DATA ANALYSIS AND INFORMATICS, IV
E. Diday et al. (editors)

REGIONS DE CONFIANCE EN ANALYSE FACTORIELLE

Gilbert SAPORTA
Département Mathématiques et Informatique
C.N.A.M.
292 rue Saint-Martin 75141 PARIS Cedex 03

Gérard HATABIAN
Direction des Etudes et Recherches
EDF
92141 CLAMART

On donne les principes et des formules simples permettant de tracer des ellipses de confiance autour des points représentatifs des catégories de variables qualitatives, sur des plans factoriels, tant en analyse des correspondances sur table de contingence qu'en analyse des correspondances multiples, ou en analyse en composantes principales. Les résultats sont comparés avec ceux obtenus par des techniques de rééchantillonnage.

1 INTRODUCTION

L'utilisation de zones de confiance autour des points d'un graphique d'analyse factorielle répond à deux préoccupations principales relativement indépendantes.
La première est l'inférence statistique usuelle : les données analysées ne constituent qu'un échantillon extrait d'une population plus vaste ; les résultats obtenus sont alors des estimations des caractéristiques de la population. Fournir pour chaque point estimé un domaine de confiance est alors une préoccupation légitime. On sait que sous des hypothèses de normalité ces domaines sont des ellipsoïdes.
La deuxième concerne la validité des résultats obtenus et se refère plutôt à la notion de stabilité vis à vis de perturbations des données ; l'hypothèse d'échantillonnage n'est alors nullement nécessaire.
Les méthodes utilisées relèvent essentiellement de deux approches, l'une classique fondée sur la normalité, l'autre sur la simulation de tableaux de données analogues à celui analysé : les méthodes de rééchantillonnage (bootstrap , jack-knife cf (3)) en font partie.

2 "L'ETAT DE L'ART"

Les travaux dans ce domaine sont encore peu nombreux et leur intégration dans des logiciels standard extrêmement rare On notera en "multidimensional scaling" J.O. RAMSAY (11) qui utilise une hypothèse de log-normalité pour les distances inter-individuelles : il en déduit un positionnement des individus ainsi que des ellipses de confiance autour de ces positions en utilisant la méthode du maximum de vraisemblance et ses propriétés asymptotiques (programme MULTISCALE).

Très récemment WEINBERG et al.(12) ont utilisé le jack-knife et le bootstrap dans les mêmes conditions et ont comparé leurs résultats avec ceux de RAMSAY.

En analyse des correspondances on note dans l'ouvrage de J.P. BENZECRI (2) page 215 la proposition suivante : tracer autour de chaque point représentatif d'une catégorie un cercle de rayon $(6/n_j)^{1/2}$ où n_j est l'effectif de la catégorie considérée, lorsque les facteurs sont normalisés. Le tracé des ellipses d'inertie de

sous-nuages proposé par BORDET est parfois utilisé comme aide à l'interprétation (ces ellipses ont pour demi-axes les racines des valeurs propres de la matrice d'inertie du sous-nuage).

Dans une optique d'étude de stabilité A.GIFI (6) et J. MEULMAN (9) utilisent la "méthode delta" et le bootstrap pour construire des ellipses à 95 % autour des points issus de graphiques plans d'analyse des correspondances. La "méthode delta" liée au jack-knife infinitésimal (cf (3)) consiste ici à écrire que les coordonnées $\underline{y}$ d'un point catégorie sont une fonction Φ (N) différentiable où N est le tableau de contingence relatif à n observations. N est une réalisation d'une variable multinomiale X d'espérance μ et de matrice de variance Σ que l'on estime à partir des nij. Comme $\sqrt{n}\ (X-\mu)$ converge vers une loi normale $N(0\ ;\ \Sigma)$ il en est de même de $\sqrt{n}\ (\underline{y} - \Phi(\mu))$ comme le montre un développement de Taylor au premier ordre d'où le nom de la méthode. La matrice de variance de $\sqrt{n}\ \underline{y}$ est donnée par

$$\Delta\ \Sigma\ \Delta'$$

où Δ est la matrice des dérivées partielles de $\underline{y}$ par rapport aux termes de X calculée en μ. Ces dérivées peuvent être obtenues numériquement en faisant varier d'une unité (d'où le lien avec le jack-knife) chaque terme de N.

