$ sont les coefficients de régression partiels du modèle; u_i est un terme aléatoire, c'est à dire la partie non expliquée par le modèle. On estime généralement $b_0, b_1, \ldots b_p$ en résolvant le système de p équations qui minimise la somme des carrés des u_i, c'est à dire qui en annule les dérivées partielles par rapport à chacun des coefficients $b_l (o \leq l \leq p)$.

Pour être directement comparables les coefficients de régression du tableau 2 sont calculés sur les variables centrées réduites :

Tableau 2

Coefficients de régression multiple (centrée réduite)

	TRACTIONS	FLEXIONS	SAUT
Poids	0,3683	0,2872	-0,2590
TourDeTaille	-0,8818	-0,8898	0,0146
Pouls	-0,0258	0,0161	-0,0546

Soient *Ytr*, *Yflx* et *Ysau* les variables prédites au moyen de ces coefficients. Si le modèle était exact, les coefficients de corrélation entre variables réelles et valeurs prédites seraient égaux à 1. Or nous constatons, tableau 3, que la corrélation de *Ytr* avec TRACTIONS est égale à 0.58. Celle de *Yflx* avec FLEXION est 0,66 et celle de *Ysau* avec SAUT est 0.23 :

Tableau 3

Coefficients de corrélation avec Ytr,Yflx et Ysau

		1	2	3	4	5	6	7	8	9
Poids	1	100								
TourDeTaille	2	87	100							
Pouls	3	-37	-35	100						
TRACTIONS	4	-39	-55	15	100					
FLEXIONS	5	-49	-65	23	70	100				
SAUT	6	-23	-19	3	50	67	100			
Ytr	7	-67	-95	26	58	66	15	100		
Yflx	8	-75	-98	34	58	66	16	99	100	
Ysau	9	-97	-82	15	36	46	23	63	69	100

C'est donc que :

- soit le modèle ne contient pas assez de régresseurs,
- soit les relations ne sont pas linéaires ou les régresseurs ne sont pas les bons.

Le signe des coefficients de la régression multiple est souvent différent de celui des corrélations correspondantes. Par exemple au tableau 3 les variables TRACTIONS, FLEXIONS et SAUT sont corrélées négativement à *Poids* et *TourDeTaille*, et positivement à *Pouls*; or au tableau 2 les coefficients de régression de TRACTION et FLEXION sont positifs pour le *Poids*, et celui de SAUT est positif pour *TourDeTaille* et négatif pour le pouls. En fait ces entités mathématiques ne sont pas directement comparables. Le coefficient d'un régresseur X correspond à la *corrélation partielle* de ce dernier avec Y lorsque *tous* les autres régresseurs sont maintenus constants (voir ci-après), tandis que les corrélations totales prennent en compte le fait que les régresseurs sont quelque peu colinéaires (particulièrement *Poids* et *TourDeTaille*)

Dans le tableau 4 sont rassemblées les corrélations partielles des variables à expliquer avec chacune des variables explicatives lorsqu'on retire l'influence des deux autres variables explicatives. Ainsi, la corrélation partielle de TRACTIONS avec *Poids* lorsqu'on retire l'influence de *TourDeTaille* et *Pouls*, est égale à 0.216; la corrélation partielle de TRACTIONS avec *TourDeTaille* lorsqu'on retire l'influence de *Poids* et *Pouls*, est égale à -0,470 etc...

Tableau 4	**Corrélations partielles**		
	TRACTIONS	FLEXIONS	SAUT
Poids	0,216	0,184	-0,129
TourDeTaille	-0,470	-0,503	0,007
Pouls	-0.030	0,020	-0.520

Les signes du tableau 4 correspondent maintenant à ceux du tableau 2.

3 - La régression assistée par CORICO

3.1 - La méthode CORICO

La méthode CORICO, décrite par Lesty et Buat-Ménard [1982], a fait l'objet de plusieurs applications d'analyse de données industrielles (par exemple Lesty et Coindoz [1988], Séchet [1991]) et médicales (Malpéalat-Domaine [1986], Rolland [1988], Lesty et al. [1992], Kujas et al. [1996]). Mais son application à la régression multiple est encore inédite.

Le logiciel analyse méthodiquement les corrélations totales et partielles, et délivre un schéma synthétique des liens non-redondants entre les variables. Grâce aux corrélations partielles, CORICO détermine si la forte corrélation entre deux variables ne découle pas d'une commune dépendance à un troisième terme. Il détecte aussi l'action simultanée de plusieurs variables sur un paramètre, malgré la faible corrélation de celui-ci avec chacune d'entre-elles (voir l'algorithme en annexe).

L'architecture des liens apparaît sur le schéma et leur lecture repose sur des conventions simples : *trait plein* = corrélation positive remarquable; *trait pointillé* = corrélation négative remarquable. On entend par remarquable un lien qui ne peut être entièrement expliqué par une variable intermédiaire parmi celles disponibles. Le problème toujours délicat de donner un nom à des axes factoriels ne se pose pas ici : la figure n'en comporte pas. Ainsi il est possible de représenter schématiquement (sur une sphère à 3 dimensions) pour une lisibilité maximale, des objets de dimensions quelconques. Cela est précieux en particulier dans un plan d'expérience où de nombreux facteurs orthogonaux ou quasi orthogonaux font rapidement monter la dimension. Les liens découlent directement des proximités réelles dans l'espace à n dimensions. Les positions sur la sphère sont aussi disposées si possible (voir annexe) en tenant compte de ces proximités; cependant l'interprétation repose essentiellement sur les liens tracés. La robustesse de CORICO vient du découplage total entre la représentation des liens et celle des positions.

3.2 Recherche d'interactions au sens de CORICO

La notion d'interaction est étendue ici aux fonctions f(A,B). Le tableau 5 montre un plan d'expérience à 4 essais, conçu pour déterminer l'effet sur une réponse Y, des facteurs A et B, à deux niveaux, ("-" symbolise le niveau faible et "+" symbolise le niveau fort) :

Tableau 5

	A	B	A.B	A*B	A&B	Y
essai 1	-	-	+	-	-	...
essai 2	-	+	-	+	-	...
essai 3	+	-	-	+	-	...
essai 4	+	+	+	-	+	...

A.B représente l'effet croisé. A*B = -A.B représente l'interaction logique "A ou-exclusif B". A&B représente l'interaction logique "A et B". A, B et A.B (ou A*B si l'on préfère) sont orthogonaux. Mais A&B n'est pas orthogonal aux précédents. Il n'est donc pas utilisé dans une analyse de variance qui requiert des facteurs orthogonaux. CORICO au contraire peut l'intégrer dans l'analyse y compris lorsque A et B ne sont pas orthogonaux, d'où un moyen d'éviter les confusions d'interactions (alias), alors moins fréquentes. Dans le cas plus général de variables continues, l'interaction A&B peut être définie comme suit :

Soient a et b les variables A et B centrées-réduites, et soient a_{min} et b_{min} les valeurs absolues des minimums respectifs de a et b :

$A_i \& B_i = (a_{min} + a_i)(b_{min} + b_i)$ pour $1 \le i \le n$ où n est le nombre d'observations,

tandis que $A_i * B_i = -a_i b_i$. En réalité, CORICO n'utilise pas les valeurs centrées-réduites de A et B, mais une autre unité qui équivaut à un simple changement d'échelle : les "corrélations variables-instants" (voir annexe), afin d'optimiser les opérations.

Sur ce principe, on peut définir d'autres types d'interactions. Elles apparaissent sur le schéma si elles ont un effet sur la réponse. On peut également considérer le logarithme, le rang, ou une fonction sinusoïdale de chacun des facteurs etc... Certaines n'existent que si A et B ont au moins trois niveaux ou sont continues. En voici une liste non exhaustive :

Tableau 6 Table des symboles utilisés sur les schémas de CORICO

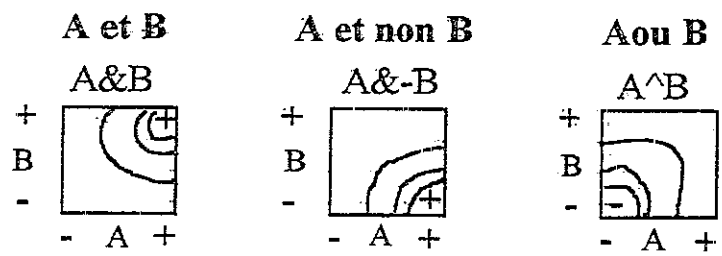
f(A,B)	Signification	La réponse Y est forte lorsque...
A*B	A ou-exclusif B	...A est fort et B faible ou A est faible et B fort
A^B	A ou B	...A est fort ou B est fort
A^-B	A ou non B	...A est fort ou B est faible
A&B	A et B	...A et B sont forts
A&-B	A et non B	...A est fort et B est faible
A]B	A modulé par B	...A est fort si B est fort
A]-B	A modulé par non B	...A est fort si B est faible
A'B	ni A ni B (sens large)	...ni A ni B ne sont extrêmes (ils sont moyens)
A!B	ni A ni B (sens strict)	...ni A ni B ne sont extrêmes (ils sont strictement moyens)
A#B	A comme B	...A varie comme B
A+B	"A plus B"	...la somme de A et B (centrés-réduits) est forte
A-B	"A moins B"	...la différence de A et B (centrés-réduits) est forte

Tableau 6 bis

f(A,B)	Formules pour toute observation i
A*B	$-a_i b_i$
A^B	$-(a_{max} - a_i)(b_{max} - b_i)$
A^-B	$-(a_{max} - a_i)(b_{min} + b_i)$
A&B	$(a_{min} + a_i)(b_{min} + b_i)$
A&-B	$(a_{min} + a_i)(b_{max} - b_i)$
A]B	$a_i(b_{min} + b_i)$
A]-B	$a_i(b_{max} - b_i)$
A'B	$-a^2_i - b^2_i$
A!B	$(a^2_{min} + a_i)(b^2_{min} + b_i)$
A#B	$-(a_i - b_i)^2$
A+B	$a_i + b_i$
A-B	$a_i - b_i$

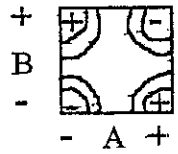
Les formules du tableau 6 bis ne sont pas forcément intuitives, mais le sens logique (les surfaces de réponse, figure 1) des fonctions f(A,B) est tout à fait intuitif si A et B sont orthogonaux. A l'opposé, lorsque A et B sont parfaitement corrélés, les graphes de la **figure 12** en diront plus qu'un long discours (dans ce cas, on ne connaît que la diagonale de la surface de réponse)

Figure 1



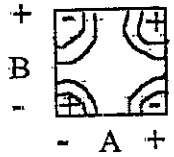
A ou-exclusif B

A*B



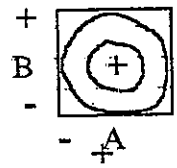
Effet croisé

A.B

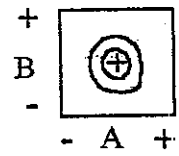


Ni A ni B ne sont extrêmes

A'B

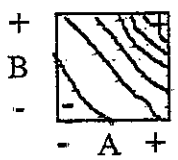


A!B



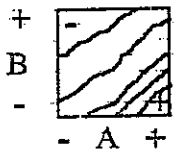
Addition

A+B



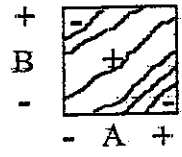
Soustraction

A-B



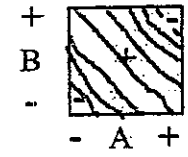
A comme B

A#B



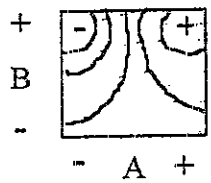
A comme non B

A#-B



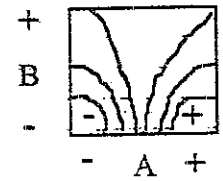
A si B

A]B



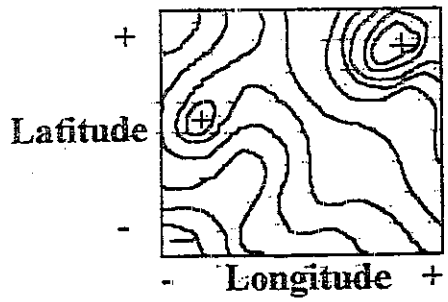
A si non B

A]-B



On parle d'effet d'interaction chaque fois qu'une variable est conditionnée par les valeurs simultanées de deux autres variables. Ainsi, l'altitude d'une voiture dépend de sa latitude et de sa longitude. La carte des altitudes en fonction de la longitude et de la latitude, présente des montagnes (+) et des vallées (-) :

Figure 2



Cependant il existe une infinité de dispositions possibles de montagnes et de vallées. Dans CORICO on s'intéresse seulement à celles qui peuvent attirer l'attention sur des phénomènes simples ou logiques, ou qui présentent un caractère de généralité susceptible d'intéresser tout utilisateur.

