EXPLICATIONS D'UN TABLEAU PAR UN AUTRE

LE PROGRAMME RESDIF

Roger LAFOSSE

Laboratoire de Statistique et Probabilités - Université Paul-SABATIER
118 route de Narbonne - 31062 Toulouse Cédex

RESUME: Les deux tableaux à analyser sont formés des mêmes mesures effectuées sur de mêmes individus. Ces tableaux peuvent être relatifs à deux époques différentes (données temporelles), ou relatifs à deux juges (données sensorielles). Deux analyses sont mises en œuvre, décomposant la variance totale de l'un des deux tableaux en parts expliquées et en parts non expliquées par l'autre tableau. Ainsi, une première explication correspond à une reconnaissance de forme de deux nuages de points-individus, à une étude de la ressemblance (puis de la différence) entre deux époques ou entre deux juges. Dans une deuxième analyse, les variables explicatives des parts expliquées sont les composantes principales de variables instrumentales, définies par Rao en 1964 (A.C.P.V.I.). Le programme met en œuvre les deux approches, la comparaison de leurs résultats pouvant enrichir l'étude.

Mots clés: Analyse en composantes principales, analyse canonique, régression linéaire simple, analyse en composantes principales de variables instrumentales, rotation procruste.

1. Position du problème

L'analyse canonique de deux tableaux et l'A.C.P.V.I. (analyse en composantes principales de variables instrumentales, Rao, 1964) sont deux analyses connues, adaptées à une étude comparative de deux paquets de variables quantitatives. Les deux jeux de variables doivent être mesurés sur de mêmes individus, mais ne sont pas forcément formés des mêmes variables. *En fait, même s'ils l'étaient, ces analyses ne tiennent pas compte de cette nature particulière des données. En effet, l'analyse canonique est une analyse entre deux espaces de variables engendrés; l'A.C.P.V.I. est une analyse explicative d'un tableau par un espace engendré.*

Pour nous, non seulement les individus de même nom occupent le même numéro de ligne dans l'un et l'autre tableau, mais les variables aussi, pour les colonnes. On dit que les individus sont appariés, et les variables aussi. Les deux tableaux sont dits totalement appariés.

La rotation procruste consiste à comparer deux nuages centrés de points-individus appariés. Pour cela, on regarde si l'un ressemble à l'autre pris pour cible, après l'avoir soumis à une rotation visant à le superposer à la cible. Le critère utilisé est celui des moindres carrés entre points appariés. Les résultats sont évidemment inchangés si, au lieu des tableaux initiaux, on avait transformé ces tableaux par isométries pour faire tourner les nuages autour de leur centre gravité commun. Cette approche est surtout intéressante sur des tableaux totalement appariés.

L'analyse de communité a été définie par Lafosse (1989). C'est l'analyse rappelée dans cet exposé, la mise en œuvre ayant conduit à des développements nouveaux. Le néologisme "communité" est adopté pour désigner ce qu'un tableau peut avoir en commun avec un autre. Plus précisément, c'est une façon de désigner une part expliquée contenue dans un tableau, l'explication provenant de l'autre tableau. Dans cette explication, on veut que chaque variable d'un tableau soit expliquée par la variable correspondante dans l'autre tableau. Mais cela n'est pas vraiment simple: chaque variable est liée aux autres dans chacun des deux tableaux. En plus ces liens ne sont pas identiques dans les deux tableaux. L'analyse est basée sur une simplification préliminaire de ces liens.

En A.C.P.V.I., une variable d'un tableau est expliquée par n'importe quelle variable de l'autre tableau. Le prolongement donné par Johansson à l'A.C.P.V.I. nous intéresse tout particulièrement. Il consiste à remarquer que toute l'explication issue d'une composante explicative trouvée dans un tableau est contenue dans une composante expliquée correspondante dans l'autre tableau.

On pourrait dire que l'analyse de communité est une analyse procrustéenne sur tableaux totalement appariés, où le classement des composantes principales a été réalisé selon un critère de variance expliquée maximale. Cette variance expliquée est obtenue par régressions linéaires simples. L'appariement des deux variables d'une régression est total. Ces régressions sont donc bien particulières, et adaptées à notre propos.

Cette analyse, vue comme une étude de ressemblance entre tableaux, se prolonge naturellement par une étude de différence. Cette étude de différence est moins simple à aborder que l'étude faite par l'analyse en composantes principales (A.C.P.) du tableau résultat de la différence arithmétique des deux tableaux.
Mais cette dernière A.C.P. ne permet pas de dégager une part ressemblante complémentaire de la différence considérée. En communité, la variance totale d'un tableau est décomposée en deux parties; l'une est associée à la ressemblance, l'autre à la différence.
Dans l'A.C.P. indiquée ci-dessus, on mélange les différences entre valeurs dues à des changements d'échelle, avec des différences dues à des changements dans les positions relatives des individus. Pour éviter cela, pourrait-on alors plutôt envisager l'A.C.P. de la différence arithmétique, après avoir réduit les variables des deux tableaux? Mais ce serait

admettre que les changements en variance peuvent s'aborder variable par variable, indépendamment des liens existant entre ces variables. L'analyse de communité vise à surmonter de telles difficultés.

Quand aucune confusion n'est possible, l'analyse de communité désigne à la fois l'analyse de ressemblance et celle complémentaire de différence.
Cependant, les résultats relatifs à la différence entre tableaux peuvent changer beaucoup si on échange les rôles des deux tableaux, alors que ceux obtenus pour analyser la ressemblance sont inchangés.

L'analyse de communité ne peut se confondre à celle de Johansson-Rao (nous dirons brièvement A.C.P.V.I.), car l'échange des deux tableaux correspond vraiment à deux analyses totalement différentes en A.C.P.V.I.

Supposons que les deux tableaux soient associés à deux époques différentes. Quelles sont alors les variables ayant le moins changé dans le temps? Y a-t-il des individus qui ont une évolution temporelle hors du commun? La prise en compte de l'information sur l'évolution temporelle des liens améliore-t-elle l'explication du présent par le passé, explication qui est donnée en A.C.P.V.I. sans cette prise en compte?

Si deux juges utilisent les mêmes critères pour évaluer les mêmes objets, quels sont les critères les plus communs aux deux juges, hors notion de moyenne et d'échelle? les plus différents? quels sont alors les objets les plus jugés de manière semblable? Et si on supposait, qu'en fait un juge utilise ses propres critères et l'autre les siens, les réponses aux questions précédentes en seraient-elles modifiées?

L'analyse de communité fournit des réponses à ces questions, en partie en confrontant ses résultats à ceux d'une A.C.P.V.I.

Dans **RESDIF-FORTRAN** trois modes de fonctionnement sont proposés à l'utilisateur: analyse de communité seule, A.C.P.V.I. seule, ou les deux analyses successivement, avec une comparaison de résultats (options 1,2 ou 3).
Il est aussi possible, ayant modifié l'un des tableaux à lire en entrée, de l'utiliser pour effectuer l'analyse canonique

2. Le programme

Le programme a été écrit en language Matlab (sur Mac II) et en language Fortran (version 1.1, sur PC).
Dans ces écritures, il est proposé en conversationnel. Il est non interactif en Fortran, et peu en Matlab.
Dans la version 2.3 du logiciel suisse hautement interactif EDA du Pr. E. Horber, l'analyse de communité et l'A.C.P.V.I. sont l'objet de la commande "resd".
Une écriture en S+ est prévue.

En FORTRAN, il est proposé aussi sous forme d'un programme batch, paramétrable et intégrable aux logiciels permettant des rajouts de programmes fortran. La version batch actuelle du programme est issue directement de la version conversationnelle. Elle n'a pas été écrite toujours de façon optimale quant à la place occupée en mémoire, et pour cette raison l'utilisateur de petit matériel peut se voir limité, quant à la taille des tableaux, plus vite qu'il ne pouvait le penser.

A priori, le programme RESDIF-FORTRAN est conçu pour un maximum de 999 individus et 26 variables. Ces valeurs peuvent être poussées plus loin pour les calculs. Si les graphiques sont quand même demandés: l'individu 2567 sera codé comme l'individu 567; les variables venant à la suite de la 26-ième variable Z, seront codées par les caractères ascii qui suivent habituellement Z. Les sorties des graphiques se font dans des carrés, une échelle commune étant prise sur les deux axes. La taille des carrés peut être choisie petite (2 pouces) ou plus grande (8 pouces ou 21 cm.). Une lecture à l'écran, donne souvent une image déformée, le carré apparaissant rectangulaire.

Dans un fichier "lisez-moi", la disquette-PC de RESDIF-FORTRAN contient les indications sur les autres fichiers: sources batch et conversationnel, exécutables batch et conversationnel, de paramètres, des données test, des résultats relatifs à ces données. La disquette fait partie de la bibliothèque du club Modulad, ou est disponible en s'adressant à l'auteur.