Le bootstrap utilisé par GIFI et MEULMAN revient à simuler B fois la loi multinomiale d'effectif n et de probabilités $p_{ij} = \frac{n_{ij}}{n}$ seul n est fixé : on estime alors la matrice de variance des coordonnées d'une catégorie à l'aide des B réalisations et on trace les ellipses de confiance à l'aide de la loi normale correspondante mêlant ainsi stabilité et inférence comme dans (12) mais le recours à la loi normale n'est pas justifié.

M.J. GREENACRE (8) (chapitre 8) effectue lui, des rééchantillonnages bootstrap séparement pour les lignes ou les colonnes d'un tableau de contingence, ce qui revient à simuler autant de multinomiales que de lignes ou de colonnes de N : on fixe donc ici les marges en ligne ou en colonne. Cet auteur n'utilise pas d'hypothèse de normalité donc ne construit pas de régions de confiance elliptiques : il trace autour de chaque point l'enveloppe convexe de ses "réplications". Cette enveloppe n'étant pas robuste lorsque le nombre de réplications augmente, il suggère de prendre la technique de "pelage" étudiée par GREEN (7). GREENACRE propose une distinction intéressante entre stabilité interne étudiée par le jack-knife liée à l'influence des individus et stabilité externe étudiée par le bootstrap. Remarquons pour terminer ce bref panorama l'absence de propositions en ACP excepté un article de K.R. GABRIEL. (5).

3 POUR UNE UTILISATION RAISONNEE DE LA LOI NORMALE.

Notre travail a porté sur la détermination d'ellipses de confiance dans des plans factoriels autour des points représentatifs des catégories de variables nominales tant en analyse en composantes principales qu'en analyse des correspondances. L'usage de la loi normale multidimensionnelle que nous avons utilisé concurremment avec le bootstrap possède en effet, dans ce cas des justifications théoriques solides : le principe barycentrique fait que le point représentatif d'une catégorie concernant n_j individus est le centre de gravité $\underline{g}$ des n_j points représentatifs des individus ; si n_j est suffisamment grand (quelques dizaines comme on le sait..) le théorème central limite multidimensionnel s'applique et $\underline{g}$ est distribué approximativement selon une loi normale.

Ce raisonnement suppose bien entendu de se placer dans l'optique inférentielle de l'estimation de la position d'un point moyen : la similitude des résultats obtenus par rééchantillonnage montre en fait que l'hypothèse de normalité peut avantageusement être utilisée dans une optique d'étude de stabilité : l'économie de calcul est alors considérable.

4 ELLIPSES DE CONFIANCE POUR DES CATEGORIES DE VARIABLES NOMINALES EN ACP ET EN ACM.

4-1 Rappels de théorie normale.

Si $\underline{x}$ est un vecteur aléatoire suivant une loi $N_p\ (\underline{\mu}\ ;\ \Sigma)$ le centre de gravité d'un nuage de n réalisations indépendantes de $\underline{x}$ est tel que :

$$\sqrt{n}\ (\underline{g} - \underline{\mu}) \text{ suit une loi } N_p\ (0\ ;\ \Sigma)$$

$$\text{et } n\ (\underline{g} - \underline{\mu})'\ \Sigma^{-1}\ (\underline{g} - \underline{\mu}) \text{ suit une loi } \chi^2_p$$

On en déduit que l'ellipsoïde de confiance pour $\underline{\mu}$ si Σ est connu a pour équation :

$$(\underline{g} - \underline{\mu})'\ \Sigma^{-1}\ (\underline{g} - \underline{\mu}) \ < \ \frac{\chi^2_{p\ (1-\alpha)}}{n}$$

Lorsque Σ est inconnue, on utilise le fait que $(n-1)\ (\underline{g}-\underline{\mu})'\ V^{-1}(\underline{g} -\underline{\mu})$ suit un $T^2_p\ (n-1)$ de Hotelling où V est la matrice de variance du nuage.