NOTE: Autrement il y aurait danger de s'enferrer dans une recherche sans fin. Par exemple, on pourrait chercher si l'interaction Log(A)/Exp(B**3) a une influence sur telle ou telle variable. Mais il y a une infinité de formules mathématiques. C'est à l'utilisateur de saisir en tant que données celles auxquelles il pense.

Il est assez rare de pouvoir tracer la « carte » des valeurs d'une variable Y en fonction de deux variables A et B. Ainsi l'altitude d'un avion n'est pas conditionnée par le couple latitude-longitude, car pour une une latitude et une longitude donnée, l'avion peut se trouver à des altitudes différentes à deux instants différents. La surface de réponse dessinée d'après les moyennes des altitudes pour chaque couple latitude-longitude est dépourvue de sens physique.

A condition de travailler sur les données brutes et non sur les moyennes des mesures répétées, la répétition (refaire passer plusieurs fois l'avion sur les mêmes coordonnées A,B) permet d'éviter de conclure à la présence d'une interaction f(A,B), en fait inexistante.

3.3 - Relation entre les interactions logiques

Une variable de la forme $\alpha A + \beta$ possède des variations identiques à celle de la variable A, quelles que soient les constantes α et β. En effet CORICO ramène toutes les variables à une unité commune d'évaluation (la corrélation *variable-instant*, voir annexe). Les cubes représentatifs de ces variables sont donc confondus sur le schéma de CORICO. Résumons cela sous forme symbolique :

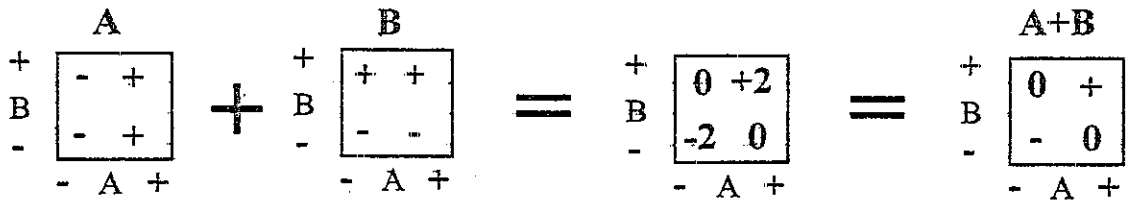
$$A \equiv \alpha A + \beta$$

En particulier : $A \equiv 2A = A + A$

Par contre $A - A = 0$ pour toutes les valeurs de A. C'est donc une constante qui n'apparaît pas à la surface de la sphère de CORICO (elle se trouve, si l'on veut, au centre de la sphère avec toutes les constantes)

Qu'en est-il d'une variable de la forme A + B ?

Pour plus de clarté, raisonnons sur les surfaces de réponses des variables à deux niveaux : leur somme est une variable à trois niveaux équidistants :

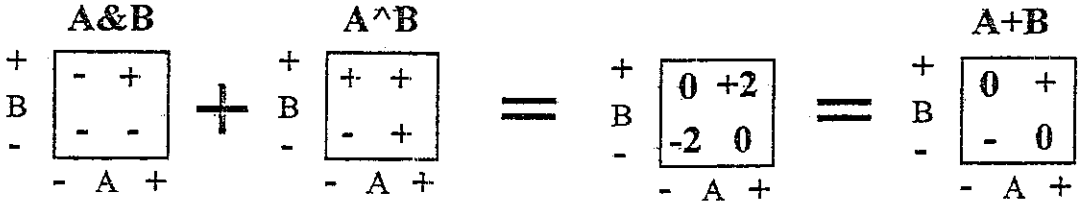


Sur la sphère de CORICO, la variable A+B se trouve au milieu de l'arc de grand cercle joignant A et B, de même que toutes les variables de la forme :

$$\alpha (A+B) + \beta (A+B) + \gamma$$

Par contre, la variable A+2B serait plus proche de B sur ce même arc, etc...

On aurait également obtenu A+B en ajoutant les interactions A&B et A^B en effet:



la variable A+B se trouve donc au milieu d'un autre arc de grand cercle qui joint A&B et A^B:

$$A\&B + A\hat{}B \equiv A+B$$

De la même façon on peut établir la relation :

$$A + B + A*B \equiv A\hat{}B$$

Elle signifie que, sur la sphère de CORICO, la variable A^B se trouve au centre du triangle dont les sommets sont A, B et A*B. On montrerait de même que:

$$A + B - A*B = A + B + \overline{A\,B} \equiv A\&B$$

Ces relations, établies sur des variables à 2 niveaux restent valables quel que soit le nombre de niveaux. Par exemple, le plan d'expérience complet pour deux variables à 7 niveaux nécessite 49 essais qu'on peut répartir de la manière suivante:

A: 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7

B: 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 6 7 7 7 7 7 7 7

Note: puisqu'il s'agit d'un plan complet, les variables A et B sont orthogonales (leur corrélation est nulle).

A partir de ces variables, CORICO fabrique diverses interactions, dont les positions relatives ont été tracées sur la **figure 3** ***(à la fin de cet article, ainsi que les figures suivantes)***. Certaines interactions n'ont pas été représentées pour alléger la figure. En outre nous n'avons pas tracé les liens d'oppositions (traits pointillés). Par exemple A*-B = A.B se trouve à l'opposé de A*B par rapport au centre de la sphère:

$$A*\text{-}B = \text{-}A*B$$

De même, selon les règles de la logique :

$$\text{-}(A\hat{}B) = \text{-}A \,\&\, \text{-}B$$
$$\text{-}(A\hat{}\text{-}B) = \text{-}A\&B$$
etc.

Au vu de la figure 3, il est facile d'établir les relations :

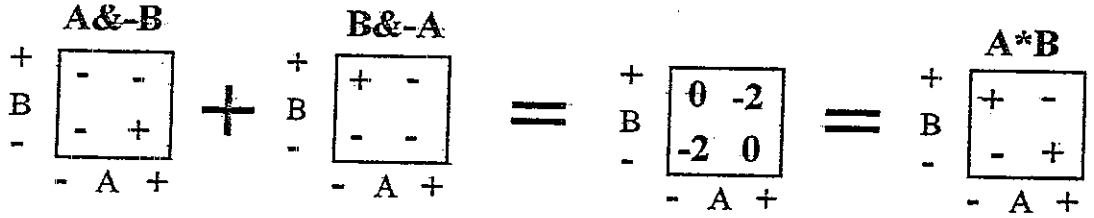
$$A + A*B \equiv A]\text{-}B$$

En effet, A]-B se trouve au milieu de l'arc joignant A et A*B. De même:

$$A + A.B \equiv A]B$$
$$B + A*B \equiv B]\text{-}A$$
$$A\hat{}\text{-}B + A\&\text{-}B \equiv A\text{-}B$$
etc.

Si nous avions introduit les variables A'B et A!B, elles auraient été posées tant bien que mal sur la sphère. En effet leurs positions exactes par rapport aux précédentes nécessitent en toute rigueur la sphère à 4 dimensions.

On a encore la relation:



A*B est donc à égale distance de A&-B et B&-A sur l'arc de grand cercle qui les joint. Lorsque B tend vers A, A&-B tend vers B&-A. Il en résulte que

$$A\&\text{-}A \equiv A*A$$

La **figure 4** montre les positions relatives des interactions de type A&A, A*A, A^A, A]A, etc...

Ce type d'interaction ne sera détecté que si A possède au moins trois niveaux. Quand A n'a que deux niveaux, l'interaction A&A est égale à A, tandis que A*A est une constante. En revanche, si A possède trois niveaux, la variable A*A est maximum lorsque A est moyenne. Pour détecter l'interaction A'B il faut en principe au moins 6 essais (3 niveaux pour A, 2 niveaux pour B) et de préférence 9 essais (3 niveaux pour A, 3 niveaux pour B). Lorsque les essais varient librement, et non pas sur des niveaux imposés, CORICO dispose de plus d'indices, tel un bon détective, pour aboutir au résultat.

$$A*A \equiv -A\&A \equiv A'\text{-}A$$

$$A!A \equiv -A!A \equiv -A!\text{-}A$$

$$A \equiv A+A$$

$$-A \equiv -A - A$$

NOTE : Puisque la variable A est exprimée dans l'unité de mesure *Corrélations variables-instants*, qui est un cosinus, Son arc-cosinus se comporte comme un angle. Les variables cos(acos(A)+π/2) et cos(acos(A)-π/2 se trouvent donc en quadrature avec A sur la sphère (orthogonales avec A), exactement comme A*-A et A*A sont orthogonales à A. Nous avons:

$A*\text{-}A = A\,A \equiv \cos(\text{acos}(A) + \pi/2) = \cos(\text{acos}(\text{-}A) + \pi/2)$

$A*A \equiv \cos(\text{acos}(A) - \pi/2) = \cos(\text{acos}(\text{-}A) - \pi/2)$

3.4 Analyse exploratoire

Soit Y la réponse d'un plan à deux facteurs orthogonaux A et B. Chercher si par exemple Y égale A B + A / Log(B) s'apparente à la pêche à la ligne, car on pourrait envisager une infinité d'autres formules mathématiques. Et si par chance nous tombions juste, la formule ne serait pas forcément facile à expliquer.

CORICO effectue au contraire une sorte de pêche au filet. Essayant toutes les interactions définie plus haut, il ne trace sur le schéma que les interactions remarquables. Si Y apparaît lié à l'une des interactions qui constituent les noeuds du filet, il devient plus aisé d'avancer une explication. En effet chaque interaction a un sens logique. A partir de quoi on peut, plus facilement, établir une relation mathématique si cela est requis.

Souvent le poisson échappe à nos plans. Mais il existe peut-être un facteur D, incontrôlé, qui permet de lancer deux autres filets constitués par les interactions de A avec D d'une part, et par celles de B avec D d'autre part. Quand D n'est orthogonal ni avec A ni avec B, ces filets ne couvrent pas aussi bien la sphère que celui de la figure 3, mais, dans leur

zone d'opération, les mailles sont beaucoup plus serrées (les filets ne se recouvrent pas sur l'hypersphère à n dimensions).

Les facteurs incontrôlés sont en général plus nombreux que les facteurs contrôlés. Ce sont souvent des facteurs essentiels, mais il serait coûteux, voire impossible, de les contrôler au prix de dispositifs délicats. La méthode permet donc d'en tirer parti. A l'issue de cette exploration, on peut choisir le sous-ensemble parmi les variables disponibles, qui entrera dans le modèle de la réponse Y.

3.5 Analyse et choix des régresseurs des performances en gymnastique

Lorsqu'on applique aux données brutes le logiciel CORICO, on obtient le schéma de la **figure 5**. Les trois variables à expliquer sont liées. De même *Poids* et *Tour de taille*. Ces liens ne sont pas forcément linéaires. Mais il est clair que FLEXIONS dépend plus, négativement (ligne pointillée), du *Tour de taille* que du *Poids*. Quant au *Pouls*, il n'exerce pas d'influence remarquable. Enfin, il n'y a pas d'interactions telles que définies au §3.2.