Nous avons choisi pour cet article le matériel et le logiciel qui nous offraient le plus de souplesse et les sorties les plus performantes, donc Matlab. Mais l'utilisateur du programme Fortran n'aura pas de difficulté à rapprocher les sorties de cet article avec les siennes. Les quelques spécificités du programme **RESDIF-FORTRAN** (RES comme ressemblance, DIF comme différence), sont indiquées dans chaque paragraphe.

3. Présentation de l'exemple et choix du tableau à expliquer

On illustre d'un exemple la présentation de la méthode.

Les p=7 analyses médicales considérées ont été mesurées en 1979, puis en 1984, sur les mêmes n=500 patients extraits d'un fichier d'étude prospective de la Sécurité Sociale, initialement formé de 26 analyses médicales. Ces patients de plus de 40 ans, qui ne sont pas malades a priori, sont des salariés ou des retraités ayant droit à un bilan médical gratuit de contrôle tous les cinq ans.
Les analyses médicales choisies, l'ont été parce qu'elles sont assez familières au grand public dont l'auteur fait partie, et non parce que ce sous-groupe est particulièrement judicieux d'un point de vue médical.

L'analyse de communité ne fait pas complètement jouer un même rôle aux deux tableaux. Quand on compare un objet à un autre objet pris pour référence, on a un certain discours. Ce discours n'aurait pas été tout à fait le même si le premier objet avait été pris pour référence. Il n'en est pas faux pour autant, mêmes si les deux discours possibles ne sont réellement identiques que si les objets le sont. Le choix du tableau à expliquer peut ainsi être arbitraire.
Ici le choix est facile. C'est le tableau passé qui est susceptible d'expliquer le présent, et donc le tableau X_2 sera toujours dans cet article le tableau expliqué.

On parlera de ce qui sépare les deux tableaux en terme d'évolution explicable et non explicable, de l'époque 1 à l'époque 2, de l'année 1979 à l'année 1984. Pour des données sensorielles, on se serait exprimé en terme de ressemblance et de différence entre les deux tableaux.

L'étude faite a pour objet de montrer quel parti peut être tiré de la méthode elle-même. Autrement, par exemple, on aurait pu analyser en quoi l'évolution temporelle diffère selon le sexe, en comparant les résultats issus d'une analyse effectuée sur les hommes avec ceux issus d'une analyse faite sur les femmes.

Dans un premier temps, on se sert de la technique en vue de désigner les analyses médicales manifestement les plus stables dans le temps: celles dont les liens aux autres se sont le moins modifiés. A contrario, on veut connaître celles qui ont évolué de façon sensible. On détecte les patients qui ont un comportement exceptionnel parce qu'ils

diffèrent fortement du comportement moyen aux deux époques, et d'autres parce que leur évolution dans le temps s'écarte trop de l'évolution moyenne, selon certaines analyses.

Les analyses médicales sont les suivantes:
Calcium (Ca), Cholestérol (Ch), Tension Artérielle maximum (TA), Acide Urique (AU), Protides (Pr), Sodium (So), et Poids (Pd).

Exemple de mesures relevées sur un patient:
Ca=2,35 Ch=4,4 TA=11 AU=226 Pr=65 So=137 Pd=76.

X_2 est le tableau des observations pour l'année 1984. X_1 est celui pour l'année 1979. Les 7 variables de X_1 (analyses médicales, en colonnes) sont centrées, ainsi que celles de X_2.

La question de savoir si une variable a évolué en moyenne de façon significative est abordée en Section 9.

Le programme **RESDIF-FORTRAN** demande en entrée le nom du fichier du tableau à expliquer (tableau nommé **X1 dans le programme**), puis celui du tableau explicatif. Il sort les lignes 1 et n du tableau à expliquer, les moyennes et variances initiales des variables des deux tableaux.

4. Problèmes de standardisation

4.1 Réductions

On peut ne pas vouloir réduire les variables avant analyse. C'est que les variables seraient homogènes, relatives à un instrument de mesure commun. Sinon, comme en A.C.P., on attribue aux variables un même poids avant d'effectuer l'analyse.
La réduction des variables appliquée sur chaque tableau séparément serait ici maladroite. Nous citons ci-dessous trois modes de réduction répondant en fait à trois analyses différentes

Supposons que l'on attribue à chacune des variables de X_2 un poids identique, en **réduisant les variables de X_2**. Ensuite usons des coefficients calculés pour opérer cette réduction, en effectuant les mêmes changements d'échelle sur les variables correspondantes respectivement dans X_1, de façon à conserver l'information sur l'évolution des variances. Alors, il s'agit d'expliquer le jeu des corrélations des variables du présent, à partir du passé.

Autrement, on pourrait plutôt **réduire X_1**, puis user des changements d'échelle sur les variables correspondantes de X_2; alors il s'agirait d'appréhender l'évolution temporelle à partir du seul jeu des corrélations internes au tableau X_1.

Sur notre exemple, nous décidons de réduire les variables en évitant de restreindre l'étude de l'évolution temporelle à des corrélations, explicatives ou à expliquer. Pour cela, on divise chacune des variables de X_1, et celle qui lui correspond dans X_2, par la racine carrée de la moyenne arithmétique de leurs variances. Cela revient à réduire les p variables du tableau de 2n individus, obtenu par **concaténation** des tableaux centrés X_1 et X_2. Il s'agira ainsi d'expliquer la variabilité de X_2 à partir d'une évolution temporelle des variances et des corrélations. Notons que les variables de X_2 possédant alors la variabilité la plus élevée sont celles dont la variance s'est accrue entre 1979 et 1984.

Les nouveaux tableaux centrés et réduits sont encore nommés X_2 et X_1 et sont ceux considérés par la suite.

Il n'est pas interdit de réduire avec des étendues interquartiles, par exemple, plutôt que des écarts-type. Il y a alors une perte d'homogénéité avec les outils intervenant dans la méthode, mais cela permet d'attribuer aux variables des poids pouvant être jugés intéressants parce que fonction d'une population médiane. Les résultats en seraient un peu changés et pourraient être comparés à ceux issus d'un mode de réduction plus classique. Diviser par les moyennes géométriques des écarts-type des variables est une autre procédure par concaténation qui pourrait être appliquée.
La richesse de l'outil exploratoire peut ainsi provenir plus de l'utilisateur que de la méthode elle-même.

Dans **RESDIF-FORTRAN**, trois options paramétrables sont proposées: pas de réduction, réduction par concaténation (moyenne arithmétique) des tableaux, et réduction de X_1 seul avec changements d'échelle correspondants sur X_2.

4.2 Transformations

En fait, et cela n'est pas étranger au problème de la standardisation visant à attribuer un même poids aux variables, avant de centrer et réduire il peut être conseillé de transformer certaines variables, pour les ramener toutes à un même type de distribution.

Le plus souvent, il s'agit de déterminer les variables de l'un des deux tableaux (X_2 en l'occurence), qui possédent des distributions très asymétriques, comparées aux distributions des autres variables, les transformations étant choisies façon à leur donner une allure sensiblement plus gaussienne. Les outils mis en œuvre dans l'analyse, moyennes, variances et corrélations, sont alors dans un cadre idéal d'utilisation.

Pour les analyses choisies, toutes les distributions ont une allure suffisamment gaussienne pour rendre la démarche inutile: la modification apportée serait négligeable sur les résultats de l'analyse.

5. L'analyse de communité

5.1 Variables et axes principaux de l'analyse

Soit $\{e_i\}$, i=1,...,p une base orthonormée de R^p.
Un vecteur e est noté e dans sa représentation matricielle en colonne, ${}^te$ dans celle en ligne. Soit I la métrique identité de R^p, dont la matrice est de dimension p.
On a l'écriture matricielle:

$$I = \Sigma\, e_i\, {}^te_i \quad \text{et} \quad X_2 = X_2\, I = \Sigma\, X_2(e_i)\, {}^te_i \qquad (1)$$

Soit D la matrice diagonale des poids affectés aux patients: la matrice unité dans R^n multipliée par 1/n.
L'A.C.P. du triplet (X_2, I, D) revient à définir une décomposition (exacte ou approchée) de X_2 du type (1), si on prend pour vecteurs e_i les axes principaux et pour vecteurs $X_2(e_i)$ les composantes principales, i = 1,...,q ($q \leq p$). Ici, une autre base nous intéresse, sachant que, quelle que soit la base orthonormée, la variance totale de X_2 vaut Σ var$[X_2(e_i)]$.

Notons $V_{21} = {}^tX_2DX_1$ la matrice des covariances des variables de X_2 avec celles de X_1 et u_{2i} les vecteurs propres orthonormés de la matrice symétrique $V_{21}\,{}^tV_{21}$, associés aux valeurs propres non nulles.