L'ellipsoïde de confiance a alors pour équation :

$$(\underline{g} - \underline{\mu})\ V^{-1}\ (\underline{g} - \underline{\mu}) \ \leqslant \ \frac{p}{n-p}\ F_{p\ ;\ n-p\ \ (1-\alpha)}$$

Pour des zones dans le plan, on prendra p = 2. L'utilisation des fractiles du χ^2 au lieu de ceux de la variable de Fisher-Snedecor conduit à sous-estimer le domaine de confiance si $n < 50$. Pour $n > 50$ on peut prendre comme second membre :

$\frac{6}{n}$ pour une ellipse à 95 % et $\frac{4.3}{n}$ pour une ellipse à 90 %.

4.2 Ellipse de confiance pour une catégorie supplémentaire.

Considérons une variable supplémentaire qualitative dans une ACP ou une ACM (correspondance multiple). Soit une de ses catégories concernant n_j individus. Il est alors facile de calculer le centre de gravité de ces n_j individus et la matrice de variance-covariance V des coordonnées de ces individus sur deux axes factoriels. Comme les n_j individus sont pris "sans remise" parmi les n individus étudiés, la matrice ^{j} de variance des coordonnées du centre de gravité doit être multipliée par le coefficient d'exhaustivité $\frac{n-n_j}{n-1}$ et l'ellipse de confiance a donc pour équation :

$$1 \qquad (\underline{g} - \underline{\mu})' \, V^{-1} \, (\underline{g} - \underline{\mu}) < \frac{2}{n_j - 2} \, \frac{n - n_j}{n-1} \, F_{2 \; ; \; n_j - 2} \, (1-\alpha)$$

Utiliser un cercle de rayon $(6/n_j)^{1/2}$ constitue donc une approximation grossière pour diverses raisons : les fractiles du χ^2 ne sont pas valables pour les petites tailles ; rendre sphérique le nuage dans son ensemble ne rend pas les sous-nuages sphériques, ils ont en tout état de cause une variabilité inférieure en moyenne à celle du nuage entier et leurs directions principales ne sont pas nécessairement les mêmes. Enfin, le coefficient d'exhaustivité joue un rôle d'autant plus grand que la catégorie est à fort effectif : on remarquera que si $n_j = n$ on est sûr que $\underline{g} = \underline{\mu} = \underline{0}$ la variabilité pour $\underline{g}$ est alors nulle.

La construction d'ellipses de confiance étend la technique des valeurs-tests utilisée dans le logiciel SPAD (10) où l'on vérifie axe par axe et indépendemment si le point moyen d'une catégorie a une coordonnée différente de zéro, c'est à dire diffère de la moyenne générale. On a alors ici des résultats plus précis sur la dispersion de chaque catégorie et la corrélation locale des composantes.

L'exemple suivant montre une application à une question supplémentaire (la préférence partisane) d'une enquête AESOP (Association pour l'Etude des Structures de l'Opinion Publique voir (1)) portant sur 1 012 individus : les variables actives sont des thèmes de conflit notés sur une échelle de 1 à 5 (depuis "pas du tout d'accord" à "tout à fait d'accord"), la méthode utilisée est l'ACP.