Approfondissons l'étude en retirant ce qui dans toutes les variables est lié linéairement à la FLEXION (**figure 6**, corrélations partielles par rapport à la FLEXION. On représente la corrélation partielle sachant la variable retirée). Il apparaît maintenant une dépendance positive entre le SAUT et le *Tour de taille*. Cette relation indique le facteur secondaire, favorable au SAUT: le saut est certes favorisé par une bonne FLEXION (figure 5) laquelle est meilleure si la personne est mince (souplesse accrue) et donc légère. Mais d'un autre côté (figure 6) un fort *Tour de taille*, abstraction faite du manque de souplesse, est lié à une musculature puissante propice au SAUT.

En revanche le *Poids* n'est jamais directement lié, sur les deux schémas, aux performances de gymnastique (positivement ou négativement). C'est pourtant lui qui, des trois variables explicatives, possède la plus forte corrélation totale (négativement) avec le SAUT. Mais cette valeur n'est pas *remarquable* pour CORICO, une fois retirées les influences qui lui sont mêlées. On aurait pu penser d'autre part que, abstraction faite des problèmes de flexion, l'inertie associée au poids favorise le SAUT (pour un même élan le corps est moins facile à arrêter), mais cela n'apparaît pas figure 6 : la résistance de l'air, seule force de frein susceptible d'intervenir, peut donc être négligée.

Le *Pouls* n'est lié à rien sur les deux figures. On peut le retirer sans grand dommage pour ne conserver que deux régresseurs. Au tableau 7, les corrélations des variables ainsi prédites par le modèle, à savoir *tpoiTr*, *tpoiFlx* et *tpoiSau* avec TRACTIONS, FLEXIONS et SAUT, sont respectivement 0.58, 0.66 et 0.23, pratiquement égales à celles obtenues au tableau 3 avec les trois régresseurs.

Tableau 7
Coefficients de corrélation avec tpoiTr, tpoiFlx et tpoiSau

		1	2	3	4	5	6	7	8	9
Poids	1	100								
TourDeTaille	2	87	100							
Pouls	3	-37	-35	100						
TRACTIONS	4	-39	-55	15	100					
FLEXIONS	5	-49	-65	23	70	100				
SAUT	6	-23	-19	3	50	67	100			
tpoiTr	7	-67	-95	30	58	66	14	100		
tpoiFlx	8	-75	-98	32	58	66	16	99	100	
tpoiSau	9	-100	-85	36	37	47	23	63	71	100

Dans cet exemple simple, CORICO permet rapidement d'éliminer une variable inutile au modèle de régression multiple. L'intérêt de cette approche augmente avec le nombre de régresseurs. Les chiffres de corrélations totales et partielles sont alors trop nombreux pour être regardés un à un.

3.6 - Les difficultés de la régression multiple

Les difficultés classiques doivent être rappelées afin de situer l'apport de CORICO. Dans une régression multiple (PLS ou non), des coefficients sont trouvés même si les régresseurs Xi sont choisis au hasard. On peut toujours remplacer un nuage de point par une droite. Bien plus, on trouvera un modèle exact de l'échantillon à expliquer (corrélation=1 avec la valeur prédite) si le nombre p de variables explicatives est égal au nombre n d'observations. Car on dispose alors d'autant d'équations que d'inconnues. Il ne s'en suit pas que la prédiction est valable pour un autre échantillon. Le problème se pose d'abord d'éliminer les régresseurs sans rapport avec la question.

A cela s'ajoutent le problème de la non-orthogonalité des régresseurs, et celui des points suspects.

Quand deux régresseurs Xk et Xj ne sont pas orthogonaux (corrélation différente de 0), la relation entre les coefficients bk et bj de la régression rend illusoire leur interprétation séparée. Par le jeu des interrelations entre régresseurs, on obtient parfois un coefficient de régression négatif sur Xi, alors que Y et Xi sont corrélés positivement. Certains régresseurs sont redondants. Les techniques de sélection ascendantes, descendantes ou mixtes sont parfois contradictoires: l'ordre d'introduction des variables est un procédé analytique; rien ne garantit qu'il ait un sens physique. La régression multiple est donc mieux adaptée au plan d'expérience dont les facteurs sont orthogonaux. Même dans ce cas les coefficients peuvent être biaisés par la présence de facteurs cachés avec phénomènes de confusions (alias). Enfin il faut prendre en compte de possibles interactions entre régresseurs.

En résumé : un coefficient de régression n'a pas en soi de sens physique car il dépend de l'utilisateur (selon le choix des variables introduites dans le modèle). Le coefficient de corrélation au contraire, même s'il n'a pas de sens isolément, ne dépend pas du modèle ni de l'utilisateur. C'est donc un point d'appui solide pour l'interprétation.

Quant au problème des points suspects, le calcul des contributions de chaque observation est utile (par exemple en les supprimant chacune à son tour), à condition d'avoir le loisir de le dépouiller.

Face à ces difficultés, CORICO (voir annexe), s'appuyant seulement sur les données brutes et sur les corrélations (qui ne dépendent pas de lui),

- peut nous aider dans le choix des régresseurs, en se focalisant sur les points essentiels,
- n'est pas gêné par les colinéarités (il tire parti des corrélations partielles pour éliminer les redondances),
- explore méthodiquement l'incidence des points suspects et des interactions,
- permet, en cas de mélange de populations identifiées par une variable indicatrice, de découvrir leurs modèles respectifs.

4 - Application à l'analyse de variance

L'analyse de variance est un cas particulier de la régression multiple lorsque les variables sont qualitatives. Le tableau 8, tiré de Tomassone, Lesquoy et Millier [1983] contient à la fois des variables qualitatives et des variables quantitatives.

Tableau 8 Gain de poids des animaux

	Dose	PoidsInitial	Boeuf	Céréales	Porc	GAINpoids
e1	0	1	1	0	0	6
e2	0	3	1	0	0	8
e3	0	6	1	0	0	9
e4	0	9	1	0	0	11
e5	0	10	1	0	0	11
e6	0	1	0	1	0	6
e7	0	3	0	1	0	7
e8	0	5	0	1	0	8
e9	0	7	0	1	0	9
e10	0	9	0	1	0	9
e11	0	1	0	0	1	3
e12	0	3	0	0	1	3
e13	0	5	0	0	1	6
e14	0	8	0	0	1	7
e15	0	10	0	0	1	10
e16	1	1	1	0	0	16
e17	1	2	1	0	0	17
e18	1	4	1	0	0	17
e19	1	6	1	0	0	18
e20	1	9	1	0	0	18
e21	1	1	0	1	0	9
e22	1	3	0	1	0	9
e23	1	4	0	1	0	10
e24	1	7	0	1	0	10
e25	1	10	0	1	0	12
e26	1	2	0	0	1	16
e27	1	4	0	0	1	16
e28	1	5	0	0	1	17
e29	1	8	0	0	1	17
e30	1	10	0	0	1	18

On analyse le gain de poïds de 30 animaux nourris de 6 façons différentes. Les facteurs sont :

- La dose d'aliment, à 2 niveaux: basse (0) et élevée (1).
- L'origine des aliments, à 3 niveaux: *boeuf, céréales, porcs*. L'origine a été ventilée sur 3 colonnes. Pour la régression, on n'utilise que les 2 premières modalités de cette variable. En effet la troisième est parfaitement connue à l'aide des précédentes.
- Le poids initial avant expérimentation

La régression multiple brute donne l'équation suivante :

$$GAINprédit = 5.24 + 0.43\ PoïdsInitial + 7.28\ Dose + 2.02\ Boeuf - 2.14\ Céréales$$

Le coefficient de détermination R^2 est 0.8161 (carré du coefficient de corrélation entre GAINpoids et GAINprédit). Si l'on ajoute au modèle un régresseur quelconque distinct des précédents, par exemple un terme d'interaction, pertinent ou non, le coefficient R^2 sera forcément supérieur. Voici à titre d'exemple, tableau 9, les valeurs de R^2 pour divers choix du cinquième régresseur parmi des interactions possibles :

Tableau 9

5 ième régresseur	R^2
Poids*boeuf	0.8173
Poids*céréales	0.8174
Dose*boeuf	0.8299
Dose*céréales	0.9318

Une équation à 6 régresseurs comprenant les 4 facteurs initiaux et les 2 dernières interactions du tableau, donne $R^2 = 0.974$ etc... Mais, plus sont nombreux les régresseurs choisis au hasard et plus hasardeuse sera la prédiction sur un nouvel échantillon.

Il est préférable de ne conserver que des régresseurs pertinents, même si le coefficient R^2 est un peu moins bon. La méthode CORICO avec recherche des interactions liées au gain de poids donne la **figure 7**. Le logiciel envisage 220 "interactions" entre les 5 variables explicatives (voir §3.2 : 15 de types *,&,^,!, 25 de types],]-,&-,^-, 20 de type -, 10 de types +,',#, 5 logarithmes et 5 rangs). Or une seule est apparue remarquable : *Dose&-Céréales* (aliasée négativement avec *Céréales^-Dose*, ce qui ne surprend pas). Cette variable, fabriquée comme il est dit au §3.2, souligne l'importance de la conjonction de protéines *sans* céréales *et* d'une forte dose d'aliment. L'équation de la régression devient :

$$GAINprédit = 6.38 + 0.045\ PoidsInitial + 2.2\ Dose + 2.02\ Boeuf - 1.68\ Céréales + 19.36\ Dose\&-céréale$$

Alors $R^2 = 0.9668$.

Mais la figure 7 suggère de simplifier la régression pour ne conserver que les facteurs *PoidsInitial, Boeuf, Dose* et *Dose&-Céréales,* liés directement au gain. Alors $R^2 = 0.9543$.

En fait, la variable *Dose &- Céréale* contient implicitement la plupart de l'information relative aux facteurs *Dose, Boeuf, porc* et *céréales* (comme ce ne sont pas des céréales, c'est forcément du boeuf ou du porc). Conservons seulement les régresseurs *PoidsInitial* et *Dose&-Céréales*. L'équation devient:

$$GAINprédit = 8.9 + 0.419\ PoidsInitial + 22.68\ Dose\&\text{-}céréale$$

$R^2 = 0.892$ reste très bon par rapport à celui de la régression multiple brute (remarquons qu'avec le seul régresseur *Dose&-céréale,* le coefficient R^2 tombe à 0.812 tandis que l'équation à deux régresseurs *PoidsInitial* et *Dose*céréale* donne le très mauvais coefficient $R^2 = 0.214$. L'opérateur "&" colle donc beaucoup mieux à la réalité que l'opérateur "*"). La réduction du nombre de régresseurs indispensables peut faciliter les prévisions ultérieures.

5 - Expériences de modélisation économique

A partir de données mensuelles **(figure 9)**, 285 mois, de janvier 1973 à septembre 1996, on dispose

d'indicateurs financiers :

Monétaire	placements monétaires à court terme (cash)
Obligations	placements à long terme
ACTIONS	indice du marché des actions américain

d'indicateurs sur la monnaie :

Masse	masse monétaire
Change	cours du Dollar par rapport à un panier de monnaies

et d'indicateurs économiques :

t	temps
Production	indice de production industrielle
Prix	prix à la consommation
Chômage	taux de Chômage
Balance	balance commerciale

Le temps n'explique rien, mais nous l'avons rangé parmi ce dernier groupe car il peut refléter une croissance ou un vieillissement qui sont la résultante d'une multitude de facteurs non pris en compte explicitement.

Procédons à deux expériences de modélisation : l'une pour prévoir les actions, l'autre pour prévoir le chômage.