On note R la rotation procruste vers X_2, à savoir

$$R = {}^tV_{21}[V_{21}\,{}^tV_{21}]^{-1/2},$$

mettant en correspondance, sans la modifier, la structure intracovariante de X_1 avec celle de X_2:

si on note x_{2j} la variable en position j dans X_2 et x_{1j} celle en position j dans X_1R, on a

$$\text{cov}(x_{2j}, x_{1i}) = \text{cov}(x_{2i}, x_{1j}), \text{ pour tout i et j} = 1,\ldots,p.$$

En analyse de communité des triplets (X_2, I, D) et (X_1, I, D), la comparaison des deux tableaux est ramenée à celle de deux systèmes de variables $\{\xi_{1i}\}$ et $\{\xi_{2i}\}$, combinaisons linéaires des variables des tableaux respectifs X_2 et X_1R, associables aux décompositions de type (1) suivantes:

$$X_1 = (\Sigma\, \xi_{1i}\, {}^tu_{1i}),\ X_1R = (\Sigma\, \xi_{1i}\, {}^tu_{2i}) \text{ et } X_2 = (\Sigma\, \xi_{2i}\, {}^tu_{2i}).$$

Pour simplifier l'exposé, on suppose que l'indice i va de 1 jusqu'à p. Nous reviendrons sur cette question plus tard.

Comme $X_2 u_{2i} = \xi_{2i}$ et $X_1R\, u_{2i} = \xi_{1i}$ pour tout i, ξ_{2i} et ξ_{1i} sont définies par la même combinaison linéaire (les coefficients sont les composantes du vecteur u_{2i}) des variables de leur tableau respectif.

De plus, ${}^tX_2D\xi_{1i} = {}^t(X_1R)D\xi_{2i}$, et les variables ξ_{2i} et ξ_{1i} possèdent les mêmes liens covariants croisés avec chacune des variables observée aux instants 1 et 2 (tout comme x_{1i} et x_{2i}).

Pour ces deux raisons, ξ_{2i} et ξ_{1i} sont bien les deux expressions d'une même variable aux deux époques. C'est pourquoi elles servent de supports à notre étude de ressemblance.

En A.C.P., les variables principales sont non corrélées et permettent l'étude des liens entre variables,
en analyse de communité,

$$\forall\, i \neq j,\ \xi_{2i} \text{ est non corrélée à } \xi_{1j},$$

et les couples principaux, non corrélés quant aux liens croisés, servent à l'étude des liens entre deux tableaux.

Ces non corrélations croisées, vérifiées aussi entre couples canoniques de l'analyse canonique de deux tableaux, nous semblent plus fondamentales que les autres non corrélations que vérifient par ailleurs les variables canoniques (et que ne vérifient pas les variables principales de l'analyse de la communité). En effet, elles signifient que, pour i fixé, le seul lien que possède ξ_{2i} avec X_1 est le lien que ξ_{2i} possède avec ξ_{1i}.

De plus, les ξ_{2i} participent à la décomposition de la variance totale de X_2, et les ξ_{1i} à celle de X_1.

Finalement, la correspondance entre les variables de l'un et l'autre tableau est non seulement préservée, mais encore simplifiée à travers cette correspondance entre variables principales qui les remplacent. Cette simplification est à la base de tous les développements donnés à la méthode.

5.2 Classement des couples principaux

Invariance par changement d'échelles

Soit $T_1 = \Sigma\ (t_j \xi_{1j}\ {}^t u_{2j})$ le tableau obtenu à partir de X_1R en multipliant les variables ξ_{1j} par les coefficients $t_j > 0$. Cette opération revient à se proposer un changement d'échelle sur les données initiales X_1 dans chacune des directions orthogonales u_{1j} (celles superposées aux u_{2j} par R) en remplaçant $\text{var}(\xi_{1j})$ par $t_j^2\ \text{var}(\xi_{1j})$.
Si l'analyse procruste de X_2 et X_1 (ou de de X_2 et X_1R) mène à la définition des couples principaux (ξ_{2j}, ξ_{1j}), celle de X_2 et T_1 mène aux couples $(\xi_{2j}, t_i \xi_{1j})$.
Aux coefficients t_j près, l'analyse est donc indépendante d'une infinité de similitudes applicables au tableau X_1.
Pour la rendre plus intrinsèque, indépendante des variances des variables ξ_{1j}, il n'est donc pas souhaitable de classer les couples selon les valeurs propres $\text{cov}(\xi_{2j}, \xi_{1j})$ de $(V_{21}\ {}^t V_{21})^{1/2}$.

Classement

Notons cette fois $Y_1 = \Sigma\ (a_j \xi_{1j}\ {}^t u_{2j})$ le tableau obtenu à partir de X_1R en multipliant chaque variable ξ_{1j} par le coefficient a_j de régression linéaire simple de ξ_{2j} en ξ_{1j}. On note $\Sigma_{21} = {}^t X_2 D Y_1$.

Les couples principaux sont classés selon les variances expliquées $\rho_j^2\ \text{var}\xi_{2j}$ des régressions de ξ_{2j} en ξ_{1j}.

En effet, pour reconnaître en quoi X_1 peut ressembler à X_2, on agit par rotation et dilatations sur le nuage de points individus X_1 pour rapprocher sa forme de celle du nuage X_2. Le critère intervenant dans ces opérations est celui des moindres carrés des distances entre individus appariés.

Le tableau Y_1 est alors la forme donnée à X_1 pour désigner cette partie ressemblante à X_2.
Les valeurs propres de

$$ {}^t X_2 D Y_1 = (\Sigma_{21}\ {}^t \Sigma_{21})^{1/2}, $$

représentent des parts de communité entre X_1 et X_2. Ce sont les parts expliquées

$$ \rho_j^2\ \text{var}\xi_{2j} = \text{cov}(a_j \xi_{1j}, \xi_{2j}). $$

Et comme $\text{tr}\ ({}^t Y_1 D Y_1) = \text{tr}\ ({}^t Y_1 D X_2)$, ces valeurs décomposent l'inertie du nuage ajusté.

Si on échangeait les rôles joués par les deux tableaux, le classement pourrait être différent, se faisant selon les mesures $\rho_j^2\ \text{var}\xi_{1j}$. Mais ce sont les <u>mêmes</u> coefficients de corrélations linéaires ρ_j qui interviendraient. C'est dire que le changement de discours alors induit, provient du changement des variances, pouvant influer sur le classement, et non des corrélations entre variables principales.

La notion même de ressemblance exprimée dans une mesure de communité est contenue dans la valeur ρ_j^2. La quantité $\text{var}\xi_{2j}$ exprime l'importance de la variabilité de X_2 concernée par cette ressemblance.

Une valeur ρ_j^2 élevée caractérise une direction commune u_{2j} sur laquelle les deux nuages appariés projetés sont proches, hors considération d'échelle et de moyenne.

Le rapport des variances totales de Y_1 et X_2 est un indice de communité globale entre les deux tableaux initiaux C'est le carré de l'indice d'association de Lingoes et Shönemann (1974) entre Y_1 et X_2 Ici il vaut 0,4247, valeur lue en bas de la troisième colonne de la table 1.

com	cumuls	cumul-var	rot.angles	corr. **2
1.5051	49.6992	21.1059	0.9978	0.7325
0.4426	64.3132	27.3121	0.9955	0.3172
0.4419	78.9061	33.5093	0.9930	0.5578
0.2245	86.3176	36.6568	0.9850	0.2597
0.2104	93.2660	39.6076	0.9828	0.3051
0.1391	97.8600	41.5585	0.9925	0.1913
0.0648	100.0000	42.4674	0.9816	0.1067

Table 1 *Mesures de communité. Cumuls Cumuls des parts de variance expliquée de X_2. Cosinus des angles procrustes. Corrélations au carré des variables des couples principaux (intensités de la ressemblance)*

5.3 Analyse de la variance en communité

Dans le schéma ci-dessous, pour j fixé quelconque, l'époque 2 est une variables ξ_{2j}, et l'époque 1 la variable correspondante ξ_{1j}. Les longueurs des vecteurs représentent des écarts-type, et les cosinus des angles désignent les corrélations entre variables. Les interprétations relatives à la longueur de ξ_{2j}, à la corrélation ρ_j et à la longueur de la variable projetée sur ξ_{1j} ont été faites en 5.2 Il reste deux longueurs à interpréter dans ce schéma, correspondant à ce qui est désigné comme autonome et comme non expliqué.