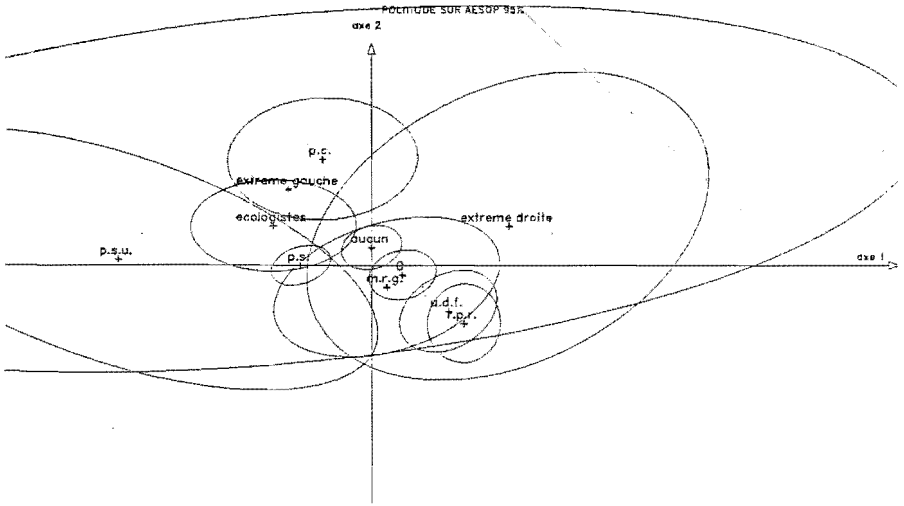

Figure 1. Ellipses à 95 % déduites de la loi normale.

On constate une très grande dispersion des catégories à faible effectif (extrême-droite concernait 8 individus), une séparation nette entre PS, PC, MRG et le groupement UDF-RPR qui apparait assez homogène. On note que les composantes principales apparaissent non corrélées quelle que soit la préférence partisane.

Une approche basée sur le bootstrap a également été utilisée : on a procédé à B tirages avec remise à l'intérieur de chaque catégorie de la variable qualitative étudiée d'où à chaque fois B nouveaux points moyens. La figure suivante donne pour chaque catégorie la moyenne de ses B points moyens et l'ellipse homothétique de l'ellipse d'inertie des B centres qui contient 95 % d'entre eux.

Il ne s'agit donc pas ici d'ellipses de confiance au sens inférentiel qui devraient contenir 95 % de tous les centres et pas seulement de ceux obtenus par réechantillonnage. On note que ces ellipses ont mêmes orientations que celles obtenues par la loi normale même pour les catégories à faible effectif. Les résultats sont donc concordants et les expériences réalisées n'ont montré que des différences minimes entre les cas B = 300 et B = 50.

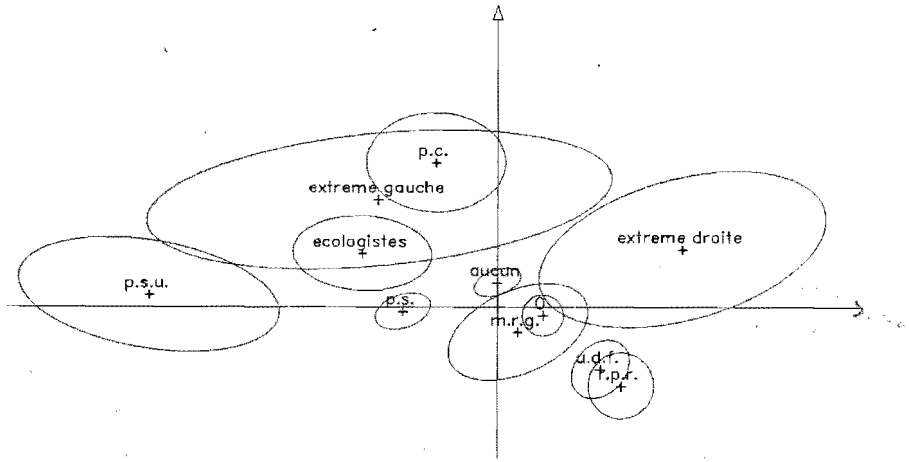

Figure 2. Ellipses homothétiques des ellipses d'inertie contenant 95 % des réplications des barycentres.