5.1 - Indice du marché des actions américain

La régression multiple simple de ACTIONS sur les 9 régresseurs centrés-réduits donne le modèle suivant :

$$Y1 = -1.638\,t + 0.814\,Production + 2.006\,Prix + 0.093\,Chômage + 0.038\,Balance + 0.059\,Monétaire - 0.444\,Obligations - 0.200\,Masse + 0.015\,Change$$

Le pouvoir prédictif du modèle peut être apprécié par le coefficient de détermination R^2, carré du coefficient de corrélation entre ACTIONS et sa prédiction Y1:

$$R^2 = 0.965$$

Cependant la **figure 10,** d'après CORICO avec recherche des interactions liées à ACTIONS, montre que *ACTIONS* est liée négativement au *Monétaire* et , positivement, à *Change* et à *t&t*. L'effet *t&t* a été trouvé parmi 702 fonctions non linéaires de deux parmi 9 facteurs. Son allure en fonction de t est représentée **figure 12.**

La régression multiple sur ces trois régresseurs donne le modèle :

$$Y2 = 0.925\,t\&t - 0.038\,Monétaire + 0.057\,Change$$

avec $R^2 = 0.974$

La prédiction est donc meilleure, et avec moins de régresseurs. Cependant le facteur *t&t* joue un rôle prépondérant. Voyons l'effet du retrait de cette composante, **figure 11 :**

ACTIONS est désormais lié à *Prix*. Les liens à *Change* et à *Monétaire* ont disparus. Ainsi, ces deux derniers facteurs étaient surtout liés à ACTIONS par leur composante dépendant de *t&t*. Ecrivons donc le modèle suivant :

$$Y3 = 1.292\,t\&t - 0.322\,Prix$$

avec $R^2 = 0.981$

La prédiction Y3 est encore meilleure, avec seulement deux régresseurs. Comparons les modèles Y1 et Y3. Dans le premier, le coefficient du régresseur t était négatif, alors que dans Y3, le coefficient de *t&t* est positif. En effet dans le modèle Y1 l'effet du temps est réparti sur plusieurs régresseurs. *Production, Masse* et *Prix* sont directement liés au temps, figure 10. Mais *t&t* est sans doute un meilleur indicateur de la croissance générale du pays que ces trois variables prises séparément ou en interaction.

5.2 - Un modèle du chômage

La régression du chômage sur les 9 autres variables donne $R^2 = 0.943$, mais sur la **figure 10** le chômage est lié à *Obligations* et, négativement, au *Change*. Une régression bâtie sur ces deux variables centrées-réduites donne :

$$Y1Chôm = 0.257\,Obligations - 0.384\,Change$$

avec $R^2 = 0.318$

L'analyse des données, avec recherche des interactions liées au chômage, donne la **figure 13**, où *Chômage* est lié à *Obligations*, à *Change* et à l'"interaction" *Production - Prix* qui est la différence calculée entre les données centrées-réduites. Pour réduire le chômage il conviendrait donc d'augmenter le différentiel entre la production industrielle et le prix à la consommation. Avec ces régresseurs on obtient :

$$Y2Chôm = -0.0429\ Obligations - 0.290\ Change - 0.766\ Prod\text{-}Prix$$

avec $R^2 = 0.777$

Dans Y2Chôm, l'interaction *Prod-Prix* joue le rôle majeur. Retirons sa composante, **figure 14.** Comme on retire ce qui les distingue, *Production* et *Prix* deviennent exactement corrélés (le second n'est pas dessiné par CORICO pour alléger figure et calculs). Le lien avec *Obligations* disparaît aussi ! Il existait donc surtout par ce qui est commun à *Obligations* et *Prod-Prix*. Mais un lien apparaît avec *Masse*. D'où le modèle :

$$Y3Chôm = -0.263\ Masse - 0.090\ Change - 0.838\ Prod\text{-}Prix$$

avec $R^2 = 0.815$

D'après les coefficients de la régression, *Change* joue le rôle le moins important (d'ailleurs il est lié à Obligations sur toutes les figures). Supprimons-le. L'équation devient :

$$Y4Chôm = -0.319\ Masse - 0.873\ Prod\text{-}Prix$$

avec $R^2 = 0.811$

Ce qui reste assez bon, compte tenu de la réduction du nombre des régresseurs

S'agissant des "interactions" de type "+" ou "-", on peut considérer les termes séparément, pour augmenter le degré de liberté. L'équation devient :

$$Y4Chômbis = -1.771\ Masse - 2.775\ Production + 4.254\ Prix$$

avec $R^2 = 0.865$

Il est intéressant de comparer les valeurs des coefficients de corrélations du chômage avec ces différents facteurs. Les variables qui, séparément, ont la corrélation la plus forte avec le chômage, ne sont pas, finalement, retenues dans le modèle :

Tableau 10

	R
Obligations	0.460
Change	-0.520
Masse	-0.235
Production	-0.395
Prix	-0.152
Prod-Prix	-0.842

Si l'on découpe la période 1973-96 en deux périodes : "Avant" (jan-73 à déc-84) et "Après" (jan-85 à sep-96), le modèle Y4Chômbis donne respectivement 0.870 et 0.801 pour R^2, tandis que Y1chôm donne 0.359 et 0.097.

6 - Simulation

Une simulation à partir de la régression pas à pas serait difficile car les logiciels disponibles ne sont pas conçus pour détecter les interactions au sens de CORICO. Les difficultés classiques des techniques ascendantes et descendantes sont rappelées au §3 6. L'algorithme donné en annexe montre que dans CORICO le choix des régresseurs ne dépend pas de l'ordre des opérations. Ceci est encore plus précieux quand le nombre de régresseurs est grand (711 dans l'exemple précédent : 9 variables plus 702 interactions).

6.1 - Simulation de ACTIONS

Pour simuler la prédiction de ACTIONS, recherchons un modèle dans la période "Avant" (jan-73 à déc-84), afin de le tester sur la période "Après" (jan-85 à sep-96).

Pour la période "Avant", la régression multiple simple sur les 9 régresseurs centrés réduits donne le modèle suivant :

$$Y1avant = -3.970\ t + 0.954\ Production + 3.415\ Prix + 0.255\ Chômage + 0.036\ Balance - 0.205\ Monétaire - 0.605\ Obligations + 1.09\ Masse - 0.0396\ Change$$

avec $R^2 = 0.862$

Cependant l'analyse CORICO, avec recherche d'interactions liées à ACTIONS, montre, **figure 15**, que sur cette période, *ACTIONS* est liée seulement à t]t. D'où le modèle :

$$Y2avant = 0.914\ t]t$$

avec $R^2 = 0.835$

Pour savoir s'il existe des influences secondaires, retirons, **figure 16**, la composante t]t. Alors ACTIONS apparaît lié à *Obligations* et à Production. Le modèle devient :

$$Y3avant = 0.134\ Production - 1.62\ Obligations + 0.943\ t]t$$

avec $R^2 = 0.845$

Maintenant nous pouvons tester ces différents modèles sur la période "Après":

Tableau 11

Modèle	R^2
Y1avant	0.862
Y2avant	0.857
Y3avant	0.906
Y3	0.937

Le modèle *Y3avant* avec seulement trois régresseurs donne une meilleure prédiction que le modèle calculé sur les 9 régresseurs d'origine. La prédiction marche même mieux sur la période "Après" que sur la période "Avant". Par contre, il est normal que le modèle Y3, calculé précédemment sur l'ensemble des données soit meilleur pour la période "Après", puisque ce n'est plus alors une prédiction.

6.2 - Simulation du chômage

Sur la période "Avant", la régression multiple simple sur les 9 régresseurs centrés réduits donne le modèle suivant :

$$Y1chômAvant = 2.093\ t\ -1.232\ Production + 0.988\ Prix + 0.068\ Balance - 0.062\ Monétaire + 0.046\ ACTIONS - 0.108\ Obligations - 1.397\ Masse - 0.167\ Change$$

avec $R^2 = 0.975$

Or, d'après CORICO, avec recherche des interactions liées au chômage, **figure 17**, le chômage est lié sur cette période à *t-Production, Prix, Monétaire-Obligations* et *Change*. Le modèle sur ces facteurs centrés réduit est :

$$Y2chômAvant = -0.066\ Monétaire\text{-}Obligations + 0.0013\ Change + 0.775\ t\text{-}Production$$

avec $R^2 = 0.960$

Le facteur le plus influent est t-Production. Retirons sa composante, **figure 18**. Alors disparaissent tous les autres liens précédents, tandis qu'un lien avec t apparaît. D'où le modèle :

$$Y3chômAvant = 0.359\ t + 0.819\ t\text{-}Production$$

avec $R^2 = 0.962$

Le coefficient de détermination s'est légèrement amélioré avec seulement deux régresseurs. S'agissant d'une interaction de type "-" on pourrait considérer les termes séparément, mais ici, contrairement à Y4chômbis, le degré de liberté n'augmente pas puisqu'il n'y a toujours que deux régresseurs. Dans ce cas l'équation devient :

$$Y3chômAvantbis = 1.847\ t - 1.487\ Production$$

avec $R^2 = 0.962$

Le coefficient de détermination ne change pas, et c'est normal d'un point de vue mathématique. Mais nous allons voir que ce modèle a un moins bon pouvoir prédictif.

Maintenant nous pouvons tester les différents modèles sur la période "Après":

Tableau 12

Modèle	R^2
Y1chômAvant	0.298
Y2chômAvant	0.299
Y3chômAvant	0.242
Y3chômAvantbis	0.129
Y4chôm	0.694
Y4chômbis	0.802

Au vu des résultats, le modèle *Y3chômAvantbis* colle sans doute de moins près à la réalité physique que le modèle *Y3chômAvant*. En effet CORICO a d'abord souligné l'importance du différentiel entre le temps et la Production industrielle (rappellons qu'on

travaille sur les valeurs centrées réduites, c'est à dire qu'on s'intéresse seulement aux variations, abstraction faite des unités de mesure).

La prédiction de la période "Après" à partir de la période "Avant" est moins satisfaisante pour le chômage que celle que nous avons obtenue pour ACTIONS. (Y4chômbis est bon car il est calculé sur les deux périodes : ce n'est pas une prédiction). Ceci doit nous rendre prudent; il est des cas où le contexte a trop varié pour que la prédiction soit possible. Le chômage dépend sans doute d'autres facteurs non donnés. En effet ce corpus de données n'était pas spécifiquement destiné à l'analyse du chômage

7 - Discussion

Le choix des régresseurs comporte toujours un aspect automatique et un aspect volontaire. Il se ramène à trois étapes, éventuellement répétées :

Etape 1 : La construction du schéma est purement mathématique, donc programmable de façon automatique ou mécanique. L'utilisateur considère l'organisation de liens découverts sans son intervention. Selon cet ordre purement spatial on peut souvent rejoindre un même point par divers chemins. En présence d'un lien AB, l'algorithme de CORICO ne permet pas de décider si A est cause de B ou l'inverse. Et le calcul ne change pas si l'on remplace les noms réels des variables par des symboles abstraits tels X1, X2, ...Xm.

Etape 2 : Le choix de la réponse Y à expliquer et des variables "explicatives" est volontaire, et ne saurait être programmé d'avance car il dépend essentiellement du sens attribué aux noms des variables. C'est l'utilisateur qui détermine, parmi les variables liées à Y sur le schéma, quelles sont les candidates "explicatives". Et certes, il existe un risque de ranger parmi les causes (les régresseurs) ce qui n'est peut-être qu'une des manifestations de l'effet Y. Cependant un algorithme automatique ne peut nous remplacer ici. L'ordre de hiérarchie et de subordination diffère entièrement de l'ordre spatial du schéma. Au reste, on peut juxtaposer ces deux ordres au moyen de liens orientés (module ORIENT de CORICO).

Etape 3 : Les régresseurs une fois choisis, l'estimation des coefficients du modèle est automatique. Il reste à appliquer le modèle à un problème nouveau

8 - Conclusion

La méthode CORICO constitue une alternative aux méthodes ascendante, descendante ou mixte. L'accès aux interactions de type logique élargit le champ de recherche. Le pouvoir prédictif ne découle pas de l'orthogonalité des régresseurs mais de leur capacité à coller ou non à la réalité. Cela est précieux lorsqu'il est trop coûteux ou matériellement impossible d'orthogonaliser l'ensemble des variables d'un plan d'expérience à facteurs quantitatifs et qualitatifs. La méthode pourrait s'appliquer également aux réseaux neuronaux où se pose le problème du choix des variables d'entrée.

