Retour au menu

ANALYSE FACTORIELLE DES CORRESPONDANCES

par M. ROUX
Université de Paris - France

1ère partie

QUELQUES FORMULES MATHEMATIQUES COMMENTEES EN GUISE DE PREAMBULE.

Avant d'exposer nos formules précisons nos notations. Nous appelerons $k(i,j)$ les nombres portés à l'intersection de la ligne i et de la colonne j du tableau à analyser; par un abus d'écriture dont les mathématiques sont coutumiers, nous nous servirons de la même lettre k pour désigner les sommes des lignes et les sommes des colonnes de ce tableau.

$$k(i) = \sum_j k(i,j) \qquad k(j) = \sum'_i k(i,j) .$$

La lettre k employée sans parenthèses désignera la somme de tous les termes du tableau:

$$k = \sum'_i j \sum'_k (i,j) = \sum'_i k(i) = \sum'_j k(j) .$$

Mais c'est en réalité le tableau des fréquences associées au tableau précédent qui va être l'objet de notre attention:

$$f(i,j) = k(i,j)/k$$
$$f(i) = k(j)/k = \sum_i f(i,j)$$
$$f(i) = k(j)/k = \sum_i f(i,j) .$$

Les $f(i)$ et $f(j)$ sont les fréquences marginales. L'analyse factorielle a pour but l'étude des profils des n individus donnés, par rapport aux p variables de références, c'est à dire des ensembles de mombres:

$$(f(i,1)/f(i), f(i,2)/f(i), \ldots, f(ip)/f(i)) ,$$

l'indice n pouvant varier de 1 à n. On remarque, en passant que chacun des nombres figurant dans la liste ci-dessus est identique.

(Equivalence distributionnelle).

Retour au menu

Compte tenu de ces notations et remarques, l'analyse factorielle des correspondances consiste à fabriquer des fonctions $F(i)$ et $G(j)$, appelées facteurs définies, le premier ensemble (observations) pour les Fk et sur le second (variables) pour les G_k, et satisfaisant à l'équation suivante:

$$(1) \qquad f(i):f(i)\,f(j)\,[1+\textstyle\sum_k \lambda_k^{-\frac{1}{2}} F_k(i)\,G_k(j)]\,.$$

Cette formule peut être considérée indifféramment comme une décomposition du tableau initial, qui est une fonction de deux variables, à l'aide de fonction qui, elles, ne dépendant que d'une seule variable, ou bien comme une formule de reconstitution des données à partir des résultats qui sont les facteurs.

En outre ces facteurs doivent satisfaire aux contraintes suivantes:

$$(2) \qquad \textstyle\sum_i (\lambda_k\lambda_{k'})^{-\frac{1}{2}} f(i)\,F_k(i)\,F_{k'}(i) = \delta_{kk'}$$

$$(3) \qquad \textstyle\sum'_j (\lambda_k\lambda_{k'})^{-\frac{1}{2}} f(j)\,G_k(j)\,G_{k'}(j) = \delta_{kk'}$$

où $\delta_{kk'}$ est le symbole de Kronecker égal à 0 si $k \neq k'$ ou à 1 si $k = k'$.

Dans ce dernier cas ces formules s'écrivent encore:

$$(2') \qquad \textstyle\sum_i f(i)\,F_k^2(i) = \lambda_k$$

$$(3') \qquad \textstyle\sum_j f(j)\,G_k^2(j) = \lambda_k\,.$$

Ces nombres $\lambda_{k'}$ qui représentent donc les variances, au sens large, des facteurs, sont ce que l'on appelle en mécanique les moments d'inertie du nuage I des n points ayant pour coordonnées les n profils définis ci-dessus et pour masses les $f(i)$. De plus, les formules étant parfaitement symétriques entre i et j, ces nombres K sont aussi les moments d'inertie du nuage J des p profils des variables ayant pour masse les $f(j)$.

