REVUE DE STATISTIQUE APPLIQUÉE

TOMÀS ALUJA
MANUEL MARTI

Discussion

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 83-85.
<http://www.numdam.org/item?id=RSA_1987__35_3_83_0>

DISCUSSION

Tomàs ALUJA et Manuel MARTI
Departement de Statistique et Recherche Opérationnelle
Universitat Politècnica de Catalunya
c/Pau Gargallo 5, 08028 Barcelone, Espagne

Les 5 articles précédents représentent de remarquables contributions à l'élaboration d'une stratégie complète d'analyse des données complexes, par emploi itératif des approximations exploratoire et modélisation.

Comme l'a remarqué L. LEBART (LEBART 1985), le statisticien se trouve confronté à des situations qui peuvent être décrites selon les trois caractéristiques suivantes : données structurées par opposition à données amorphes, données univariées par opposition à données multivariées et utilisation d'une approche exploratoire par opposition à une approche confirmatoire. Habituellement les situations ne seront pas pures, comme celles qui se trouveraient aux sommets d'un cube défini par les trois caractéristiques précédentes. Dans la pratique, les données ne sont pas complètement amorphes, car on possède une certaine connaissance sur les données (soit qu'on puisse les diviser en sous-groupes homogènes, soit qu'on puisse établir une liaison temporelle ou spatiale, soit qu'on connaisse l'existence de facteurs sous-jacents qu'on voudrait confirmer, etc.). D'autre part, dans la plupart des cas l'approche n'est pas totalement exploratoire ni totalement confirmatoire, elle conjugue des éléments des 2 types d'approches, et enfin, par contre, la réalité est presque toujours multivariée.

Dans ces situations, le problème du statisticien consiste à trouver la stratégie optimale d'utilisation des méthodes (COPPI 1986) de façon à mettre à profit l'information a priori, ou acquise en cours d'analyse, pour essayer d'en savoir plus. Sur ce point notre expérience nous montre que l'approche exploratoire et la modélisation sont deux phases d'une même analyse, et correspondent à des étapes différentes de la recherche; d'abord on veut voir quelles données on a, les représenter, détecter les anomalies, et suggérer des structures possibles permanentes. Après, on désire confirmer ces hypothèses, expliquer les « patterns » détectés. Cette stratégie implique l'utilisation complémentaire des deux approches, en particulier l'utilisation des techniques de représentation d'un corpus multidimensionnel et de la réduction de la dimensionnalité comme guide et simplification pour la phase de modélisation.

Cette démarche est appliquée dans les articles qui présentent des études de cas. WORSLEY montre sur un bon exemple comment l'analyse exploratoire peut guider la formulation du modèle; il faut remarquer, cependant, le fait que, malgré que le tableau concerne trois variables, il est analysé comme un tableau de

contingence simple. Dans ce cas sont connues les analogies entre les représentations graphiques et les termes d'association du modèle log-linéaire (LAURO 1982, GOODMAN 1986). Par contre, les représentations de l'ACM, peuvent induire des associations fausses, par exemple, lorsque il existe deux associations simples entre trois variables, les représentations sur les premiers plans factoriels montreraient une association fausse entre les variables non associées (ALUJA et MARTI 1986); cela découle du fait que l'ACM analyse les tableaux marginaux de l'hyper-cube de contingence, donc, il analyse les relations entre couples de variables indépendamment du reste des variables du tableau; cette remarque est aussi valable pour les articles de BACCINI, MATHIEU et MONDOT et AITKIN, FRANCIS et RAYNAL, et explique le fait que les résultats de l'ACM et du modèle log-linéaire souvent ne coïncident pas.

Les articles de FALGUEROLLES — HEIJDEN et CAUSSINUS — FALGUEROLLES proposent un pas en avant quant à la stratégie d'utilisation des méthodes exploratoires pour l'analyse des résidus d'un modèle. CAUSSINUS — FALGUEROLLES, présentent dans leur article deux cas de mise en œuvre de cette stratégie pour l'analyse de deux tableaux carrés pour lesquels l'ACP s'avère peu satisfaisante (ce type de tableaux de données est assez fréquent, comme est le cas des matrices de flux réciproques : mobilité, échanges économiques, etc.). La modélisation proposée permet d'identifier un effet ligne, un effet colonne, un effet de similarité mutuelle entre ligne et colonne, et le résidu, qui représente la partie asymétrique de la proximité entre ligne et colonne; de plus, cette stratégie permet la représentation de la structure de ces deux matrices (similarité mutuelle et résidus). Cette stratégie est une démarche puissante et flexible pour l'analyse de ce qu'il reste dans les résidus, après avoir éliminé les effets spécifiés dans le modèle; c'est donc une façon de prendre en compte la structure connue des données dans l'analyse et cela repose sur le même principe que l'analyse des corrélations partielles.

Dans le même esprit, nous avons développé les analyses, locales et partielles (ALUJA et LEBART 1985) lorsque il existe une relation entre les individus représentable à l'aide d'un graphe; ces analyses permettent aussi de dégager d'une façon plus précise les véritables associations existantes dans une table de contingence multiple. D'autre part, l'emboitement des démarches esquissées (méthodes exploratoires préalables à la formulation du modèle et méthodes exploratoires pour l'analyse des résidus d'un modèle), signifie l'inclusion de ces méthodes dans la méthodologie statistique.

Références complémentaires

T. ALUJA et M. MARTI (1986). — « Complementarity between log-linear models and correspondence analysis », *II Catalan Symposium on Statistics,* Barcelona, Vol 1 (Invited lectures), 113-140.

T. ALUJA ET L. LEBART (1985). — « Factorial analysis upon a graph », *DLP Bullet. Tech. du CESIA,* Vol. 3, 5-19.

COPPI (1986). — « Comparaison de méthodes et comparaison de stratégies d'analyse : quelques réflexions sur l'analyse d'un gros fichier de données sociologiques » *Publications du Lab. de Stat. et Prob. n° 02-86,* pp. 27-35. Université Paul Sabatier. Toulouse.

L. LEBART (1985). — « Quelques progrès récents dans la pratique de l'analyse des données », Invited lecture at the International Meeting of Statistics in the Basque Country, *IMSIBAC III.* Bilbao.

N.C. LAURO et A. DECARLI (1982). — « Correspondence analysis and log-linear models in multiway contingency tables study. Some remarks on experimental data » *Metron n° 1-2,* pp. 213-234.