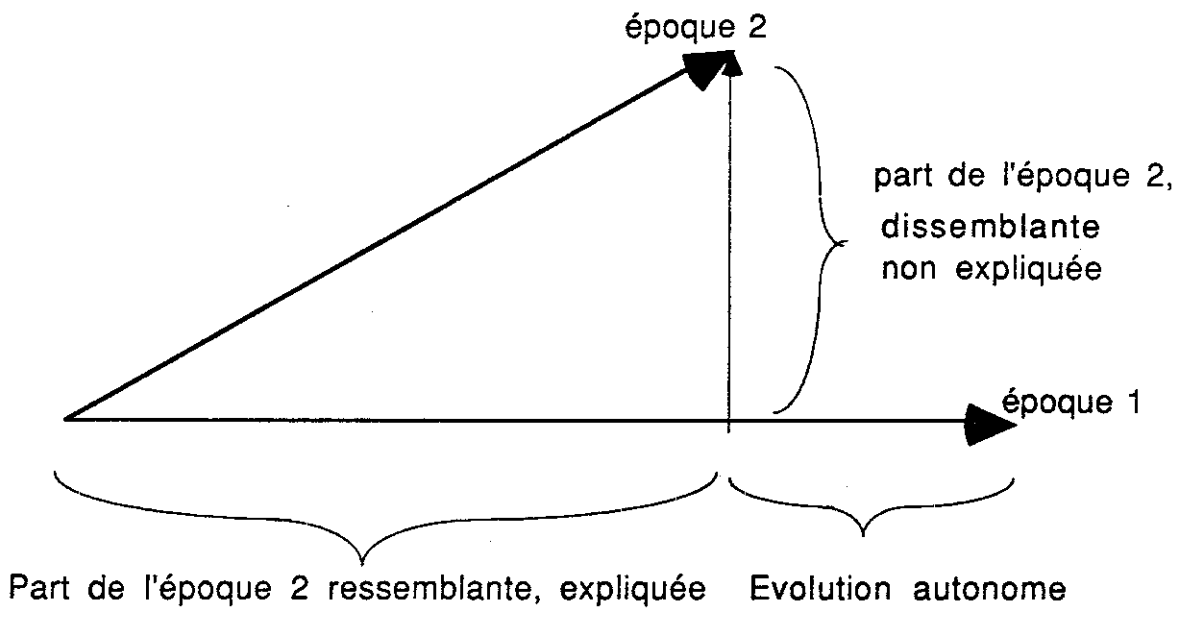


Schéma explicatif de l'analyse de la variance en analyse de communité

Evolutions en variance autonomes

On a déjà souligné que l'analyse de communité de X_2 et Y_1 était identique à celle de X_2 et X_1. Elle est insensible au fait que d'autres variances que celles obtenues auraient pu être observées pour les variables ξ_{1j}. La variance expliquée ρ_j^2 varξ_{2j}, qui aurait pu être la variance de ξ_{1j}, participe totalement à l'explication de l'évolution temporelle en variance. A contrario, la différence entre la variance de ξ_{1j} et la variance ρ_j^2 varξ_{2j} est sans effet sur l'explication. Nous dirons qu'elle correspond à une évolution en variance propre à chacune des variables, autonome.

Comme il n'est pas très intéressant de faire cette évaluation sur les variables principales, on revient aux variables du tableau X_1.

Dans la table 2, sont données les mesures de ces évolutions autonomes, par le calcul des variances des variables du tableau $X_1 - (\Sigma\, a_j \xi_{1j}\, {}^t u_{1j})$. Les rapports entre variance calculées et variances observées en 1979, exprimés en pourcentage, et traduisant l'évolution en variance autonome la plus faible, sont ainsi observées pour les analyses médicales Acide Urique et Poids.

Analyses médicales	Ca	Ch	TA	AU	Pr	So	Pd
Variance en 1979	0,89	1,27	0,94	0,94	0,99	0,82	1,00
Variances autonomes	0,18	0,19	0,13	0,07	0,20	0,20	0,07
Pourcentages	20%	15%	14%	8%	20%	25%	7%

Table 2 *Les pourcentages évaluent l'importance de l'évolution propre pour chaque variable.*

Parts non expliquées

Le tableau $(X_2 - Y_1) = \Sigma\, (\xi_{2j} - a_j\, \xi_{1j})\, {}^t u_{2j}$ représente ce qui dans X_2 est non expliqué par X_1, indépendamment des changements d'échelles autonomes précédents. En effet, une différence $(\xi_{2j} - a_j\, \xi_{1j})$ représente le vecteur des résidus de la régression de ξ_{2j} en ξ_{1j}. Ainsi on vérifie que

$$\mathrm{tr}[{}^t(X_2-Y_1)D(X_2-Y_1)] = \mathrm{tr}({}^tX_2DX_2) - \mathrm{tr}({}^tY_1DY_1)$$

La variance totale de X_2 (égale dans l'exemple à 7,13: dans l'ensemble, la variabilité augmente ici légérement, étant supérieure à 7), est décomposable en une part de variance associée à la ressemblance (42,47%, dernière valeur de la troisième colonne de la table 1) et une autre part attribuable à la dissemblance (57,53%, valeur en même position, mais dans la table 3). De la sorte, les notions de ressemblance et de dissemblance sont bien complémentaires.
Les valeurs propres de l'opérateur ${}^t(X_2 - Y_1)\, D\, (X_2 - Y_1)$, fournies dans la première colonne de la table 3, permettent de classer axes et variables principaux servant aux représentations de la dissemblance entre les tableaux (A.C.P. de $\{(X_2-Y_1), I, D\}$).

Cette dissemblance des deux tableaux provient des changements qui n'ont pu s'expliquer en phase avec le changement du jeu des corrélations internes aux tableaux. Cela est le produit d'un changement des distributions des variables, accompagné d'un changement dans le classement des patients selon les nouvelles valeurs prises par ces variables à l'époque 2. Elle peut provenir aussi éventuellement d'une part de variabilité totalement originale, propre à X_2 seul, lorsque le rang de V_{21} est inférieur strictement à celui de X_2 (ce qui ne se produit pas sur les données traitées ici, le rang de V_{21} étant égal à 7).

diff	cumuls-diff	cumuls-var
1.4116	34.4051	19.7942
0.9167	56.7474	32.6483
0.6355	72.2372	41.5600
0.3840	81.5966	46.9447
0.3377	89.8271	51.6799
0.2874	96.8315	55.7097
0.1300	100.0000	57.5326

Table 3 *Mesures de dissemblance Cumuls Cumuls des parts non expliquées de var(X_2).*

Dans **RESDIF-FORTRAN**, le rang r de V_{21} correspond au nombre de valeurs non nulles de la colonne com.

6. Choix du nombre total d'axes de représentation

Dans la première colonne de la table 1 se trouvent les mesures décomposant la part ressemblante, et dans celle de la table 3 les mesures décomposant la part dissemblante. Vue la complémentarité des deux notions, provenant de ce que le total de toutes ces mesures égale la variance de X_2, on peut choisir le nombre d'axes en observant simultanément les deux suites de valeurs.
Le nombre total d'axes adoptés est associé à une part de la variance totale de X_2, celle alors considérée dans l'ensemble des graphiques.

Par exemple, en prenant 3 axes pour la ressemblance, on englobe 78,9% de toute la variance expliquée, soit 33,5% de la variance de X_2 (ligne 3 des colonnes 2 et 3 de la table 1).
En prenant 3 axes pour la différence, on englobe 72,2% de la part de variance non explicable, soit 41,6% de la variance de X_2 (ligne 3 des colonnes 2 et 3 de la table 3).
Avec ces 6 axes est englobée 75% de la variance de X_2.

Lors d'un premier passage de l'analyse sur les données, on n'hésitera pas à surévaluer le nombre d'axes de dissemblance, pour des raisons indiquées en Section 8.2. Lors du dernier passage, sur les tableaux formés des individus non éliminés lors des premières tentatives, on aura l'attitude inverse puisqu'il s'agit de présenter l'essentiel en résumant le plus possible.

De plus, nous pouvons faire intervenir dans notre choix la lecture des valeurs de la quatrième colonne de la table 1. Ce sont les cosinus des angles des rotations nécessaires pour mettre en correspondance la structure intracovariante de X_1 avec celle de X_2 (en superposant par R le système orthonormé $\{u_{1i}\}$ au système orthonormé $\{u_{2i}\}$).

Supposons qu'au lieu des valeurs observées toutes proches de 1, on ait observé à partir d'un certain moment, dans la liste des valeurs, une succession de valeurs positives et négatives. Cela révèlerait un désordre dans la mise en correspondance des deux structures covariantes, relativement aux directions concernées.
Il aurait été raisonnable d'introduire l'hypothèse du "bruit" et, pour la poursuite de l'étude de la ressemblance, on aurait pu ne retenir que les premières dimensions traduisant une correspondance structurelle évidente.
Venant confirmer cette remarque, nous avons pu vérifier par ailleurs sur divers jeux de données, qu'en opérant une permutation circulaire sur les individus d'un seul tableau, détruisant ainsi toute notion de correspondance entre les individus des deux tableaux, le

calcul de ces cosinus donne alors des valeurs successives désordonnées entre -1 et +1 (dès le début et de façon évidente pour des jeux de dimensionnalité p faible), indiquant l'impossibilité de reconnaître par rotation l'existence d'une structure commune aux deux tableaux.