Cette terminologie se justifie à condition de préciser que la distance entre deux points x et y de 1 n'est pas donnée par la formule usuelle.

$$d(x, y) = \textstyle\sum'_j (xj - yj)^2$$

mais par la formule du x^2:

$$d(x, j) = \textstyle\sum'_j (xj - yj)^2 / f(j)\,.$$

Ce qui implique que le produit scalaire soit défini par:

$$\langle x, y\rangle = \textstyle\sum_j xj\,yj / f(j)$$

Retour au menu

quelles sont les coordonnées yj du centre du gravité G du nuage I des observations?

$$yj = \sum_i' f(i)\,[f'(i,j)/f(i)] = \sum_i' f(i,j) = f(i)\,.$$

(Par la normalisation initiale, masse totale du nuage est égale à 1). Divisons les deux termes de la formule (1) par $f(i)$:

$$f(i,j)/f(i) = f(j) + \sum_k' F_k'(i)_k^{-\frac{1}{2}} f(j)\,G_k(j)\,.$$

Le profil de i s'obtient par adjonction aux coordonnées du centre de gravité d'une somme de termes que l'on peut interpréter comme une somme vectorielle: les $k(i)$ étant les coordonnées de i dans le système d'axes constitué par ces $\sqrt{k}(j) = k^{\frac{1}{2}} f(j)\,G_k(j)$.

Faisons le produit scalaire de deux de ces axes:

$$\sqrt{k}, \sqrt{k'} = \sum_j \sqrt{k}(j)\,\sqrt{k'}(j)/f(j) = \sum_j' (k'k')^{-\frac{1}{2}}$$

$$f(j)\,G_k(j)\,G_{k'}\,,\,(j)\,.$$

Or cette somme vaut $\delta_{kk'}$ d'après la contrainte (3) à laquelle sont assujeties les fonctions G_k.

Or l'unique système d'axes ayant cette propriété est celui des axes principaux d'inertie du nuage I, ce que nous voulions démontrer. Pratiquement ces axes s'obtiennent par la recherche des vecteurs propres et des valeurs propres de la matrice S dont les termes sont donnés par la formule:

$$Sjj' = \sum_i f(i,j)\,f(i,j)/f(i)\,f(j)\,f(j)\,.$$

Une fois ces calculs faits, on pourra représenter graphiquement les positions de chacun des individus dans le système des deux premiers facteurs extraits par ordre d'importance décroissante, importance indiquée par les coefficients I, λ_k, où λ_k est la valeur propre associée au k-èmes facteurs; puis on s'interessera aux axes nn. 1 et 3 etc.

Complétons ce bref exposé par la description de quelques propriétés supplementaires. Il existe une formule canonique de passage entre l'ensemble I des observations et l'ensemble J des variables:

$$\sum_j' G_k(j) f(i,j)/f(i) = k^{\frac{1}{2}} F_k(i)$$

$$\sum_j' F_k(i)\,F(i,j)/f(j) = k^{\frac{1}{2}} G_k(j)\,.$$

En termes géométriques celà pourra s'appeler principe barycentrique:

Sur l'axe factoriel k, le point i d'abscisse $F_k(i)$ est l'homothétique dans le

Retour au menu

rapport $k^{\frac{1}{2}}$ du barycentre des points j, d'abscisse $G_k(j)$ affectés des masses $f(i, j)/f(i)$.

Autre possibilité de cette méthode: pour chaque point i on peut calculer sa contribution à la part d'inertie exprimée par un axe k; c'est la quantité $f(i)F_k^2(i)$ intervenant dans (2').

Nous avons enfin une troisième facilité à notre disposition. Supposons qu'après une première série d'observations déjà analysées se présentent quelques observations supplémentaires. On peut supposer que l'adjonction de ces nouveaux points ne perturbe pas radicalement la disposition des axes principaux d'inertie et calculer leur position dans ce système d'axe par la formule:

$$F_k(S) = k^{-\frac{1}{2}} jG_k(j) f(S_1 j)/f(S) .$$

De même s'il s'agissait d'une variable supplémentaire:

$$G_k(t) = k^{-\frac{1}{2}} iF_k(i) f(i, t)/f(t) .$$

Ce calcul peut également être utile dans le cas de données douteuses ou pour des observations d'un poids $f(i)$ excessif. Abordons, pour conclure cette première partie, les problèmes qui se posent le plus souvent dans la pratique:

1) *Quel est le nombre d'axes à extraire?*

Dans la plupart des cas on n'en extrait qu'un nombre restreint, variant entre 5 et 10, et fonction de la puissance de l'ordinateur dont on dissipe, et il est bien rare que l'on réussisse à les interprêter tous.

