RECHERCHE DES RACINES CARACTERISTIQUES EN ANALYSE MULTIDIMENTIONNELLE

PAR: ALBERT DIONNE,
UNIV. LAVAL

I - INTRODUCTION

L'étude mathématique du problème des racines caractéristiques est souvent dégoûtante pour l'étudiant en algèbre matricielle, la raison en étant que la solution du problème ne semble pas avoir d'applications pratiques ni même être sujette à interprétation concrète.

Le but de cet article est d'illustrer la très grande utilisation qu'en font les statisticiens dans trois techniques majeures d'analyse multidimensionnelle: analyse en composantes principales, analyse discriminante et analyse canonique.

II - POSITION DU PROBLEME EN ALGEBRE MATRICIELLE

Soit une matrice carrée A d'ordre p. On cherche un vecteur colonne v non nul et un scalaire λ tel que

$$Av = \lambda v \qquad (1)$$

Du point de vue géométrique, le problème consiste à trouver un vecteur v qui, multiplié par A, conserve la même orientation mais peut changer de direction (si $\lambda < 0$) et de longueur (si $|\lambda| \neq 1$).

L'expression en (1) peut s'écrire de façon équivalente

$$(A - \lambda I)v = 0 \qquad (2)$$

Ce système d'équations linéaires homogènes possède une solution non triviale si et seulement si

$$|A - \lambda I| = 0 \qquad (3)$$

Le développement de cette dernière expression donne lieu à une équation en λ de degré p, appelée équation caractéristique de A, dont les solutions $\lambda_1, \lambda_2, \ldots, \lambda_p$ sont les racines caractéristiques ou les valeurs propres de A. Pour chaque λ_i, on pourra résoudre le système d'équations

$$(A - \lambda_i I)v_i = 0 \qquad (4)$$

et trouver le vecteur propre v_i correspondant à λ_i.

III - APPLICATIONS A LA STATISTIQUE

(i) Analyse en composantes principales

À partir d'un ensemble d'objets mesurés sur plusieurs variables X_1, X_2, ..., X_p, l'analyse en composantes principales permet surtout de

a) trouver la dimension selon laquelle l'ensemble des variables possède le maximum de variation,

b) réduire la dimension de la matrice des données en regroupant des variables.

La première composante principale sera cette combinaison linéaire des variables originales

$$F_1 = v_1X_1 + v_2X_2 + \ldots + v_pX_p \qquad (5)$$

telle que la variance de F_1 soit maximale. Il est possible (Tatsuoka 1971, p. 116), à l'aide du calcul différentiel, de montrer que ceci revient à chercher la solution du système d'équations linéaires homogènes

$$(S - \lambda I)v = 0 \qquad (6)$$

dans lequel S représente la matrice d'ordre p des sommes de carrés et des produits croisés des variables X_1, X_2, ..., X_p. Si l'on a pris soin de normer le vecteur v, c'est-à-dire de le rendre unitaire $(v' \cdot v) = 1$), alors on a tout de suite

$$\text{Var } (F_1) = \lambda_1 \qquad (7)$$

λ_1 étant la plus grande racine caractéristique de la matrice S.

(ii) Analyse discriminante

Des objets issus de plusieurs groupes connus sont mesurés sur plusieurs variables X_1, X_2, ..., X_p. On recherche une combinaison linéaire des X

$$Y = v_1X_1 + v_2X_2 + \ldots + v_pX_p \qquad (8)$$

telle que les nouveaux scores Y séparent au mieux les groupes. Cette combinaison linéaire s'appelle la fonction discriminante; elle est calculée sur la base de l'appartenance connue de chaque objet à l'un des groupes.

Les principaux objectifs poursuivis par l'analyse discriminante sont:

a) savoir si les groupes diffèrent significativement les uns des autres,

b) analyser la nature des différences entre les groupes, s'il y a lieu,

c) découvrir les variables qui contribuent le plus à différencier les groupes les uns des autres.

Pour séparer au mieux les groupes, il faut que le critère de discrimination $\lambda = \frac{\text{somme des carrés entre les groupes}}{\text{somme des carrés à l'intérieur des groupes}}$ soit maximal. Pour des scores X_i donnés, λ est une fonction des coefficients inconnus $v_1, v_2, \ldots, v_p$. Tatsuoka (1971, p. 161) montre que le maximum s'obtient par calcul différentiel en résolvant le système d'équations linéaires homogènes

$$(W^{-1}B - \lambda I)v = 0 \qquad (9)$$

où W est la matrice d'ordre p des sommes de carrés et des produits croisés calculée à l'intérieur des groupes, et B est la matrice de même nature calculée entre les groupes.

Si λ_1 représente la valeur propre la plus grande de (9), alors le vecteur propre correspondant v_1 permet de trouver la combinaison linéaire des X qui sépare le mieux les groupes.

(iii) Analyse canonique

Étant donné des objets mesurés à la fois sur un ensemble de variables $X_1, X_2, \ldots, X_p$ et sur un autre ensemble de variables $Y_1, Y_2, \ldots, Y_q$, l'analyse canonique sert à

a) tester si les deux ensembles de variables sont statistiquement indépendants,

b) déterminer l'importance de chacune des variables pour expliquer l'association entre les deux ensembles de variables.

Pour atteindre ces buts, il faut d'abord chercher une combinaison linéaire de chaque ensemble de variables

$$\begin{aligned} Z &= u_1X_1 + u_2X_2 + \ldots + u_pX_p \\ W &= v_1Y_1 + v_2Y_2 + \ldots + v_pY_p \end{aligned} \qquad (10)$$

telle que la corrélation de ces deux nouvelles variables Z et W soit maximale. Or il est possible (Tatsuoka 1971, p. 184) de démontrer que la découverte des coefficients u_i et v_j s'effectue en résolvant le système d'équations linéaires homogènes

$$(S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} - \lambda^2 I)v = 0 \qquad (11)$$

dans lequel S_{xx} et S_{yy} sont des matrices de même nature qu'en (6), et $S_{xy} = S'_{yx}$ est la matrice de format p sur q des produits croisés des X et des Y.

La racine caractéristique λ_1^2 la plus élevée indique la valeur maximale du carré du coefficient de corrélation de Z et W.

IV - CONCLUSION

Les techniques multidimensionnelles décrite ici brièvement visent toutes trois à maximiser un certain critère d'optimalité: rendre maximale la variance d'une combinaison linéaire des variables originales, séparer au mieux les groupes d'objets, maximiser le carré de la corrélation entre deux ensembles de variables.

Dans tous les cas, la solution s'obtient en résolvant un système d'équations linéaires homogènes de la forme

$$(A - \lambda I)v = 0 \qquad (2)$$

Voir (6), (9) et (11).

Dire que la recherche des racines et vecteurs propres n'est qu'un exercice purement mathématique sans utilité est loin d'être conforme à la réalité du statisticien.

Référence: Tatsuoka, Maurice M.; Multivariate Analysis, Wiley, 1971.