En pratique, au minimum on devrait rejeter toutes les dimensions dès la première valeur négative observée dans la liste parcourue de haut en bas.

Ici, tous les angles sont voisins de 0 et aucun désordre n'apparaît nettement relativement à un sous-espace. De ce point de vue, les 7 dimensions sont donc conservées pour l'étude de la ressemblance.

Dans **RESDIF-FORTRAN**, batch ou conversationnel, on ne demandera pas de représentations graphiques au premier passage, puisque tous les résultats sont envoyés dans un fichier, sans apparaître à l'écran. Le premier passage sert à lire les tableaux.
Les r dimensions, correspondant au rang r de V_{21}, sont toutes conservées pour définir la ressemblance. Il n'y a donc pas suppression de dimensions au vu des cosinus précédents, qui auraient pu être mises ensuite au compte de la différence.
Le plus souvent, ces dimensions sont relatives à une variabilité négligeable.

7. Graphiques

7.1 Lectures associées à la ressemblance

On a

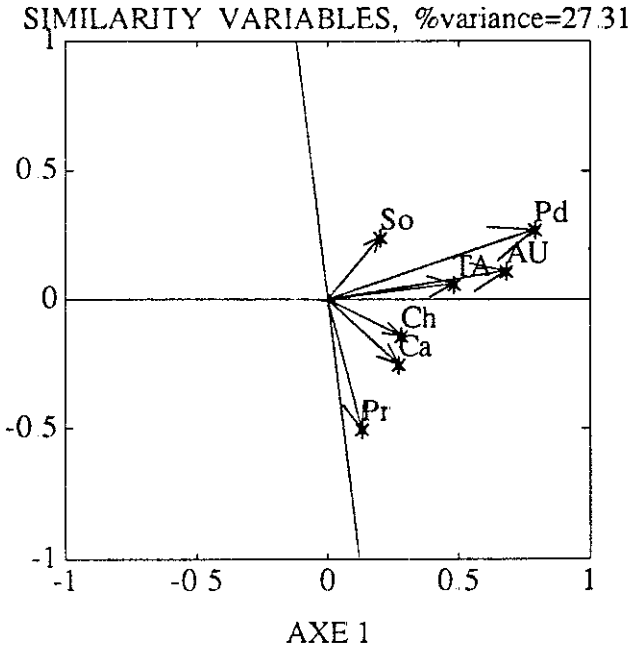
$$ {}^tX_2D\ [\ a_j\xi_{1j} / \{var(a_j\xi_{1j})\}^{1/2}] = com_j^{1/2}\ u_{2j} $$

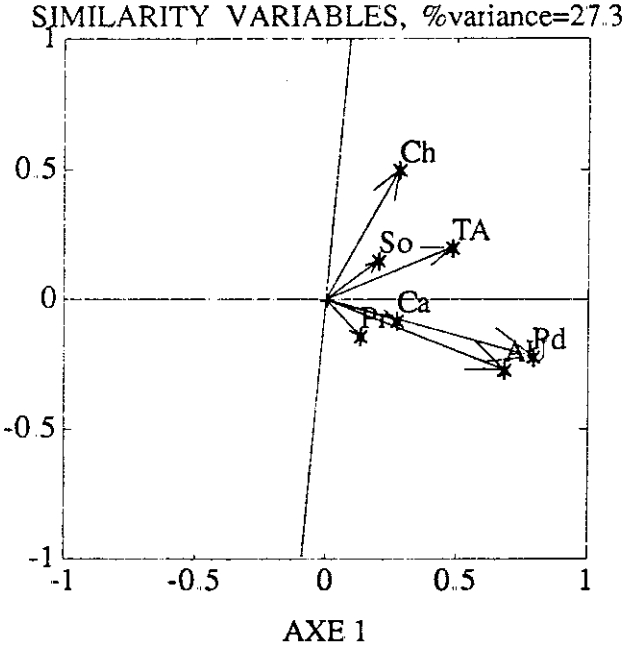
Les coordonnées $com_j^{1/2}\ u_{2j}$ servent aux représentations des points vectoriels des variables de X_2. On représente donc dans un plan explicatif ce qui est à expliquer. La même unité est choisie sur chaque axe. La longueur d'un vecteur représentatif d'une variable traduit l'importance de sa contribution à l'explication (par elle-même, à partir du passé; c'est donc l'importance de sa contribution à la stabilité temporelle).

Les repères (ξ_{1i}, ξ_{1j}), $i \neq j$, ne sont pas vraiment orthogonaux, même si, par construction, l'angle est suffisamment proche de 90° pour qu'il soit admissible de définir les plans de représentation à partir de ces vecteurs principaux. La solution adoptée ici à ce propos consiste à effectuer la représentation par projection D-orthogonale des variables sur les plans engendrés par ces repères.

L'angle entre deux axes d'un repère plan peut être interprété comme un indicateur de fiabilité du plan de représentation, d'autant meilleur qu'il est voisin de 90°.

Dans **RESDIF-FORTRAN**, la valeur du cosinus de l'angle des deux axes de représentation, souhaitée proche de zéro, est indiquée dans la ligne surmontant le graphique (ce nombre doit être nul en A.C.P.V.I.). Le calcul des coordonnées est réalisé en introduisant pour chaque plan une métrique telle que l'angle entre les deux axes soit l'angle projeté réel et les distances lues sur ces deux axes aussi. Ce n'est donc pas la métrique identité, mais une métrique qui lui est d'autant plus proche que le cosinus est proche de zéro.
La variabilité englobée dans une représentation est accessible à partir de la colonne 3.





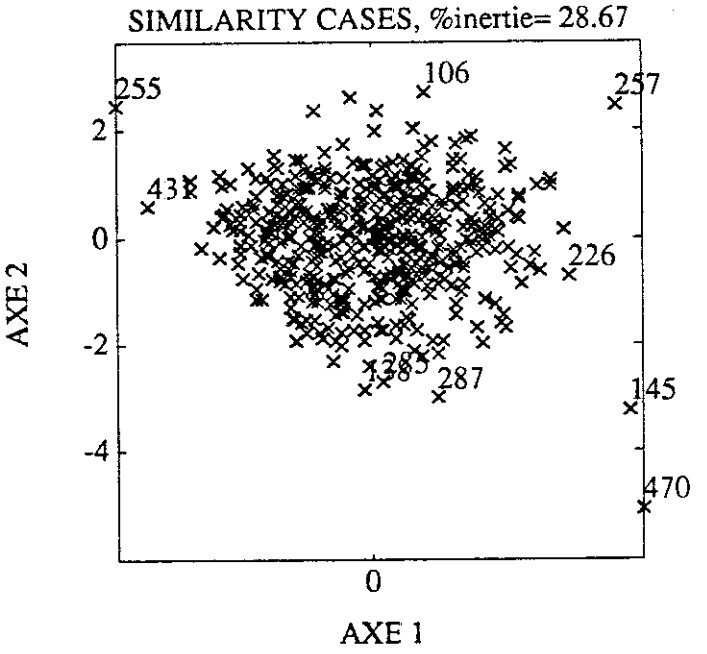
Analyses médicales	Ca	Ch	TA	AU	Pr	So	Pd
Variance en 1984	1,11	0,73	1,06	1,05	1,00	1,18	1,00

D'après les représentations précédentes, les analyses médicales évoluant dans le temps de façon la plus stable tout en étant porteuse d'une variabilité importante en 1984, sont d'abord les analyses: Acide Urique et poids
L'analyse médicale Cholestérol est porteuse d'une variabilité sensiblement plus faible que les autres analyses en 1984. Pour la rendre comparable aux autres quant à sa participation

à la stabilité proprement dite (à la ressemblance proprement dite), on peut ainsi mentalement allonger la longueur du vecteur qui la représente, de plus de 25%. Nous la jugeons finalement comme faisant partie des plus stables aussi.
Cette lecture faite pour cette variable cholestérol qui se confond assez avec l'axe 3, est confirmée par l'indice de stabilité plus important pour l'axe 3 que pour l'axe 2 (lignes 2 et 3 de la dernière colonne de la table 1).

Comme de plus les distributions sont sensiblement gaussiennes, nous pouvons dire que les trois analyses médicales précédentes sont celles qui se sont le moins modifiées quant à la nature gaussienne de leurs distributions, et/ou quant à la façon dont elles ordonnent les patients.