Il existe cependant des épreuves de validité basées sur des simulations obtenues par tirages au hasard de tableaux « analogues » au tableau initial.

2) *Les problèmes de stabilité: à partir de quelle table d'échantillons obtient-on des axes stables?*

Ceci peut encore se traiter expérimentalement en réduisant l'échantillon dont on dispose de 10%, puis de 2% etc. et en calculant les corrélations entre ces nouveaux facteurs et ceux qui sont issus de l'échantillon initial. Si on enregistre des bonnes corrélations c'est que le nombre d'observations faites est insuffisant, dans ce cas contraire les conclusions risquent d'être fausses ou même impossibles à tirer.

3) *Les problèmes des données initiales sont tellement importants qu'ils nécessiteraient de longs developpements.*

Donnons-en seulement les deux grands principes:

a) Homogénéité: nous entendons par là que deux nombres figurant dans le tableau initial doivent être, autant que possible, des grandeurs comparables.

Retour au menu

En particulier les tableaux où figurent des variables mesurées par des unités différentes, par exemple des longueurs et des poids doivent faire l'objet d'un traitement préalable.

Dont le plus simple consiste à faire des classes de valeurs. A plus forte raison si l'un a des mélanges de variables qualitatives et quantitatives.

Ces dernières doivent impérativement être découpées en classes dont chacune d'elles sera considérée comme une variable qualitative.

b) Exhaustivité: nous ne voulons pas dire qu'il faille faire toutes les observations possibles dans un domaine, mais que l'échantillon retenu soit bien représentatif de toutes ces variations, réellement existantes dans le domaine considéré. Ceci implique que ce domaine soit clairement délimité, par des frontières aussi naturelles que possible.

2ème partie

UNE EXPERIENCE D'ANALYSE FACTORIELLE EN PHYTOTECOLOGIE.

Il s'agit de l'approfondissement d'une étude faite par un de nos collègues F. ROMANE du centre d'études phytosociologiques et écologiques (C.E.P.E.) de Montpellier (France), étude portant sur 443 relevés appartenant à un transect allant de Montpellier au Vignan suit une direction Nord-Nord-Est à Sud Ouest.

Outre un répertoire de 500 espèces végétales ce travail comporte l'enregistrement de 26 variables écologiques telles que l'altitude, l'exposition, la géomorphologie, la nature du sol, la nature de la roche mère, le degré de recouvrement, la distance à la mer, etc.

Le découpage nécessaire de ces 26 variables écologiques en classe a donné 257 variables logiques que nous appelleront modalités. On voit que ces données sont d'une taille respectable.

Pour des raisons d'encombrement ROMANE avait préféré faire une sélection raisonnable de 120 espèces parmi les 500 répertoriées. Début de son travail n'était pas la découverte des relatives entre écologie et floristique, mais la comparaison de diverses méthodes d'analyse multidimensionnelle notamment l'analyse en composantes principales et l'analyse des correspondances, et il concluait en faveur de cette dernière, aussi bien en ce qui concerne le tableau des données floristiques que celui des données écologiques. Ce dernier est constitué de la façon suivante: chaque ligne représente une espèce et chaque colonne une des 257 modalités décrivant le milieu. A l'intersection de la ligne i et de la colonne j de ce tableau figure le nombre de relevés contenant l'espèce i et présentant la modalité j on constatera en passant qu'on a bien l'homogénéité souhaitée dans notre première partie. Le but de notre travail est double; tout d'abord vérifier que

Retour au menu

la sélection opérée sur les espèces n'a pas trop entaché le résultat, ensuite comparer les résultats fournis par l'analyse de ce tableau à ceux que l'on obtient avec le tableau floristique, où chaque case (i, j) contient un 1 ou un 0 suivant que l'espèce i figure ou non dans le relevé j.

