J. J. Daudin

Discussion

Revue de statistique appliquée, tome 35, nº 3 (1987), p. 85-87
<http://www.numdam.org/item?id=RSA_1987__35_3_85_1>

J.J. DAUDIN

Département de Mathématique et Informatique
INAPG, 16 rue Claude Bernard 75005, Paris

Les travaux effectués par les équipes du Laboratoire de statistique et de probabilités de l'Université Paul Sabatier et le centre de statistique appliquée de l'Université de Lancaster forment une contribution importante au développement de la statistique sous plusieurs aspects : non seulement ils permettent de mieux saisir les différences et les complémentarités existant entre les approches françaises et britanniques, mais aussi ils ouvrent la voie à une démarche statistique originale qui associe et coordonne modèles et analyses factorielle et/ou typologique. Si l'idée d'utiliser conjointement des outils d'analyse des données et des modèles statistiques est présente depuis quelques années, comme par exemple l'analyse des interactions d'une analyse de variance à l'aide d'une classification automatique, ou encore la conception de l'analyse factorielle des correspondances comme une analyse des résidus par rapport au modèle d'indépendance, l'exploitation systématique et raisonnée de cette idée n'avait par contre jamais été développée aussi complètement.

On pourrait opposer les approches britannique et française par la notion de modèle statistique, la première en faisant un usage intensif alors que la seconde la rejetterait. Cette opposition est trop rapide et artificielle.

D'une part, on peut présenter beaucoup de méthodes d'analyse factorielle sous la forme d'un modèle statistique (CAUSSINUS (1985)), d'autre part la notion de modèle statistique recouvre des pratiques très différentes : on peut utiliser un modèle statistique pour confirmer ou réfuter une théorie concernant la discipline scientifique pour laquelle ont été récoltées les données, mais aussi pour prédire des variables non mesurées ou encore pour décrire des données sur lesquelles on ne possède pas de théorie globale cohérente.

Lorsqu'il s'agit de données complexes il est rare que l'on dispose d'une théorie globale qui puisse être formalisée sous la forme d'un modèle statistique. On se trouve donc généralement dans la situation où ce dernier est un moyen de décrire les données de façon judicieuse, c'est-à-dire en extrayant l'information significative du bruit dû aux erreurs de mesure ou à la variabilité d'échantillonnage. Mais c'est précisément ce même but qui est poursuivi, par d'autres voies, en analyse factorielle. Dans un cas comme dans l'autre on veut décomposer les données en 2 composantes sous la forme

$$\text{Données} = \text{« effets »} + \text{bruit blanc}$$

Les deux approches sont donc concurrentes, au sens où elles cherchent toutes deux à réaliser un même objectif par des voies différentes. La différence essentielle entre les deux approches est la forme mathématique que prend la première composante : axes factoriels dans l'approche française et espace paramétrique du type du modèle linéaire dans l'approche britannique. Ce point est développé dans (DAUDIN, TRÉCOURT, TOMASSONE (1984)).

La qualité d'une méthode statistique descriptive se mesure à son aptitude à décomposer correctement les deux composantes et à la qualité de la description de la première composante qui est généralement complexe. H. CAUSSINUS et A. De FALGUEROLLES montrent clairement que l'on peut utiliser les deux types de description *conjointement sur deux éléments séparés de la première composante.*

Les mêmes auteurs développent en introduction la conception de modèle statistique dans un contexte d'analyse exploratoire. Ce point est important sur le plan méthodologique car les termes de modèle et de résidu n'ont pas, dans ce contexte, le sens qui leur est accordé classiquement. En particulier le terme de résidu a ici une acception large : il contient à la fois le résidu au sens classique et des éléments de la première composante non pris en compte par le modèle. L'intérêt de la démarche est que si les aspects les plus forts et les plus simples de la première composante des données peuvent être bien pris en compte par un petit nombre de paramètres, les aspects plus diffus et plus compliqués sont mieux synthétisés par une analyse factorielle.

L'ensemble des travaux suscite un grand nombre de questions, mais je me limite à deux commentaires supplémentaires :

La présentation des résultats se fait plutôt sous la forme de chiffres (estimation des paramètres du modèle) dans l'approche anglaise alors que l'approche française utilise des représentations graphiques issues des axes des analyses factorielles. Dans les deux cas, il s'agit de décrire, résumer, synthétiser des données complexes, mais la forme donnée à la description est différente.

Cependant, il est souvent possible d'utiliser les estimations des paramètres pour construire des graphiques comme le font CAUSSINUS et De FALGUEROLLES (voir aussi AITKIN (1984) ou DAUDIN et TRÉCOURT (1980) où une utilisation entièrement descriptive du modèle loglinéaire est faite); il semble qu'une représentation graphique soit une bonne façon de transmettre les résultats d'une analyse de données complexes, dans la mesure où les paramètres sont alors très nombreux, et où les chiffres bruts n'étant pas réutilisés dans un modèle prédictif, il n'est pas nécessaire de retenir précisément la valeur de chacun d'eux, mais plutôt les ordres de grandeur et les relations d'ordre, informations qui apparaissent bien dans un graphique.

Ma deuxième remarque concerne l'utilisation du modèle loglinéaire; AITKIN, FRANCIS et RAYNAL n'ont pas pu l'utiliser sur leurs données « vues les grandes dimensions des tables de contingence et leur taux de cellules vides ». En réalité, certains logiciels mieux adaptés à ce modèle que GLIM permettent d'utiliser ce modèle sur des tables d'assez grandes dimensions dans la mesure où il n'est pas indispensable de conserver toute la table de contingence mais seulement certaines tables marginales. Pour la même raison les cellules vides ne créent pas trop de problèmes si on se limite à des modèles raisonnables. Par contre un des problèmes de ce modèle est l'exploitation des résultats que l'on peut en faire : dans ce type de situation on obtient un grand nombre de paramètres décrivant la première composante des données et il est difficile de synthétiser cette information. On rejoint ici le problème abordé précédemment.

Références

M. AITKIN (1984). — Modélisation mathématique de l'enquête communautaire sur les forces de travail. In Développements récents dans l'analyse de grands ensembles de données, *Information de l'Eurostat, numéro spécial,* Luxembourg.

H. CAUSSINUS (1985). — Quelques réflexions sur la part des modèles probabilistes en Analyse des données. *Quatrièmes journées Internationales Analyse des données et Informatique.* INRIA. Edition provisoire.

J.J. DAUDIN et P. TRÉCOURT (1980). — Analyse factorielle des correspondances et modèle loglinéaire, comparaison des deux méthodes sur un exemple. *Revue de Statistique Appliquée,* XXVIII, 1, 5-24.

J.J. DAUDIN, R. TOMASSONE et P. TRÉCOURT (1984). — Analyse d'enquêtes à grande échelle. In Développements récents dans l'analyse de grands ensembles de données, *Information de l'Eurostat, numéro spécial,* Luxembourg.

—