Les représentations des individus de X_2 sont bâties à l'aide des repères issus du système $\{u_{2j}\}$. Comme $X_2 u_{2j} = \xi_{2j}$, les coordonnées des patients de 1984 sur l'axe j sont les composantes des variables principales ξ_{2j}. Le centre de gravité du nuage corespond à un comportement moyen, comme en A C P.



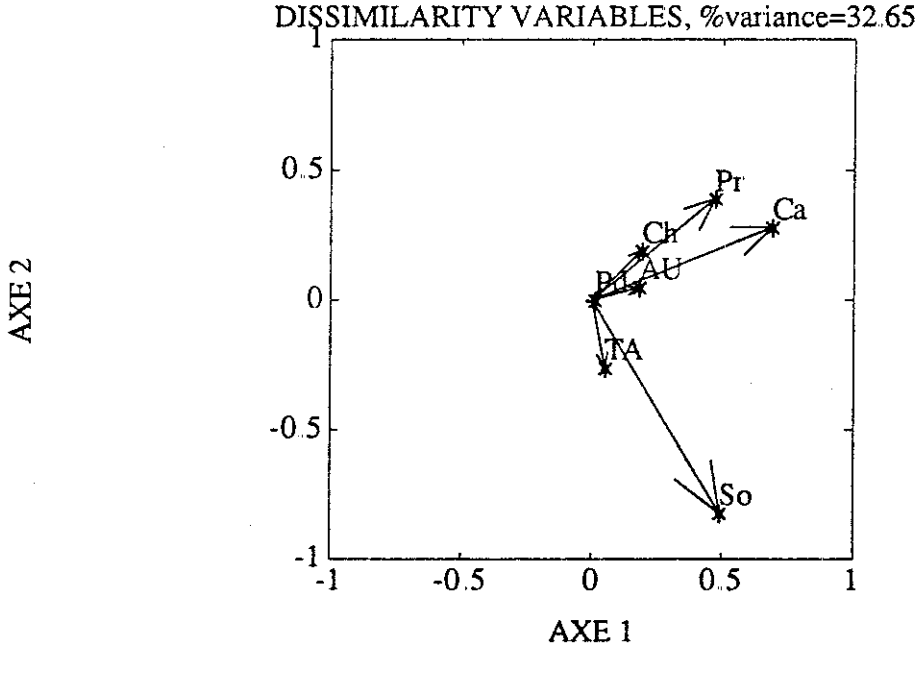
Sur le graphique ci-dessus, les patients 145, 255, 257, et surtout 470 sont assez démarqués pour qu'on puisse considérer comme excessive leur influence dans la détermination des axes de représentation, et se poser la question d'une analyse sans leur présence. Le patient 470 est dans une direction (droite allant de l'origine au point) qui est à peu près orthogonale à celle de la variable Sodium repérée dans le plan (1,2) correspondant des variables. A priori, c'est putôt pour les autres variables, contribuant à cette direction, que ce patient possède une forte valeur, en 1984 (et en 1979, puisque ce premier plan caractérise la stabilité).

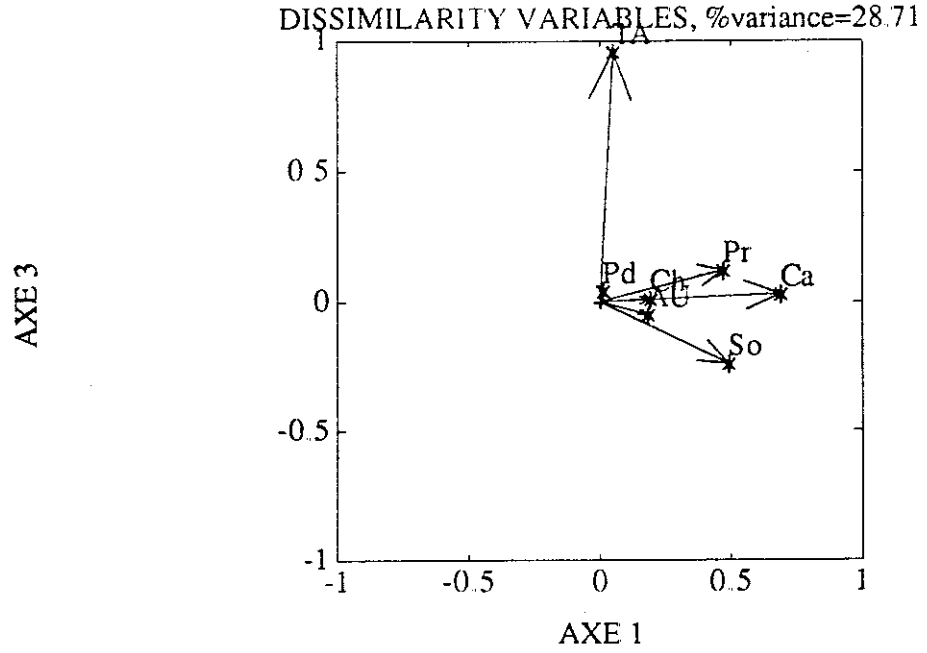
Il est vrai que l'on vient de faire coïncider mentalement un repère orthonormé avec un repère qui ne l'est pas tout à fait, mais ce défaut de lisibilité a peu de conséquences, la double lecture étant valide en moyenne, sujette à caution en particulier: elle doit être confirmée par un retour aux données, pour observer les valeurs des variables pour les patients en question.

Dans **RESDIF-FORTRAN**, les 5 points les plus éloignés de l'origine dans un plan de représentation d'individus, sont nommés dans un tableau qui précède le graphique, avec les valeurs de leurs coordonnées. Un paramètre, à fixer par l'utilisateur, permet de se dispenser ou non des représentations des individus. Pour une meilleure lecture des graphes, il manque un indice de qualité de la représentation dans un plan donné, pour chaque variable, ou individu, basé sur la distance du point au plan

7.2 Lectures associées à la différence

D'après les premiers plans qui suivent, les analyses médicales les plus responsables de l'instabilité temporelle sont Calcium et Protides et, peu liée à celles-ci, Sodium. Dans le plan (1,3), on repère aussi la Tension Artérielle, très bien représentée et presque constitutive de l'axe 3 à elle seule.
Ces quatre analyses médicales sont donc celles qui caractérisent le plus une modification quant à leurs distributions (modification de distributions à considérer indépendamment de l'évolution des moyennes ou des variances), et/ou quant à la façon dont elles ordonnent les patients, tout en étant associées à une forte variabilité en 1984.





Dans les représentations des individus, le tableau Y_1 est représenté par le centre de gravité du nuages des patients, et la distance lue entre l'origine et un patient correspond à la vision de l'écart projeté entre un individu de X_2 et l'individu correspondant de Y_1.
En effet, on aurait pu projeter sur les plans utilisés les points de X_2 et ceux de Y_1. Les distances entre points appariés auraient été les distances précédentes, le trajet entre les couples de points ayant même direction et même sens que les trajets allant de l'origine aux patients dans les représentations de l'A.C.P.. Mais ce graphe serait moins lisible.

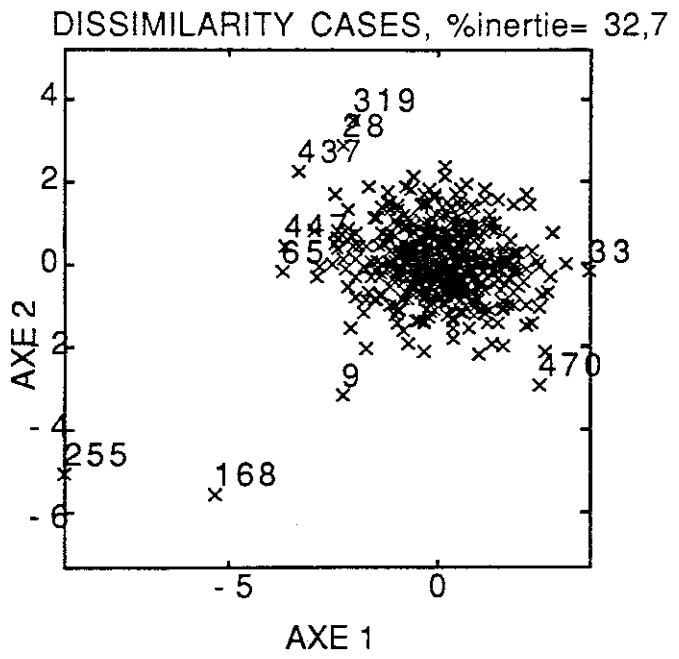
La mise en correspondance d'une direction attachée à un patient éloigné de l'origine, avec les directions et sens des variables dans les graphes duaux des variables, permet de désigner les analyses médicales ayant a priori le plus contribué à la présence de ce patient cause de dissemblance entre les deux tableaux

A priori et par construction, de tels patients devraient avoir pour certaines variables des valeurs assez grandes en 1984 et assez petites en 1979, à moins que ce soit assez petites en 1984 et assez grandes en 1979, toute proportion gardée.
Si un patient est éloigné de l'origine de façon exceptionnelle, comme le sont ci-dessous 255 et 168, son influence sur l'étude ne peut être négligée, tant du point de vue de la ressemblance que de la différence, et l'usager de la méthode doit recommencer l'analyse sans sa présence.