B = 50

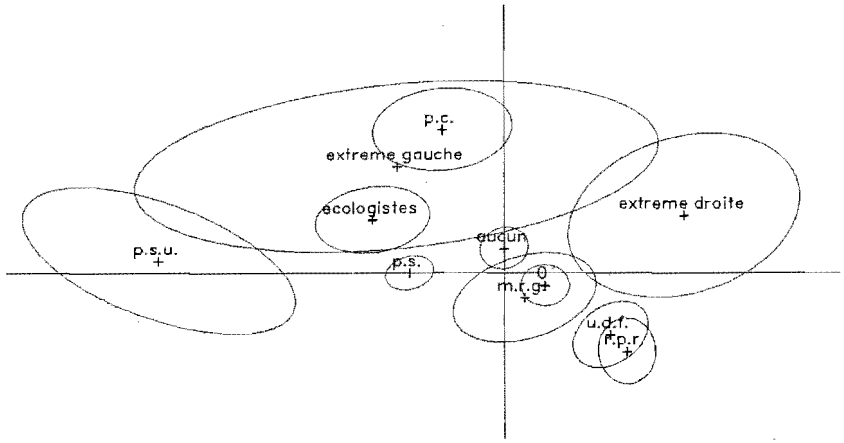

Figure 3. Même principe avec B = 300

4. 3 Ellipses de confiance pour des catégories actives en analyse des correspondances multiple.

Les coordonnées des individus sur les axes factoriels ne sont pas modifiées par des variations des variables supplémentaires : cela justifie partiellement les procédés du paragraphe précédent (il faudrait en toute rigueur également des individus supplémentaires). L'extension aux variables actives soulève des problèmes théoriques en particulier quand on veut comparer des résultats de réechantillonnage puisque les axes principaux varient avec le tableau simulé. Lorsque ces axes ne sont pas stables, ce qui se produit si les valeurs propres sont confondues ou très proches, on ne peut comparer les configurations obtenues et il convient d'agir avec précaution. J. MEULMAN (9) propose d'effectuer avant les calculs d'ellipses des rotations procustéennes de façon à amener chacune des B configurations le plus près possible de la configuration initiale.
Lorsque les axes sont stables, il suffit de projeter en éléments supplémentaires les moyennes des échantillons bootstrappés sur le plan de référence défini par le tableau initial, ce qui évite de recalculer les axes principaux (voir (8)).

5. ELLIPSES DE CONFIANCE EN ANALYSE DES CORRESPONDANCES SUR TABLEAU DE CONTINGENCE.

Pour obtenir des ellipses de confiance autour des points issus de la représentation simultanée des lignes et des colonnes d'un tableau de contingence, il suffit de se ramener au cas précédent en utilisant l'équivalence entre l'AFC du tableau disjonctif et l'AFC du tableau de contingence.

Les conventions de normalisation variant selon les ouvrages et les logiciels, nous commencerons par définir les notres.

5. 1. Notations

X_1 et X_2 désignent les tableaux d'indicatrices associés à deux variables qualitatives et $N = X_1' X_2$ le tableau de contingence. On pose $D_1 = X_1' X_1$ et $D_2 = X_2' X_2$.
Les coordonnées des lignes $\underline{a}$ et des colonnes $\underline{b}$ de N vérifient pour la représentation simultanée

$$D_1^{-1} N D_2^{-1} N' \underline{a} = \lambda \underline{a} \qquad \frac{1}{n} \underline{a}' D_1 \underline{a} = \lambda$$

$$D_2^{-1} N' D_1^{-1} N \underline{b} = \lambda \underline{b} \qquad \frac{1}{n} \underline{b}' D_2 \underline{b} = \lambda$$

$$\sqrt{\lambda}\ \underline{b} = D_2^{-1} N' a \qquad \sqrt{\lambda}\ \underline{a} = D_1^{-1} N \underline{b}$$

Considérons maintenant l'analyse du tableau disjonctif $(X_1 \mid X_2)$. Ses valeurs propres μ sont telles que :

$$\mu = \frac{1 + \sqrt{\lambda}}{2}$$

Le vecteur $\underline{z}$ des coordonnées des n individus sur un axe est alors proportionnel à $X_1 \underline{a} + X_2 \underline{b}$.

Pour que les coordonnées des points lignes et des points colonnes de l'analyse du tableau N coïncident avec les barycentres des catégories de l'ACM, il faut prendre :

$$\underline{z} = \frac{1}{2\mu} (X_1 \underline{a} + X_2 \underline{b})$$

On a alors $\underline{a} = D_1^{-1} X_1' \underline{z}$ et $\underline{b} = D_2^{-1} X_2' \underline{z}$

5. 2. Matrice de variance et covariance pour une catégorie.