9 - Annexe

9.1 L'algorithme de CORICO (iconographie des corrélations).

Isolément, la valeur du coefficient de corrélation n'a guère de sens. Ainsi nous obtenons une forte corrélation entre le poids des élèves d'un collège et leur note à un exercice de mathématique (le même pour tous). En effet, les élèves de 12 ans pèsent moins que ceux de 17 ans, et leurs performances en mathématique sont moins bonnes dans l'ensemble. Il faut donc souligner le lien entre le poids et l'âge, et le lien entre l'âge et la note. Le lien entre le poids et la note est redondant, et peut être écarté. Tel est le principe de CORICO, fondé sur le calcul du coefficient de corrélation partiel "à âge constant". En présence de m variables, A,B,..M, une corrélation r(A,B) est ainsi contrôlée par rapport aux m-2 autres variables. Il convient aussi de connaître l'influence de chaque observation. Le plus simple consiste à retirer cette observation pour recalculer le coefficient de corrélation. Mais ce calcul est lourd : il faut l'effectuer pour chaque observation et chaque couple de variables. CORICO procède autrement. Soit un tableau de données de m variables (colonnes) et n observations (lignes). On définit n autres variables virtuelles, nommées "instants d'observation". Un "instant i" est une colonne toujours nulle sauf à l'instant de l'observation i (la ligne i). Le coefficient de corrélation partielle r(A,B) par rapport à cet "instant", noté ici r(A,B /i), est strictement égal à la corrélation r(A,B) en l'absence de l'observation i, mais les calculs d'ordinateurs sont plus rapides. En outre, la matrice des corrélations variables-instants possède des propriétés intéressantes qui sortent du cadre de cet article. D'où l'algorithme de CORICO : pour éviter les redondances, le lien AB est tracé si et seulement si r(A,B)>SEUIL en valeur absolue, et r(A,B /Z)>SEUIL en valeur absolue, à condition que r(A,B /Z) soit du signe de r(A,B), pour tout Z parmi les variables disponibles, y compris les "instants". La méthode revient à "arpenter" la sphère à n dimensions pour y dresser la carte des variables par triangulation, avec une précision trigonométrique. Il est donc inutile de calculer les corrélations partielles par rapport à 2 variables, 3 variables etc.. Cela est même inopportun : on doit questionner les témoins séparément pour éviter les influences. Au contraire, dans les techniques ascendantes ou descendantes, qui sont séquentielles, les tests successifs ne sont pas indépendants.

9.2 La représentation graphique

La simplicité de l'algorithme est contrebalancée par une difficulté de représentation quand les variables sont nombreuses. Le nuage de points à n dimensions peut être projeté dans un plan et les liens tracés. Mais le plus souvent on obtient un écheveau difficile à lire. Aussi, à l'origine, rien ne pouvait remplacer un tracé manuel exigeant des journées entières et force usage de la gomme. Avec le temps il est apparu que la sphère, qui ne privilégie aucune position, constitue un bon support puisque le coefficient de corrélation est un cosinus dans l'espace à n dimensions. La construction automatique décrite par Lesty et Buat-Ménard [1982] a été améliorée peu à peu sur des données diverses, et le programme dépasse maintenant les 30000 lignes de code FORTRAN. Il ne repose pas sur un algorithme unique, mais sur une réponse empirique à des cas de figure multiples. Il faut souligner que l'interprétation est découplée de la position des points (voir **figure 8**) et qu'on pourrait envisager d'autres supports que la sphère. C'est l'avantage de la

représentation schématique (et non projective) de pouvoir accepter une infinité de solutions, pratiquement aussi bonnes les unes que les autres.

En pratique une première variable A est dessinée n'importe où sur la sphère. On cherche la variable B la moins corrélée à cette première. Elle sera posée sur la sphère à la distance *arcCosinus(r(A,B))* de la première. On dessine de même, par triangulation, la variable C la moins corrélée aux deux premières. Les autres points sont posés de proche en proche. Si la quatrième variable a une corrélation nulle avec les trois premières, il n'est matériellement pas possible de lui assigner une position exacte. On recalcule ses distances de façon proportionelle aux valeurs réelles. Au bout d'un certains temps, la position des premiers points est recalculée d'après les suivants. Etc... Ainsi, la figure est réajustée progressivement.

Dans la mesure du possible, deux points seront donc d'autant plus proches sur la sphère qu'ils sont plus corrélés positivement. L'arcCosinus d'une corrélation négative est un angle obtus. Il est parfois intéressant d'utiliser la valeur absolue des coefficients (ceci explique en particulier la différence entre les figures 7 et 8).

9.3 Choix du seuil

Les tables de Fisher et Yates donnent, en fonction du nombre de degré de liberté, la probabilité que le coefficient de corrélation égale ou dépasse en valeur absolue une valeur donnée. Ces tables, fondées sur la distribution gaussienne, ne sont pas requises ici. CORICO en effet ne s'intéresse pas qu'aux relations « linéaires » (coefficient proche de 1), mais à toutes les valeurs du coefficient de corrélation (cosinus dans l'espace à n dimensions). Ce sont autant d'indices permettant de reconstituer l'organisation d'un ensemble de variables.

Lorsqu'une variable dépend à la fois de plusieurs variables indépendantes, Les corrélations avec chacune sont faibles, et "non-significatives" au sens des tables précitées. Ces dépendances sont pourtant "remarquables" au sens de CORICO et, si le seuil est assez bas, chacune est soulignée par un lien tracé.

Comme il n'y a pas de raison géométrique de rejeter un coefficient de corrélation, même nul, le seuil est choisi sur un critère de bon sens ou de clarté du schéma : un seuil trop bas, et le dessin est un réseau inextricable; un seuil trop haut, et la plupart de l'information est perdue. Remarquons que, même au seuil nul, beaucoup de variables ne sont pas liées dès qu'il existe une corrélation partielle de signe contraire à celui de la corrélation totale. Par exemple, même si R(A,B) = 0.99, la corrélation partielle entre A et B, par rapport à une variable Z = A+B, est négative. On observe donc les liens AZ et ZB, mais pas le lien AB. En fait, le schéma attire notre attention sur les liens qu'aucune autre variable ou "instant" ne peut "expliquer". Pour ne pas alourdir la figure, un "instant" n'est dessiné que s'il est lié à une variable au moins. Cet *instant remarquable* met en évidence un événement rare (un pic ou un creux inexpliqué de la variable) qui, parfois, a des conséquences importantes pour l'interprétation.

Quelquefois, on s'intéresse plus précisément à une variable, et l'on désire un seuil assez bas, compte tenu du phénomène étudié. Alors, si le tracé complet apparaît trop touffu, on focalise le schéma sur les seuls liens à cette variable.

9.4 Variables, observations, instants, propriétés

CORICO manipule quatre sortes d'entités : les "variables" (qui sont les colonnes du tableau), les "observations" (qui sont les lignes), les "instants d'observations" (qui sont des variables indicatrices des observations, cf. §9.1) et enfin les "propriétés des variables" (ci- après).

Exemple :

Tableau 13

6 observ.	5 variables					6 instants					
	FRANCE	PAYS-BAS	BELGIQUE	LUXEMBOURG	G.B.	caf	nes	Thé	bis	sou	pot
Café	88.	96.	94.	97.	27.	1	0	0	0	0	0
Nescafé	42.	62.	38.	61.	86.	0	1	0	0	0	0
Thé	63.	98.	48.	86.	99.	0	0	1	0	0	0
Biscuit	76.	62.	74.	79.	91.	0	0	0	1	0	0
Soupe	53.	67.	37.	73.	55.	0	0	0	0	1	0
Potage	11.	43.	25.	12.	76.	0	0	0	0	0	1

2 propriétés des variables

	FRANCE	PAYS-BAS	BELGIQUE	LUXEMBOURG	G.B.
Latitude	48.75	52.1	50.78	48.55	52.62
Longitude	2.00	5.18	4.35	7.62	0.00

Une "propriété" est donc une variable qui ne varie pas en fonction des observations mais en fonction des variables. Outre la Latitude et la Longitude, nous pourrions introduire la Population du pays, son Produit national brut etc.

Calculons les corrélations entre les colonnes du premier tableau :

Tableau 14

		FR 1	PB 2	BEL 3	LUX 4	GB 5	caf 1	nes 2	Thé 3	bis 4	sou 5	pot 6
FRANCE	1	100										
PAYS-BAS	2	75	100									
BELGIQUE	3	90	60	100								
LUXEMBOURG	4	95	84	75	100							
G.B.	5	-33	-25	-46	-25	100						
Café	1	59	56	78	47	-83	100					
Nescafé	2	-24	-21	-28	-11	25	-20	100				
Thé	3	14	61	-9	29	49	-20	-20	100			
Biscuit	4	37	-21	40	18	34	-20	-20	-20	100		
Soupe	5	-5	-10	-29	8	-32	-20	-20	-20	-20	100	
Potage	6	-80	-65	-52	-91	7	-20	-20	-20	-20	-20	100

Le triangle supérieur de la matrice contient les corrélations variables-variables

Le triangle inférieur droit comprend les corrélations instants-instants (sans intérêt pour l'analyse, il dépend seulement du nombre d'observations).

Le rectangle inférieur gauche constitue le tableau des corrélations variables-instants. La matrice des corrélations entre ces nouvelles colonnes égale la matrice des corrélations variables-variables.

Le tableau des corrélations variables-instants montre le profil des données sous une forme qui ne dépend pas des unités de mesure. Toutes les cases du tableau sont dans la même unité. On peut calculer les corrélations aussi bien entre les lignes qu'entre les colonnes.

Ce tableau est analogue à la matrice des données centrées réduites. Toutefois les corrélations variables-instants sont comprises entre -1 et + 1, alors que les valeurs centrées réduites, qui sont le rapport écart-à-la-moyenne / écart-type, peuvent sortir de cet intervalle. Soient x(i) la valeur initiale, y(i) la valeur centrée-réduite, et r(x,i) la corrélation variable-instant. On a: $x(i) = a\ y(i) + b = a'\ r(x,i) + b'$
(a et a' sont différents, mais b' = b).

Pour savoir si le profil des corrélations variables-variables (extrait du tableau 14) est influencé par les propriétés de ces variables, juxtaposons les propriétés au tableau des corrélations :

Tableau 15 :

		Corrélations					Propriétés	
		FR	PB	BEL	LUX	GB	Latitude	Longitude
FRANCE	1	**100**	75	90	95	-33	48.75	2.00
PAYS-BAS	2	75	**100**	60	84	-25	52.1	5.18
BELGIQUE	3	90	60	**100**	75	-46	50.78	4.35
LUXEMBOURG	4	95	84	75	**100**	-25	48.55	7.62
G.B.	5	-33	-25	-46	-25	**100**	52.62	0.00

Nous pouvons considérer ce tableau, comme un nouveau tableau de données à sept colonnes. Mais cette fois les 5 variables initiales sont représentées par leur profil de corrélations, donc sur 5 lignes au lieu de 6. Ce qui permet de calculer leurs corrélations avec les propriétés Latitude et Longitude qui présentent aussi 5 valeurs.

Si le profil de corrélation de A ressemble au profil de la propriété P, c'est que plus une variable est corrélée à A (c'est à dire plus leur profil de variation se ressemblent), plus sa propriété P est forte. P "explique" les corrélations avec A.

Soit R la matrice de corrélations entre les colonnes du tableau 15. L'algorithme de CORICO, ébauché plus haut, est donc complété, en présence de propriétés, par les règles suivantes : pour éviter les redondances,

- le lien AP est *tracé* si et seulement si R(A,P)>SEUIL en valeur absolue, et R(A,P /Z)>SEUIL en valeur absolue, à condition que R(A,P /Z) soit du signe de R(A,P), pour tout Z parmi les "variables" disponibles dans le tableau 15, y compris les propriétés.

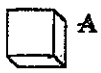
Dans le tableau 13, les propriétés sont des lignes. On peut aussi considérer les lignes du tableau de corrélations variables-instants (extrait du tableau 14) comme des propriétés des variables. Il suffit de les transposer en colonnes que l'on juxtapose au tableau 15.

En résumé : il n'est pas possible de corréler directement les observations aux variables puisque les unes sont des lignes et les autres des colonnes. En revanche, les *noms* d'observations peuvent apparaître sur la figure,

- en tant qu'instants,
- en tant que propriétés.

Les propriétés, comme les instants, sont placées en éléments supplémentaires sur la sphère. Pour ne pas alourdir le schéma, elles n'apparaissent que si elles sont remarquables, c'est-à-dire liées à une variable au moins.