Pour cela, on ne peut que conseiller de considérer un nombre un peu plus important que nécessaire de premiers plans de représentations, lors des premiers passages de l'analyse, visant à mettre en évidence ces patients particuliers. C'est notre façon d'aborder la robustesse. On aurait pu faire autrement, en adoptant en aval des estimateurs plus robustes pour V_{21}.

Ici, trois reprises furent nécessaires pour décider que de tels patients isolés n'existaient plus, sachant que nous n'avions pas éliminé a priori de l'étude les patients ayant une valeur manifestement exceptionnelle relativement à une ou plusieurs analyses médicales.

L'observation des valeurs des analyses médicales pour ces patients permet de savoir pourquoi ils sont détectés ici, alors qu'ils ne l'auraient pas été en ressemblance.



Remarque

Pour des raisons d'économie de rédaction, les résultats fournis précédemment l'ont été après avoir retiré de l'étude 13 patients ayant un comportement trop exceptionnel.

Mais les nuages des patients sont ceux du premier passage de l'analyse sur l'ensemble des données. *Cela explique pourquoi les valeurs des inerties indiquées sur les graphiques ne sont pas exactement celles des variances indiquées sur les graphiques correspondant pour les variables, et accessibles dans les tableaux fournis.*

La correspondance entre les graphiques des variables et des individus a cependant été faite, parce que les modifications produites par le retrait des 13 patients se sont avérées peu sensibles sur les graphiques des variables.

Pour la suite de l'étude, les tableaux sont ceux formés des 487 patients restant, ces tableaux étant centrés et réduits comme indiqué en Section 3. La comparaison proposée par la suite entre les graphiques des variables est donc valide à ce titre.

8. Une étude du potentiel prévisible

8.1 Position du problème et Normalisation

Nous voulons étudier dans quelle mesure il est possible de prédire X_2 à partir de X_1. Quelles sont les analyses médicales les plus prévisibles? les moins prévisibles? Nous répondons à cette question tout en condensant l'information, c'est à dire en recherchant une variabilité expliquée de X_2 importante sur un nombre de dimensions réduit.

La méthode employée revient à reprendre l'étude précédente en remplaçant l'expression "quelle part de X_2 ressemble à X_1?" par l'expression: "quelle part de X_2 est prévisible par X_1?", sans modifier le tableau X_2.

Au tableau X_1 formé des variables centrées, on substitue le tableau normalisé $X_1V_1^{-1/2}$, où $V_1 = {}^tX_1DX_1$. Cela équivaut à introduire la métrique de Mahalanobis dans l'espace

des patients et donc ainsi à ne retenir comme information sur les variables du tableau X_1 que celle d'espace engendré.

Alors, l'analyse de communité des triplets $(X_1V_1^{-1/2}, I, D)$ et (X_2, I, D), avec X_2 à expliquer, devient une analyse où une variable de X_2 peut s'expliquer par n'importe quelle variable de X_1.

Précédemment, l'explication globale a été recherchée sous une contrainte d'appariement des variables de X_1R avec celles de X_2: les variables principales correspondaient à une même combinaison linéaire de leurs tableaux respectifs. Maintenant, le tableau X_1 est remplacé par un autre et cette contrainte disparaît. Ce faisant, les parts expliquées deviennent plus importantes *dans leur ensemble*.

Les variables appelées précédemment ξ_{1i} et recalculées sur les nouveaux triplets statistiques, sont les variables explicatives définies par Rao en analyse en A.C.P.V.I. et celles appelées ξ_{2i} sont les expliquées correspondantes définies par Johansson. Le carré de l'indice de Lingoes et Shönemann précédemment cité en communité est ici l'indice de redondance de Stewart et Love. L'A.C.P. considérée en communité pour analyser la différence correspond ici à l'analyse des covariances partielles de Lebart, Morineau et Fénelon. C'est l'A.C.P. du triplet $(X_2 - X_1V_1^{-1}V_{21}, I, D)$. La matrice des covariances à diagonaliser est celle des covariances partielles $V_2 - V_{21}V_2^{-1}V_{21}$.

Comme le même programme a été utilisé, les résultats se présentent de la même façon. Cependant les lectures qui en sont faites doivent être différentes puisqu'il s'agit d'une autre analyse.

8.2 Résultats

com	cumuls	cumuls-var	rot.angle	corr.**2
1.5614	49.6020	21.8953	0.9966	0.7838
0.4809	64.8777	28.6384	0.9939	0.3513
0.4458	79.0382	34.8891	0.9917	0.5329
0.2516	87.0304	38.4170	0.9929	0.3789
0.2094	93.6813	41.3528	0.9763	0.2989
0.1362	98.0091	43.2633	0.9921	0.1660
0.0627	100.0000	44.1421	0.9799	0.0837

Table 4 *Variances expliquées - Cumuls en poucentage - Cumuls des parts expliquées de X_2 - cosinus (non interprétés ici) - Corrélations au carré entre variables expliquées-explicatives des couples principaux.*

diff	cumuls-diff	cumuls-var
1.3887	34.8615	19.4729
0.9027	57.5221	32.1306
0.6302	73.3435	40.9682
0.3709	82.6546	46.1692
0.3296	90.9287	50.7909
0.2785	97.9203	54.6963
0.0828	100.0000	55.8579

Table 5 *Parts de variance non expliquée. Cumuls en pourcentage. Parts de variance de X_2 non expliquée cumulées.*

La part de variance expliquée vaut 44,14% (colonne 3, dernière ligne de la table 4) et donc est très peu améliorée par rapport à l'approche explicative précédente, où elle valait 42,27% de la variance totale de X_2.

Les variables détectées comme les plus stables en communité, pourraient être les mieux prévues en A.C.P.V.I., les plus instables l'étant le moins bien.

Et en effet, les changements observés sur les graphiques de représentation des variables, par rapport à ceux obtenus en communité, sont absolument indécelables à l'œil, que ce soit en "ressemblance" ou en "différence". Ces graphiques ne sont donc pas fournis ici.

Cependant, l'analyse de communité est conceptuellement différente de l'analyse explicative.

Notons que les variables principales explicatives, axes des repères d'observation des variables les plus prévisibles, qui correspondaient aux variables ξ_{1i} considérées en analyse de ressemblance, sont cette fois non corrélées par construction.

Pour certains jeux de données, le calcul de l'inverse de V_1 pose problème, rendant délicate la mise en œuvre de l'A.C.P.V.I. Au vu de cette similitude des graphiques, qui est souvent rencontrée, une approche de substitution par l'analyse de communité sera souvent raisonnable.

8.3 Comparaison des deux modèles

Dans l'étude de ressemblance, nous avons le modèle sous-jacent (M_1) d'ajustement à X_2 :
$X_2 = Y_1 + E$, où E représente le tableau des résidus, et $Y_1 = X_1 (\Sigma a_j u_{1j} {}^t u_{2j})$.
Dans l'étude prévisionnelle, construite sur le même mode, mais en ayant substitué à X_1 le tableau $X_1 V_1^{-1/2}$, on a le modèle d'ajustement (M_2):
$X_2 = Z_1 + F$, où $Z_1 = X_1 (V_1^{-1} {}^t V_{21})$ et F représente les résidus.
La variance totale de F étant toujours inférieure à celle de E, le modèle d'ajustement M_2 est toujours meilleur globalement que M_1.

Mais cela ne signifie pas que chacune des variables soit mieux ajustée avec M_2.

C'est pourquoi dans la table 6 nous comparons variable par variable les variances expliquées des tableaux ajustés Y_1 et Z_1. Les rapports de variance supérieures à 1 nous indiquent quelles sont celles s'ajustant mieux avec M_1.