Considérons la catégorie correspondant à la $i^{ème}$ ligne de N et qui concerne n_i individus dont n_{ij} sont également dans la catégorie j de la 2ème variable. Ces n_{ij} individus ont pour coordonnées dans l'ACM $\frac{1}{1+\sqrt{\lambda}} (a_i + b_j)$

La variance des coordonnés des n_i points sur l'axe 1 vaut alors :

$$V_1 = \frac{1}{n_{i.}} \sum_j \frac{1}{(1+\sqrt{\lambda_1})^2} n_{ij} (a_i^{(1)} + b_j^{(1)})^2 - (a_i^{(1)})^2$$

et on a une expression identique pour V_2.

La covariance entre les composantes 1 et 2 vaut pour la catégorie i:

$$COV = \frac{1}{n_{i.}} \sum_j \frac{1}{(1+\sqrt{\lambda_1})(1+\sqrt{\lambda_2})} n_{ij} \left(a_i^{(1)} + b_j^{(1)}\right) \left(a_i^{(2)} + b_j^{(2)}\right)$$

En développant la somme et en utilisant la formule de transition, on trouve finalement :

$$V_1 = \frac{1}{(1+\sqrt{\lambda_1})^2} \left[\sum_j \frac{n_{ij}}{n_{i.}} (b_j^{(1)})^2 - \lambda_1 (a_i^{(1)})^2 \right]$$

2

$$COV = \frac{1}{(1+\sqrt{\lambda_1})(1+\sqrt{\lambda_2})} \left[\sum_j \frac{n_{ij}}{n_{i.}} b_j^{(1)} b_j^{(2)} - \sqrt{\lambda_1 \lambda_2}\, a_i^{(1)} a_i^{(2)} \right]$$

A un coefficient près dû au choix de la représentation simultanée on retrouve pour V_1 la variance pondérée des coordonnées des points colonnes.

Il reste alors à appliquer la formule 1 pour obtenir les ellipses de confiance autour de chaque point.

La figure 4 montre une illustration sur le célèbre exemple de FISHER (4) d'après des données de MAUNG croisant la couleur des yeux et des cheveux de 5 387 écoliers écossais.

Nos résultats diffèrent de (9) qui les a traités également, par l'usage de conventions d'échelles différentes et d'un réechantillonnage implicite à marges fixées au lieu de fixer seulement n.

En effet, l'usage des formules 1 et 2 séparement sur les lignes et les colonnes de N revient à les traiter indépendemment, ce qui réduit la variabilité d'ensemble. Ici, le problème de stabilité des axes ne se pose pas car λ_1 et λ_2 représentent respectivement 86,6 % et 13,1 % de l'inertie.

6 COMMENTAIRES ET CONCLUSION

Diverses difficultés subsistent non négligeables. Tout d'abord le problème déja signalé dans ce qui précède de la dépendance des axes factoriels vis à vis des variables actives. Ensuite la limitation au premier plan principal : lorsqu'il y a p > 2 axes significatifs il semble logique de travailler dans l'espace des p facteurs retenus ; il faudrait alors utiliser des ellipsoïdes à p dimensions. Pour les graphiques plans, faut-il alors considérer l'intersection de ces ellipsoïdes avec le plan principal ou sa projection ? La prise en compte d'une composante supplémentaire même non corrélée localement va en général augmenter la taille de l'ellipse par un effet de degré de liberté ; on risque alors d'aboutir à des zones de confiance démesurées.

