DONNEES SUPPLEMENTAIRES SUR L'ANALYSE DES DONNEES

Les références anglaises de G. Saporta

Annette Leclerc ayant exprimé le souhait de rassembler les références en langue anglaise relatives à l'analyse des correspondances, voici quelques éléments bibliographiques, ne prétendant pas à l'exhaustivité, qui ont été rassemblés dans le cadre du contrat DGRST n° 75.7.0230 sur les variables qualitatives obtenu par la COREF.

Les équations de l'analyse des correspondances sont connues des anglo-saxons depuis plusieurs dizaines d'années, en particulier en tant qu'analyse canonique envisagée sous l'angle du codage des modalités de deux variables qualitatives. On sait en effet que "la recherche des facteurs est équivalente à celle de couples de fonctions sur les deux ensembles en correspondance qui soient le plus corrélées entre elles". (J.P. Benzecri, L'analyse des données, T.II, B n°7). Cependant les anglo-saxons n'ont pas pensé à utiliser cette technique de codage comme une méthode d'analyse des données ; (le titre d'un récent article de Hill (14) en témoigne éloquemment) ce qui, avec les représentations graphiques et les interprétations géométriques, constitue donc l'apport original de l'école française.

La plus ancienne référence connue remonte à 1935 où Hirschfeld (1) cherchant à coder simultanément deux variables qualitatives afin que pour chacune la régression de l'une sur l'autre soit linéaire, aboutit aux équations de l'analyse des correspondances, et obtint la relation exprimant le ϕ^2 comme somme des carrés des corrélations canoniques.

Par la suite, en 1940, Fisher (2), ignorant probablement (1), proposa la même technique de codage mais en vue cette fois-ci de trouver une fonction discriminante destinée à prédire une variable qualitative par une autre, problème qu'il relia à l'analyse canonique de Hotelling. Maung (3) appliqua l'idée de (2) à un cas concret et donna la formule de reconstitution des p_{ij} au moyen des $p_{i.}, p_{.j}$ et des facteurs, dans le cas d'un tableau 3x3 ; cette formule lui ayant été signalée par Fisher.

La même année Guttman (4) développa semble-t-il la même méthode mais nous n'avons pu consulter cette publication qui est apparemment introuvable.

En 1952 Williams poursuivant les travaux de Fisher et Maung proposa un traitement très clair du problème du codage (6) que Lancaster prolongea en examinant le cas d'une table de contingence issue d'une distribution binormale (7) puis en traitant le cas de variables quelconques (8) effectuant dès 1958 ce qu'on appellera plus tard une "analyse des correspondances continues". On trouve en particulier dans (8) la formule de reconstitution de la loi conjointe pour les variables continues qui généralise celle de Fisher-Maung.

On trouve ensuite une mention dans Torgerson (9) et une nouvelle publication de Guttman (10) qui procèdent de la même démarche. Kendall et Stuart (11) (cités par A. Leclerc) consacrèrent dans leur traité plusieurs pages aux travaux de Williams et Lancaster, lequel reprit plus tard la question dans (12).

A part l'article de Srikantan signalé par A. Leclerc, je n'ai relevé depuis que deux publications récentes sur le sujet : celle de Hill (14) citée plus haut et un opuscule de De Leeuw (13), ces deux auteurs connaissant les travaux de J.P. Benzecri.

Il faudrait aussi signaler quelques publications ayant trait à l'analyse simultanée de plusieurs variables qualitatives dans une optique proche de l'analyse des correspondances notamment Burt (5) et De Leeuw (13), cette dernière renfermant à ce propos une bibliographie très complète.

N.B. Depuis que nous avons écrit cette note, est parue une publication de J.P. Benzecri intitulée "Histoire et préhistoire de l'analyse des données" qui précise les rapports entre l'analyse des correspondances et les travaux des écoles anglo-saxonnes. En particulier le paragraphe 3.5.2., répondant à l'article de Hill, analyse les articles de Hirschfeld, Fisher et Maung cités ici.

Références :

(1) HIRSCHFELD H.O. (1935), A connection between correlation and contingency, Proc. Camb. Phil. Soc., 31, 520-524.

(2) FISHER R.A. (1940), The precision of discriminant functions, Ann. Eugen. Lond., 10, 422-429.

(3) MAUNG K. (1941), Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of scottish school children, Ann. Eugen. Lond. 11, 189-223.

(4) GUTTMAN L. (1941), The quantification of a class of attributes : a theory and method of scale construction, in "The prediction of personal adjustment" (Paul Horst ed.), Social Science Research Bull., n° 48, 251-364, New-York.

(5) BURT C. (1950), The factorial analysis of qualitative data, Brit. J. Stat. Psych. 3, 166-185.

(6) WILLIAMS E.J. (1952), Use of scores for the analysis of association in contingency tables, Biometrika, 39, 274-289.

(7) LANCASTER H.O. (1957) Some properties of the bivariate normal distribution considered in the form of a contingency table, Biometrika, 44, 289-292.

(8) LANCASTER H.O. (1958), The structure of bivariate distribution, Ann. Math. Stat., 29, 719-736

(9) TORGERSON W.S. (1958), THeory and method of scaling, ch. 12, 338-345, New-York, Wiley.

(10) GUTTMAN L. (1959), Metricizing rank ordered or unordered data for a linear factor analysis, Sankhya, 21, 257-268.

(11) KENDALL M.G., STUART A. (1961), The advanced theory of statistics, vol.2, ch. 33, 568-574, Londres, Griffin.

(12) LANCASTER H.O. (1969), The chi-square distribution, ch.6, 85-116, New-York, Wiley.

(13) JAN DE LEEUW (1973) Canonical analysis of categorical data, Université de Leyde.

(14) HILL M.O. (1974), Correspondence analysis : a negleoted multivariate method, Appl. Stat., 23, 340-354.