Analyses médicales	Ca	Ch	TA	AU	Pr	So	Pd
Variance en 1984	1,11	0,73	1,06	1,05	1,00	1,18	1,00
Parts expliquées par M1	0,30	0,50	0,44	0,57	0,33	0,23	0,66
Parts expliquées par M2	0,24	0,38	0,41	0,67	0,36	0,19	0,91
Rapports M1 / M2	1,26	1,30	1,07	0,85	0,93	1,23	0,73

Table 6 *Comparaison des variances expliquées obtenues avec M_1 et M_2.*

Pour les analyses médicales Acide Urique et Poids, une nette amélioration apparaît en utilisant le modèle prévisionnel M_2, la prévision de ces variables étant meilleure sans la prise en compte de l'évolution dans le temps des liens entre analyses médicales, en

permettant que n'importe quelle analyse médicale du passé puisse expliquer chacune d'elles au présent.
En se référant aux coefficients des combinaisons linéaires des variables du passé, on pourrait vérifier que les variables jouant le plus grand rôle dans l'explication ou de l'Acide Urique, ou du Poids, sont ces deux variables, et aussi dans une moindre mesure, la Tension Artérielle

Pour les analyses médicales Calcium, Cholestérol et Sodium, c'est le modèle M_1 qui permet la meilleure explication, celle qui intègre l'information sur l'évolution des liens que possèdent ces analyses avec les autres analyses médicales. Elles s'expliquent alors mieux par elles-mêmes qu'en recherchant l'explication ailleurs
Par exemple, l'explication de l'état présent du Calcium se fait à partir du Calcium au passé. Les autres analyses n'interviennent dans cette explication qu'à travers les liens du calcium-présent avec les autres variables du présent, et ceux du calcium-passé avec les autres variables du passé. Pour que cette explication soit meilleure qu'en A.C.P.V.I., il a fallu:
a) un Calcium-présent plus corrélé au Calcium-passé qu'il ne pourrait l'être à toute combinaison linéaire des autres variables du passé.
b) des liens du Calcium aux autres variables, aux deux époques respectives, améliorant effectivement l'explication.

Dans **RESDIF-FORTRAN**, la comparaison entre les parts expliquées est fournie si on choisit l'option 3 correspondant à la demande d'exécution des deux analyses, communité et A.C.P.V.I.

8.4 Analyse canonique

L'utilisateur peut user du programme pour effectuer l'analyse canonique des deux tableaux. Il suffit de choisir l'option A.C.P.V.I. avec le tableau à expliquer égal à $X_2V_2^{-1/2}$. Il lui faudra auparavant construire ce tableau à partir de X_2 et prendre l'option "pas de réduction".

En "ressemblance", on obtient les termes de l'analyse canonique classique. Habituellement construits et présentés avec les métriques de Mahalanobis, les représentations des individus se font ici avec la métrique identité.
Les axes "u_{2i}" de l'analyse sont les vecteurs propres, orthonormés au sens usuel, de $V_2^{-1/2}\, V_{21} V_1^{-1}\, {}^tV_{21} V_2^{-1/2}$.
Les valeurs propres "com" sont les corrélations canoniques au carrée.
Les couples "(ξ_{i2},ξ_{i1})" sont les couples canoniques.

Les représentations des individus coïncident dans les deux approches.

Les représentations des "variables" sont les représentations des variables du tableau $X_2V_2^{-1/2}$. Toutes ces variables sont non corrélées, et il ne s'agit pas de s'intéresser à la lecture des angles, différents de 90° uniquement parce qu'on lit dans un plan ce qui se situe dans R^p. En analyse canonique, c'est plutôt la communauté qu'une variable peut présenter avec l'autre espace qui peut intéresser, et qui est lisible ici. Ainsi, une variable qui serait totalement représentée dans un plan, a une longueur de vecteur mesurant la communauté que cette variable possède avec l'ensemble des variables de l'autre tableau. Si elle peut être contenue dans l'espace engendré par le tableau X_1, sa longueur est 1. Si elle appartient à un espace orthogonal, sa longueur est 0.

L'analyse de la "différence" correspond à l'A.C.P. de $(X_2V_2^{-1/2} - X_1V_1^{-1}{}'V_{21}V_2^{-1/2})$.
La matrice de corrélations à diagonaliser est

$$(I - V_2^{-1/2}V_{21}V_1^{-1}{}'V_{21}V_2^{-1/2})$$

Cette A.C.P. correspond à la volonté de dégager et résumer tout ce qui n'est pas intervenu en communauté.

9. Une étude de l'évolution des moyennes

On intégre à l'analyse de communité considérée au début de l'exposé une étude de l'évolution dans le temps des moyennes.
L'objectif est de faire la comparaison entre moyennes des variables se correspondant dans l'un et l'autre tableau. Cependant, on veut tenir compte des liens que possédent ces variables aux autres variables
Comme la correspondance est préservée en substituant aux variables initiales les variables principales de l'analyse de communité, on reporte la comparaison sur ces variables principales. Les liens sont alors simplifiés, puisque ξ_{2j} ne correspond plus dans X_1 qu'à ξ_{1j}. De plus, bien que les variables principales d'un même tableau soient corrélées (faiblement le plus souvent), elles sont relatives à une décomposition de la variance totale. On est ainsi dans de bonnes conditions pour substituer à la comparaison multivariée, qui exige des hypothèses lourdes à vérifier même dans un contexte gaussien, plusieurs comparaisons univariées.

On applique aux tableaux la réduction des variables précédemment choisie. Mais on ne centre ni les variables de X_1, ni celles de X_2: on rajoutte aux colonnes des deux tableaux les valeurs des moyennes (réduites à l'aide des mêmes coefficients). Soient $\underline{X}_1$ et $\underline{X}_2$ les nouveaux tableaux. On pose $\xi_{2j} = \underline{X}_2\, u_{2j}$ et $\xi_{1j} = \underline{X}_1\, u_{1j}$.

La comparaison des moyennes est alors reportée sur les variables principales qui se correspondent: pour chaque j, en comparant les valeurs de ξ_{2j} à celles appariées de ξ_{1j}.

Dans notre exemple, on a vérifié que pour toutes les variables principales l'hypothèse de distribution gaussienne était admissible. On a effectué un test de l'égalité des moyennes de deux variables principales contre la différence, en effectuant le test de Student sur la moyennes des valeurs de la variable différence: $(\xi_{2j} - \xi_{1j})$.

Soit J le sous-ensemble des indices pour lesquels on a retenu une différence significative, avec un coefficient de sécurité égal à 5%. Soient $\underline{Z}_2$ et $\underline{Z}_1$ les tableaux reconstitués uniquement à partir des variables ξ_{2j} et ξ_{1j}, $j \in J$:

$$\underline{Z}_2 = \Sigma_{j\in J}\ \xi_{2j}\, u_{2j}' \text{ et } \underline{Z}_1 = \Sigma_{j\in J}\ \xi_{1j}\, u_{1j}'$$

En calculant alors les moyennes des variables de ces nouveaux tableaux, on trouve des valeurs que l'on peut considérer comme constituant des écarts réels à zéro, même s'ils sont assez faibles.

Ici, aucune variable principale n'a été rejetée, de sorte que ces moyennes sont celles des variables de $\underline{X}_2$ et $\underline{X}_1$, relatives à des écarts (réduits) à zéro significatifs et fournis ci-dessous:

Analyses médicales	Ca	Ch	TA	AU	Pr	So	Pd
Différences entre moyennes	0,41	-0,73	-0,17	-0,12	-0,38	-0,70	-0,16
Différences recalculées	0,41	-0,73	-0,17	-0,12	-0,38	-0,70	-0,16

Seule l'analyse médicale Calcium a augmenté en moyenne au cours de la période de 5 ans.

Dans **RESDIF-FORTRAN**, cette comparaison peut être obtenue en option; est fournie la valeur limite calculée en remplaçant la distribution de Student par la distribution gaussienne, comme cela est acceptable quand n est grand. Donc le programme fait un calcul faux si n est trop petit. Le programme ne teste pas la normalité des distributions des variables principales. Ensuite est fourni un tableau contenant les différences entre moyennes, pour chaque variable entre les deux époques (initiales), celles entre moyennes une fois les tests effectués (recalculées), et cela compte tenu du mode de réduction adopté.

REFERENCES

Horber E. Le logiciel EDA. *Dpt de Science Politique, Université de Genève*

Johansson J. K. (1981) An extention of Wollenberg's redundancy analysis. *Psychometrika* 46, 93-105.

Lafosse, R. (1989) Ressemblance et différence entre entre deux tableaux totalement appariés. *Statistique et Analyse des données*, vol 14 n°2, 1-24.

Lebart L., Morineau A., Fénelon J.P. (1979) Traitement des données statistiques: méthodes et programmes. *Dunod*.

Lingoes J.C., Shönemann P.H. (1974) Alternative measures of fit for the Shönemann-Carroll matric fitting algorithm. *Psychometrika*, 39, 423-429.

Rao C.R. (1964) The use and the interpretation of component analysis in applied research", *Sankya, ser.A*, 26, 329-358.

Stewart D. et Love W. (1968) A general canonical correlation index. *Psychological Bull.*, 70, 160-163.