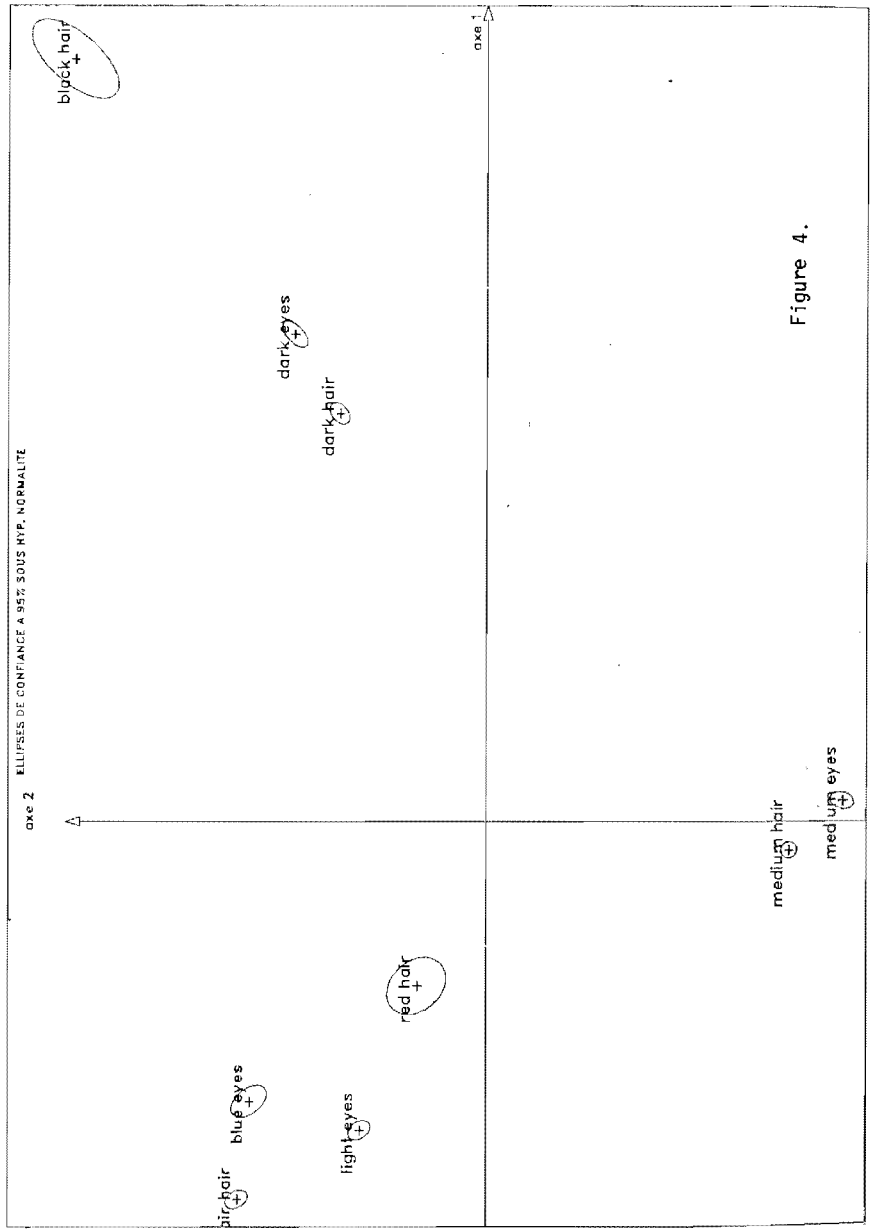

Figure 4.

Enfin, quelle interprétation donner à la mesure de l'intersection de 2 ellipses ou de 2 ellipsoïdes ?

L'utilisation d'ellipses de confiance ne vaut d'ailleurs que dans l'optique de l'estimation de barycentres de catégories. Si on cherche des zones de garde pour la dispersion des individus d'une catégorie, les ellipses n'ont aucune raison de donner les contours souhaités et d'autres techniques peuvent leur être préférées (voir (7) déjà cité).

Nous sommes néanmoins convaincus de l'utilité d'une telle démarche qui enrichit les aides à l'interprétation et qui devrait être appelée à un grand développement grace à celui des outils d'informatique graphique.

REFERENCES

(1) AESOP Les structures de l'opinion publique en 1984. Rapport interne.

(2) BENZECRI, J.P, L'analyse des données. Tome II Correspondances, (Dunod , Paris, 1973).