Cette comparaison doit interesser les spécialistes puisque la théorie défendue par M. GUINACHER, chef de file, de la phytosociologie moderne, est que la floristique permet une description du milieu ou moins précise, mais souvent meilleure que celle qui est fournie, par des variables écologiques. Nous résumerons nos résultats à l'aide des corrélations entre les cinq premiers facteurs issus des différentes analyses.

1) Comparaison des analyses écologiques sur 120 espèces et sur 500 espèces.

(Les corrélations on été calculées à partir des résultats sur les relevés). E_k désigne le K-ème axe de l'analyse sur 500 espèces tandis que F_k est le K-ème axe de l'analyse sur 120 espèces.

	E_1	E_2	E_3	E_4	E_5
F_1	0.993	—0.045	—0.094	0.001	—0.008
F_2	—0.030	0.960	—0.080	0.039	0.306
F_3	0.029	0.009	0.824	—0.403	—0.059
F_4	0.054	0.039	0.230	0.669	0.330
F_5	0.091	0.141	0.173	0.077	0.704

On constate d'excellentes corrélations entre les facteurs homologues.

2) Comparaison des analyses écologiques et floristiques.

(Les corrélations ont été calculées à partir des résultats sur les espèces). E_k désigne le k-ème axe de l'analyse écologique tandis que F_k est le K-ème axe de l'analyse floristique.

	E_1	E_2	E_3	E_4	E_5
F_1	0.908	0.141	0.046	0.141	
F_2	0.007	0.567	0.212	0.387	0.188
F_3	0.094	0.138	0.877	0.118	0.009
F_4	0.074	0.211	0.123	0.746	0.085
F_5	0.040	0.014	0.006	0.291	0.519

On observe, encore, de meilleurs corrélations entre facteurs homologues qu'entre facteurs des rangs différents, cependant deux des premières sur cinq sont anormalement basses: celle entre E_5 et F_5, ce qui se conçoit si l'on admet

Retour au menu

que la dégradation de ces coefficients soit assez rapide quand on passe du 1-er au 5-ème facteur, et celle entre E_2 et F_2, ce qui est plus grave.

Respectant notre plan de travail notre conclusion à cette deuxième partie se fera en deux temps; en ce qui concerne la réduction du nombre des espèces on peut s'estimer très satisfait de la sélection opérée. Celle-ci procure donc un grand avantage à la fois sur la taille de l'ordinateur à employer et sur le temps de calcul; elle ne pert pratiquement pas d'informaton si l'on prend la précaution de mettre les espèces en éléments supplémentaires.

En ce qui concerne la comparaison des deux types d'analyses écologiques et floristiques, nous dirons que l'hypothèse de parfaite intégration des variables du milieu par les espèces végétales est assez bien vérifiée; néamoins il reste à expliquer quelques divergences entre les deux séries de facteurs, et celà ne peut se faire que par l'interprétation des axes. L'avantage sera donné à la méthode la plus précise quant aux variations décrites, et connues par ailleurs. Nous ne pouvons donner ici ces interprétations car elles ne sont pas terminées: ce travail est en effet considérable vu la taille du problème.

On remarquera enfin que, là, encore, les résultats propres à l'une des analyses peuvent aussi être fournis par l'autre, à l'aide des éléments supplémentaires: dans l'analyse écologique les relevés peuvent être considérés comme des variables supplémentaires ayant pour valeur 0 ou 1 selon les espèces, dans l'analyse floristique les modalités des variables écologiques peuvent être mises en variables supplémentaires également.

BIBLIOGRAPHIE

BENZECRI J. P. et COLL., *L'analyse des données*, 2 vol., p. 620; Dunod, Paris, 1973.
GUNACHET M., *La phytosociologie*, p. 228, Masson, Paris, 1973.