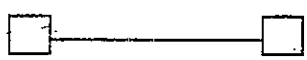
9.5 Conventions de représentation



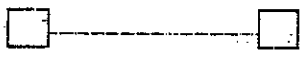
Chaque variable (colonne du tableau) est représentée par un CUBE.

Un pic (ou creux) remarquable dans le profil de variation d'une variable est représenté par un TRIANGLE. Ce pic (ou creux), qui n'est pas forcément une valeur extrême, est dit « remarquable » car non-entièrement explicable par les variations des autres variables disponibles. Il indique un événement rare ou atypique.

Une éventuelle propriété remarquable de la variable A est représentée par un PETIT CUBE OBLIQUE. Elle "explique" les ressemblances des autres variables avec A.



Un TRAIT PLEIN annonce une ressemblance entre deux variables (elles croissent et décroissent ensemble). Cette ressemblance est « remarquable » car non-entièrement explicable par les variations des autres variables disponibles



Un TRAIT POINTILLE reflète une symétrie remarquable de variation (Chaque variable croît quand l'autre décroît).

Exemple :

Tableau 16 :
Répartition des dirigeants de PME selon leur formation en fonction du secteur

	INDUSTRIE	COMMERCE	TRANSPORT	TOURISME	BTP	SERVICES
technique	46	27	25	21	51	30
gestion	10	7	15	6	8	6
commerciale	8	18	3	13	6	8
comptable	6	10	8	4	4	5
juridique	1	2	1	4	1	4
scientifique	3	3	0	3	2	22
générale	10	12	19	14	8	13
autodidacte	16	21	29	35	20	12

Source : Les PME et leurs dirigeants, Crédits d'équipement des PME.

D'après les conventions précédentes, la **figure 19** se lit comme suit :

- Les patrons de l'industrie et des BTP ont un profil de formation très semblable, particulièrement technique (petit cube = "propriété" remarquable)
- Dans les services, les "propriétés" remarquables sont la formation scientifique et (trait pointillé) le caractère non-autodidacte.

- Dans le secteur du tourisme, les patrons ressemblent d'une part à ceux du commerce et, d'autre part, à ceux du transport. Ils présentent une "propriété" autodidacte remarquable avec, en outre, une valeur remarquable pour cette formation autodidacte (c'est d'ailleurs un *pic* dans la ligne et dans la colonne)
- Les patrons du commerce présentent une "propriété" commerciale remarquable.
- Dans les transports il y a deux "propriétés" remarquables : l'une (positive) concerne la gestion, l'autre (négative) concerne la formation scientifique.

La valeur remarquable n'est pas forcément la plus forte ou plus faible valeur. En effet TOUS les secteurs présentent de relativement fortes proportions de techniciens et d'autodidactes. Mais les transports sont les seuls à être si souvent gestionnaires et si peu scientifiques, les services les seuls à présenter tant de scientifiques et si peu d'autodidactes, et le commerce le seul à présenter autant de commerciaux. Enfin, les formations comptables, juridiques et générales ne sont pas remarquées.

9.6 Limites de CORICO

a - Un lien n'est tracé qu'après une sévère série de vérifications (le calcul de toutes les corrélations partielles). Il est rare qu'il faille alors l'attribuer à une pure coïncidence, et c'est souvent un moyen a posteriori de justifier du caractère non bruité d'une série de mesures. Toutefois il ne faut jamais écarter la possibilité d'une coïncidence.

b - En présence d'un lien AB sur le schéma, on peut toujours imaginer une autre variable X, non donnée, qui viendrait s'intercaler entre A et B. Un grand nombre de variables en rapport avec le problème ne peut donc qu'affiner l'interprétation. On gagne du temps à les introduire dès la première analyse, quitte à éliminer ensuite les paramètres non pertinents.

c - CORICO se restreint aux interactions de type f(A,B). En effet l'introduction d'interactions du type f(A,B,C), même en se contentant des opérations logiques, conduit à un trop grand nombre de combinaisons et soulève des difficultés de programmation. Cependant les interactions d'ordre 2 constituent déjà un bon débroussaillage.

Remerciements

L'auteur remercie Pierre Cazes, Claude Lesty et Arnauld Ménager qui lui ont suggéré des modifications et signalé des erreurs.

Références bibliographiques

CNEXO Investigation Générale de données recueillies par le Réseau National d'Observation de la Qualité du Milieu Marin. - *Contrat n°82-2709 - Juin 1982.*

Guide de l'utilisateur de CORICO (1995), tomes 1 et 2, Coryent, Versailles

Jackson J.E. (1991) A user's guide to principal components,Wiley, New York.

Kujas M., Pléau-Varet J. and Lesty C. (1996) Assessment of Recurrence related parameters into pituitary adenoma series recurrent vs non recurrent, using multiparametric analysis CORICO software. *Electronic Journal of Pathology and Histology.*

Lesty M. et Buat-Ménard P. (1982) La Synthèse Géométrique des Corrélations Multidimensionnelles (CORICO) *Les Cahiers de l'Analyse des données*, Vol VII, n°3, pp 355-370.

Lesty M. et Coindoz M. (1988) Une méthode pour la F.M.S. des bases de connaissances de systèmes experts. Une application de CORICO. *6 ième Colloque International de Fiabilité et de Maintenabilité. Strasbourg.*

Lesty C., Chleq C., Contesso G., Jacquillat C. (1992) Nucleoli and AgNOR Proteins in 32 Cases of Primary Breast Carcinoma. Spatial Pattern of Interactions Between 50 Clinical and Histometric Criteria. *Analytical and Quantitative Cytology and Histology, vol. 14 N°3.*

Malpéalat-Domaine C. (1986) Les Invaginations Intestinales Aigües du Nourrisson et de l'Enfant, à propos de 65 cas. *Thèse de Doctorat de médecine. Paris VII, Faculté de Médecine Xavier-Bichat.*

Rolland X. (1988) L'Enzyme Linked Immunoelectrotransfer Blotting (EITB): un outil dans la lutte contre la Leishmaniose viscérale. *Thèse - Université de Technologie de Compiegne*

Tenenhaus M., Gauchi J.P., et Ménardo C. (1995): Régression PLS et Applications. *Revue de Statistique Appliquée*, vol 42 n°1, pp 7-63

Tomassone R., Lesquoy E., Millier C. (1983) La régression, nouveaux regards sur une ancienne méthode statistique. Masson.

Séchet E. (1991) Analyse Multicritère de Découpes Pyrotechniques par Cordeaux. Iconographie des Corrélations (CORICO) *Stage I.M.A. 4 à la société Aérospatiale. Institut de Mathématiques Appliquées, Université Catholique de l'Ouest, Angers.*

Figure 3 : Interactions entre variables orthogonales A et B

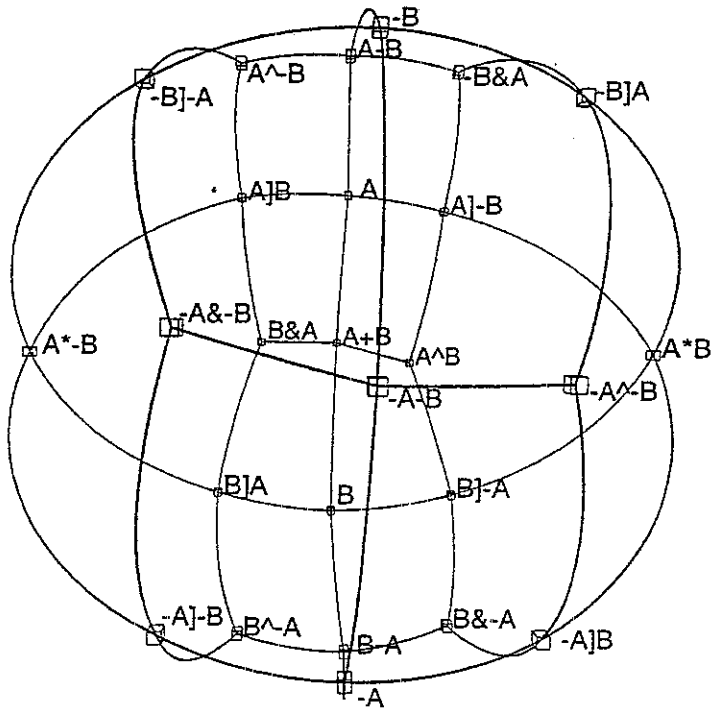


Figure 4 : Interactions entre variables colinéaires (B = A)

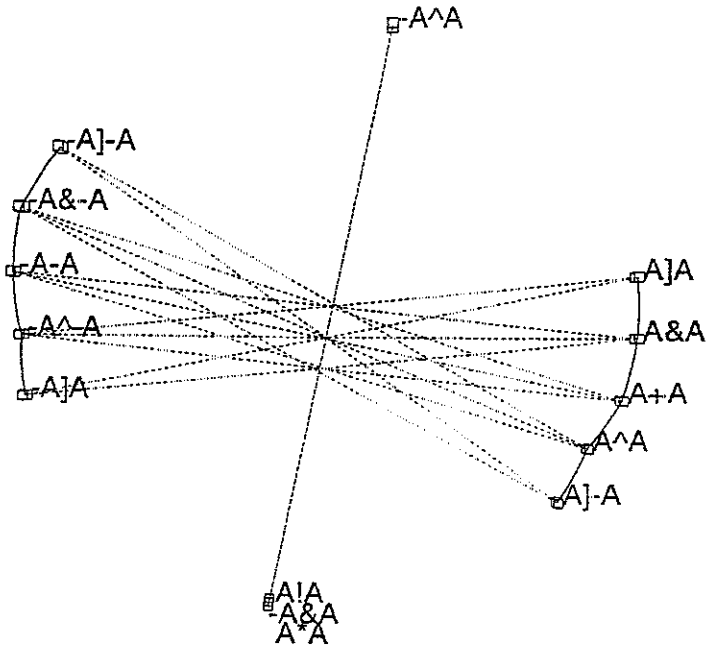


Figure 5 : Performances en gymnastique

Les performances en FLEXIONS sont liées positivement aux performances en SAUT et en TRACTIONS, et liées négativement (pointillé) au tour de taille. Le pouls n'a pas d'influence remarquable sur les performances.