(3) EFRON, B, The jack-knife, the bootstrap and other resampling plans, SIAM monograph n°38 (1982)

(4) FISHER, R.A, The precision of discriminant functions. Ann. Eugenics, 10, (1940), 422-429.

(5) GABRIEL, K.R., Biplot display of multivariate matrices for inspection of data and diagnosis. in Barnett, V. (ed) Interpreting Multivariate Data (Wiley, Londres, 1981).

(6) GIFI, A. Nonlinear multivariate analysis (Department of Data Theory, Leiden, 1981).

(7) GREEN, P.J, Peeling bivariate data, in Barnett,V. (ed) op. cit.

(8) GREENACRE, M.J., Theory and applications of correspondence analysis (Academic Press, Londres, 1984).

(9) MEULMAN, J. Correspondence analysis and stability, Table Ronde Analyse des données, Toulouse, 1984.

(10)MORINEAU,A., Note sur la caractérisation statistique d'une classe et les valeurs tests. Bulletin Technique du CESIA 2 (1984) 20-27.

(11)RAMSAY,J.O., Confidence regions for multidimensional scaling analysis, Psychometrika, 43, (1978), 145-160.

(12)WEINBERG, S.L., CARROLL, J.D., COHEN, H.S., Confidence regions for INDSCAL using the jack-knife and bootstrap techniques, Psychometrika, 49 (1984), 475-491.

DATA ANALYSIS AND INFORMATICS, IV
E. Diday et al. (editors)

THE PERMUTATIONAL LIMIT DISTRIBUTION OF GENERALIZED CANONICAL CORRELATIONS

Jan de Leeuw
Eeke van der Burg
Department of Data Theory FSW/RUL
Middelstegracht 4
2312 TW Leiden, The Netherlands

ABSTRACT

We study the permutational limit distribution of goodness-of-fit statistics computed in various forms of generalized canonical correlation analysis. In simple cases of canonical analysis the exact distribution can be computed. For somewhat more complicated forms approximations have been tabulated by *Krishnaiah* and others. For the most general forms of canonical analysis we must use resampling methods to generate the permutation distribution. Our approach is illustrated by examples of varying degree of complexity. For small examples we can actually study the quality of our approximations. For more complicated examples this is not possible, but the approximate permutation distributions themselves are a very valuable data analytical tool.

INTRODUCTION

Canonical correlation analysis is a familiar data analysis technique. It is closely related to other well known techniques such as multiple regression, discriminant analysis, analysis of variance, principal component analysis, correspondence analysis, and so on. If we define canonical analysis broadly enough, it includes the other techniques as special cases. This fact is used in Gifi (1981) to build a very general system of multivariate analysis methods, which are all versions of a very general form of canonical analysis. A short, but fairly complete, introduction to the Gifi system is given by De Leeuw (1984a). A detailed discussion of the general form of canonical analysis used in this paper is contained in Van der Burg, et al. (1984). The various forms of canonical analysis are typical data analysis techniques, in the sense that they are used in exploratory situations, emphasize graphical representation, and are seldom used for inferential purposes. This is sometimes presented as a disadvantage of this class of techniques, because we have no information about 'generalizability' or 'significance' of the results. It is shown in Gifi (1981), compare also De Leeuw (1984b), that under random sampling assumptions it is possible to derive confidence interval information for large classes of canonical correlation techniques.

If the random sampling assumptions are not appropriate, which will very often be the case, we can use the resampling framework provided by the Bootstrap or the Jackknife (Efron, 1979, 1982, Efron and Gong, 1983). This provides us with 'nonstochastic confidence interval estimates', to paraphrase a term of Freedman and Lane (1983).
In the same way we can try to find significance tests for some interesting hypotheses in this class of techniques. There are some proposals valid under random sampling assumptions in De Leeuw (1984b). As is also pointed out there, these tests have a 'nonstochastic' interpretation in a randomization framework based on permutations. Freedman and Lane (1983) present randomization versions of chi-square and F-test in a very similar framework. Edgington (1980) studies permutation tests as general data analytical tools. In this paper we work out some of the suggestions in De Leeuw (1984b), and study some permutation tests for generalized canonical analysis.