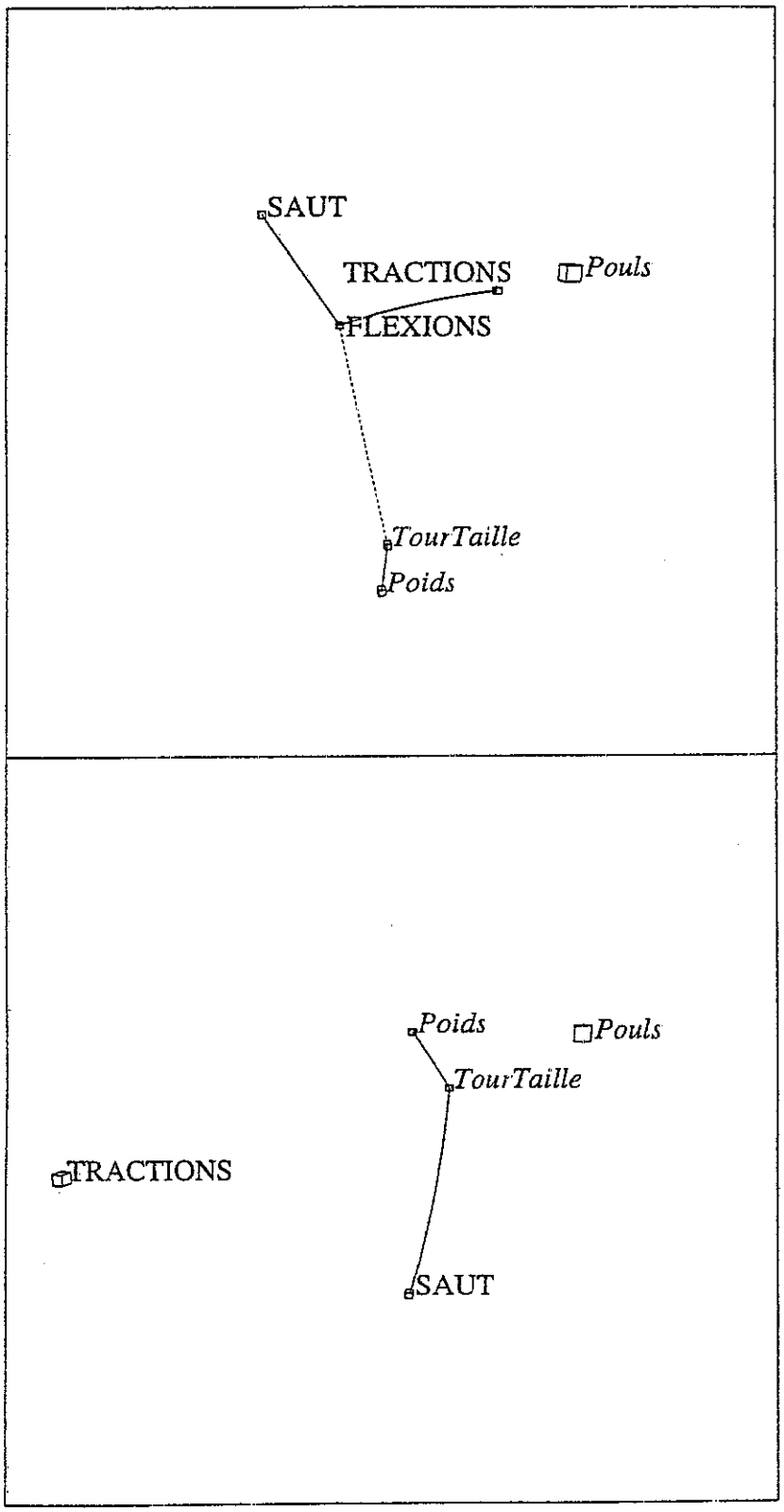
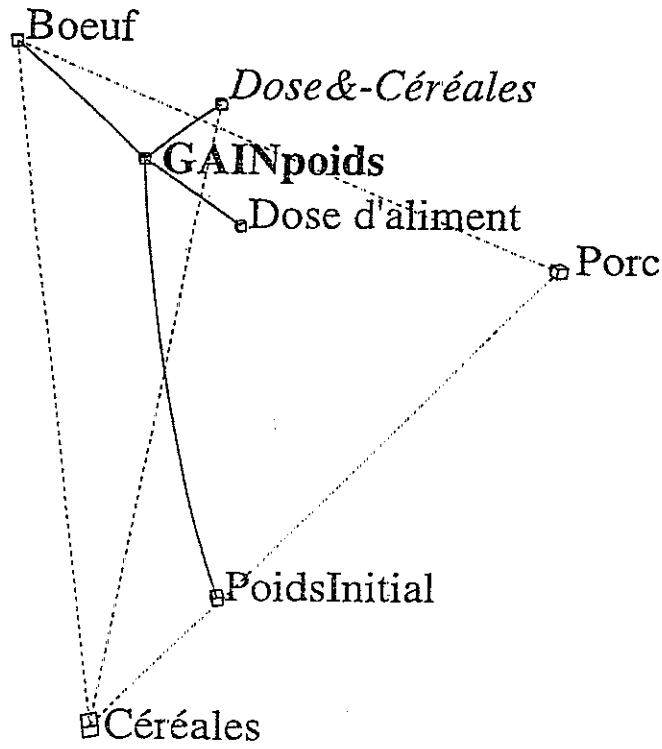


Figure 6 : Retrait de la composante de FLEXION

Une composante secondaire du SAUT apparaît, liée positivement au tour de taille

Figure 7 : Gains de poids des animaux

Traits PLEINS = Corrélations positives; Traits POINTILLES = Corrélations négatives.
Les aliments Boeuf, Porc et Céréales sont mutuellement exclusifs par définition.
Le gain de poids est fonction du poids initial et de la dose d'aliment. Il est favorisé par le Boeuf, ou, plus généralement, par une forte dose de protéines animales (non-céréales).



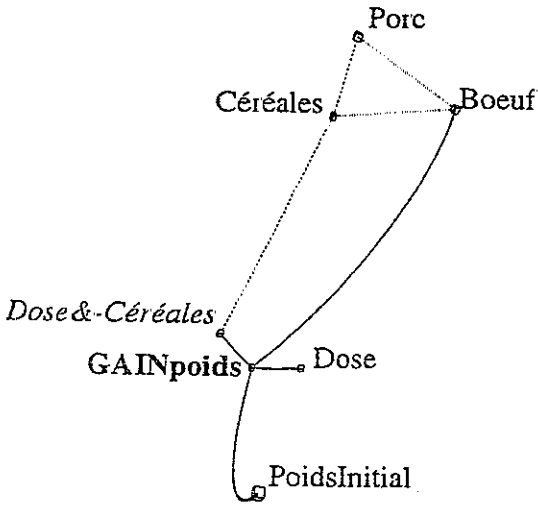
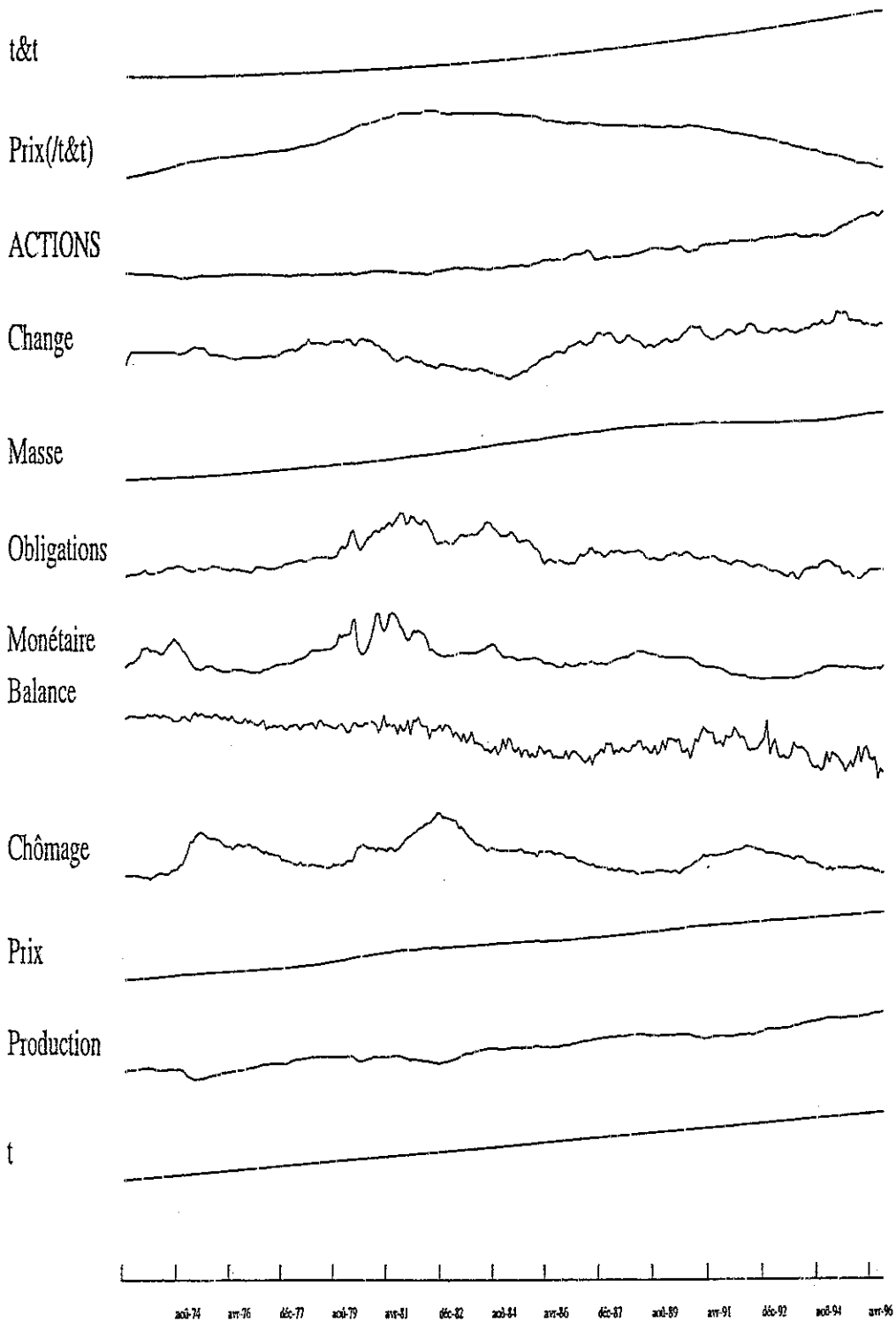


Figure 8 :

Strictement équivalente à la figure 7 quant aux liens, cette figure s'interprète de la même façon. Dans CORICO, le calcul des liens et celui des positions sont découplés. On peut jouer sur la position pour améliorer la lisibilité.

Exemple : sur la figure 7, le lien négatif Porc - céréales passe par hasard trop près de PoidsInitial. La figure 8 lève l'ambiguïté.

Figure 9 : quelques indicateurs économiques



Toutes les variables sont centrées réduites. En haut se trouve "t&t", effet accéléré du temps, détecté par CORICO comme très semblable à ACTIONS. Il est difficile d'apercevoir à l'oeil nu les diverses composantes de ces courbes. Comparer par exemple Prix, 3 ième courbe à partir du bas avec cette même variable une fois retirée la composante t&t, deuxième à partir du haut.

Figure 10 : Indice du marché des actions américain (1973-96)

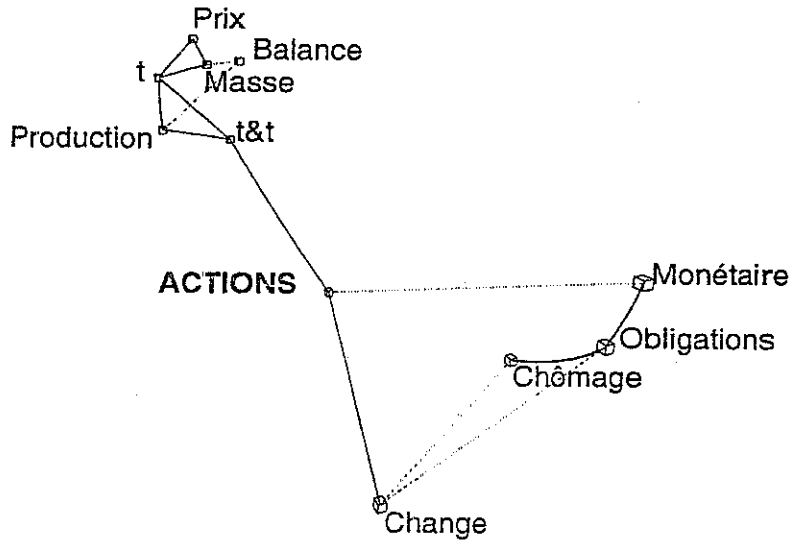


Figure 11 : Retrait de la composante t&t

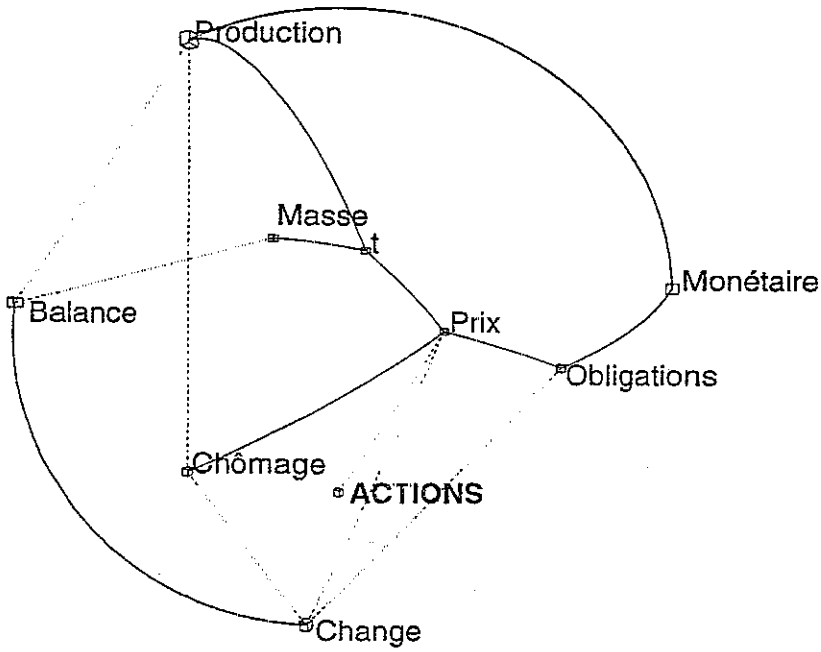
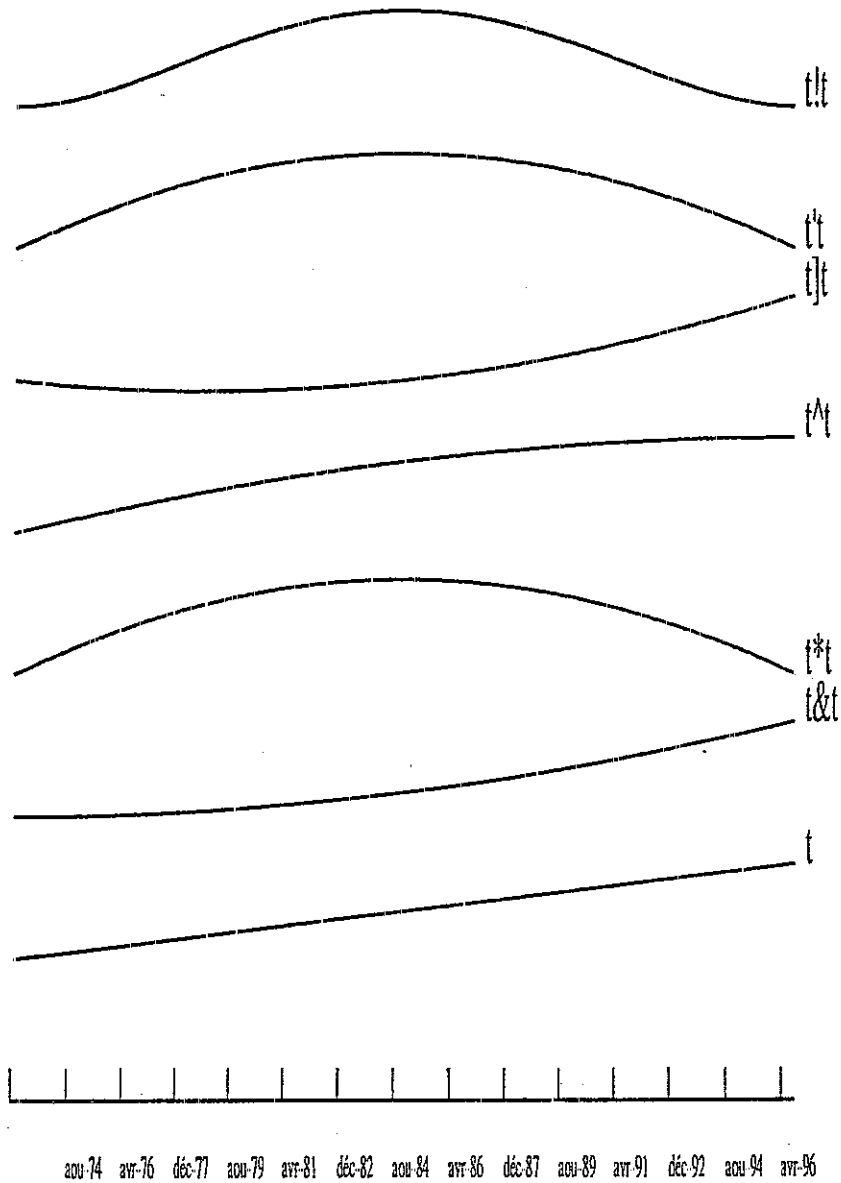


Figure 12 : quelques interactions entre variables colinéaires



Graphe des interactions f(A,B) quand A et B sont parfaitement corrélés (ici A = B = t)

Figure 13 : facteurs liés au Chômage (1973-96)

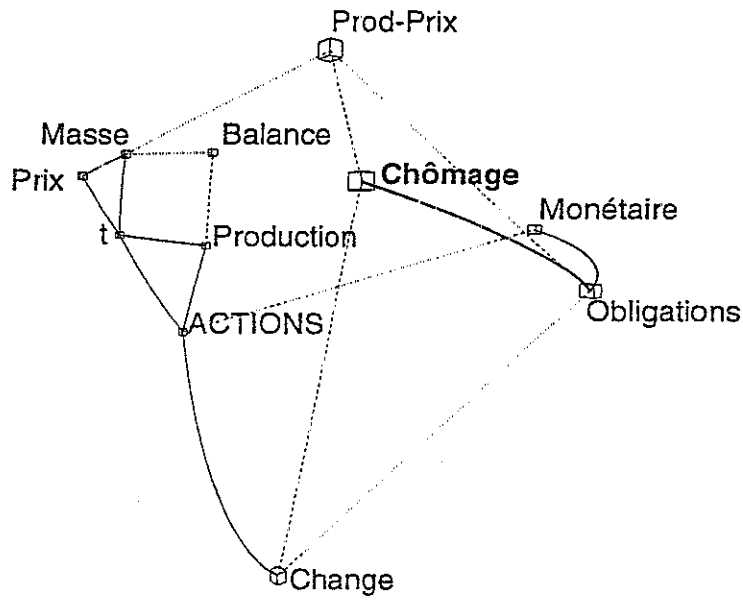


Figure 14 : Retrait de la composante Prod-Prix

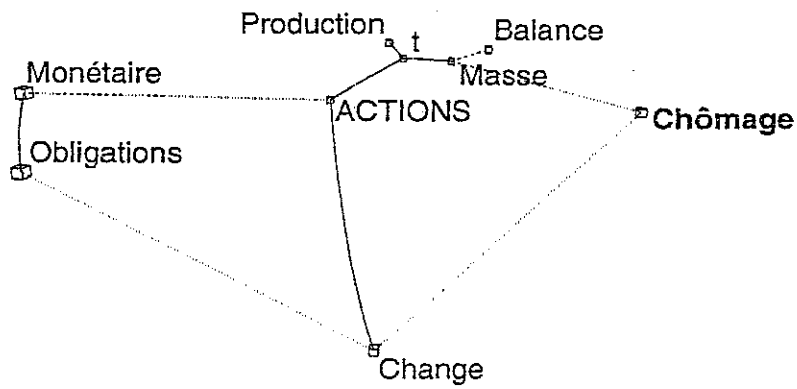


Figure 15 : Indice du marché des actions américain, jan-73 à dec-84

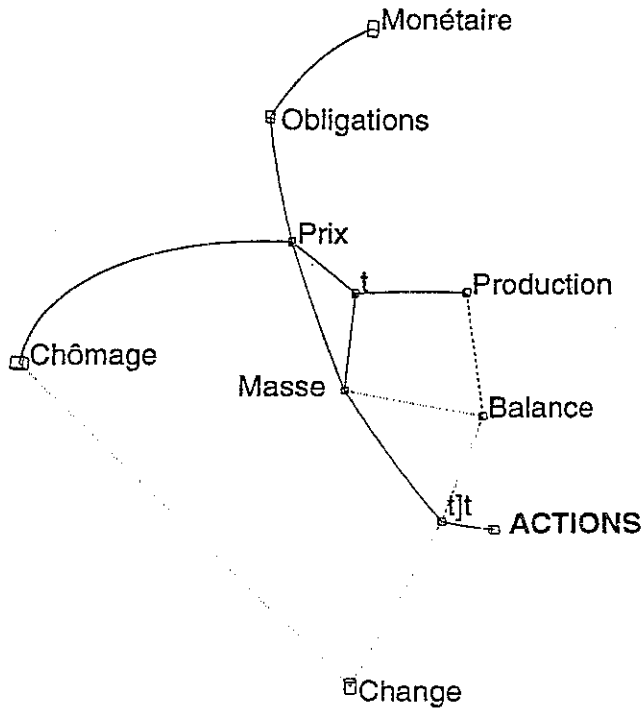


Figure 16 : Retrait de la composante t]t

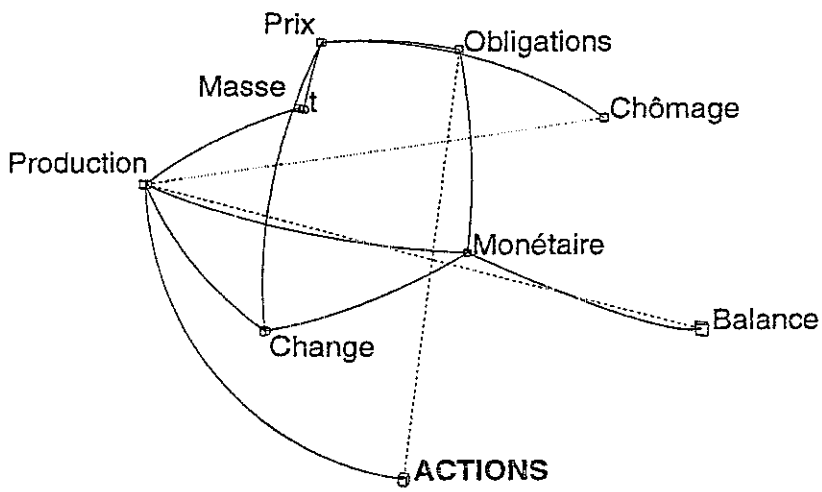


Figure 17 : Liens avec le chômage, jan-73 à déc-84

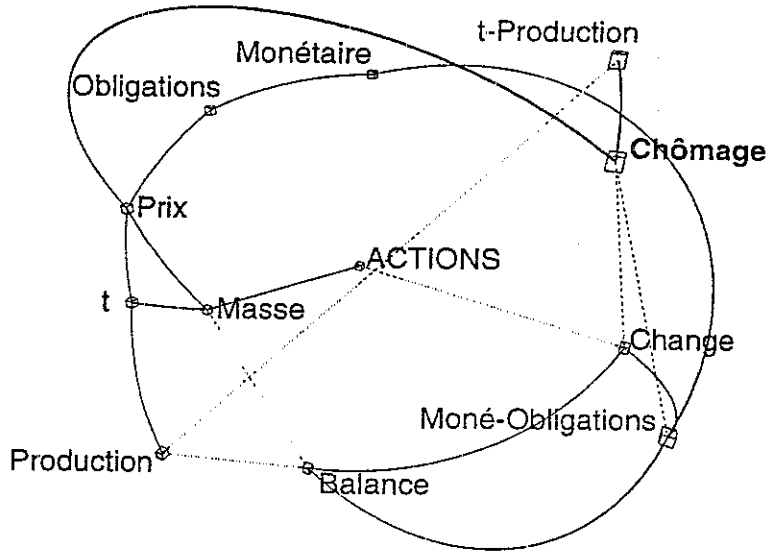


Figure 18 : Retrait de la composante t-ProdIndus

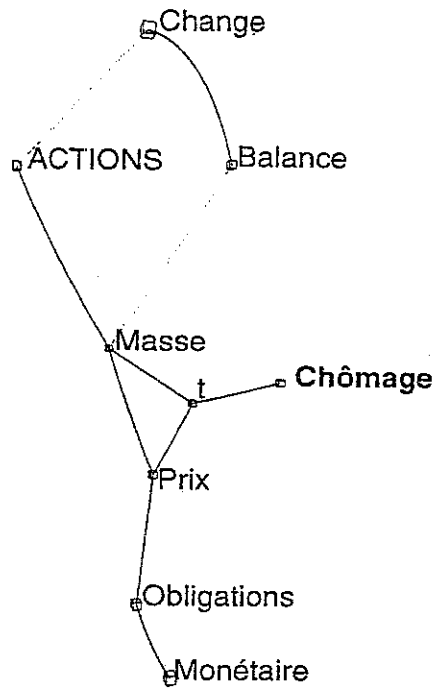


Figure 19 : Dirigeants de PME selon leurs formations et secteurs