REVUE DE STATISTIQUE APPLIQUÉE

Y. ESCOUFIER

Discussion

Revue de statistique appliquée, tome 35, n° 3 (1987), p. 89-91

<http://www.numdam.org/item?id=RSA_1987__35_3_89_1>

Y. ESCOUFIER

Unité de Biométrie, 9, Place Pierre Viala, 34060 Montpellier Cedex

Disons d'entrée que l'entreprise que représente cette publication me paraît intéressante. Je défends trop souvent l'idée qu'il n'y a pas de recherche en statistique sans traitement effectif de données pour ne pas me réjouir des travaux qui me sont soumis. Bien sûr, ce type de préoccupation donne à certains textes une fome qui est plus celle d'un compte-rendu d'activités que celle d'un article scientifique traditionnel. Il me semble bon d'accepter au moins de temps en temps une telle présentation de nos connaissances. Elle aide ceux qui ne peuvent accéder aux méthodes par la compréhension mathématique de leurs fondements à y accéder par la reproduction des démarches décrites. N'est-ce pas là une approche pédagogique reconnue ?

Pour avoir une vue générale des textes proposés à ma lecture, je prendrai pour observatoire une situation que je connais bien, l'Analyse des Correspondances. Parmi de nombreuses présentations possibles de cette méthode je choisis celle qui la voit comme une approximation au sens des moindres carrés des écarts des fréquences observées aux fréquences attendues sous l'hypothèse d'indépendance. Il y a donc trois choix. On parle de fréquences et non de logarithmes ou de racines carrées des fréquences; on se situe par rapport aux fréquences attendues sous l'hypothèse d'indépendance et non par rapport à un modèle quelconque; on utilise le critère des moindres carrés alors que d'autres parlent de déviance, de moindres valeurs absolues ou de critère minimax. Tous ces choix sont discutables et des choix alternatifs ont déjà été expérimentés. Les travaux de VOLLE ou ESCOFIER cités dans les textes en sont des exemples. Le travail de C. LAURO sur l'Analyse non symétrique des Correspondances relève de la même approche. Le modèle Log-linéaire est un exemple de choix différents : dans un contexte probabiliste, on choisit de travailler avec les logarithmes des fréquences pour le critère du maximum de vraisemblance.

Deux points me paraissent importants dans cette remise en cause des choix initiaux. Le premier concerne la cohérence qui parait nécessaire entre les éléments qui concourent à construire la démarche. Prenons un exemple : Ajuster des P_{ij} par des

$$\hat{P}_{ij} = \alpha_i \beta_j \exp\left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j}\right)$$

au sens du maximum de vraisemblance puis étudier les écarts $P_{ij} - \hat{P}_{ij}$ par une décomposition en valeurs singulières c'est-à-dire au sens des moindres carrés me paraît curieux. Pourquoi ne pas conserver les paramètres trouvés et ajuster des

$$\hat{\hat{P}}_{ij} = \alpha_i \beta_j \exp\left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j} + \sum_{\alpha=k+1}^{v} \sqrt{\lambda_\alpha}\, \Psi_{\alpha i}\, \phi_{\alpha j}\right)$$

au sens du maximum de vraisemblance ?

Le second point important est, me semble-t-il, de ne pas restreindre les approches concevables à celles rendues possibles aujourd'hui par des programmes disponibles. Prenons ici aussi un exemple directement tiré de l'article de K.J. WORSLEY : par ses formules (1) et (2) l'auteur rappelle des liens qui existent entre les solutions de l'Analyse des Correspondances et celles du modèle log-multiplicatif. Analysant les résultats de l'Analyse des Correspondances, il en déduit que des paramètres qu'il note u_{1i} et u_{2i} pourraient dans un but de simplicité être contraints à ne prendre que certaines valeurs. Il introduit alors ces contraintes dans le modèle log-linéaire, c'est-à-dire dans une approximation fondée sur le critère du maximum de vraisemblance. Pourquoi ne pas le faire directement dans l'approximation au sens des moindres carrés ? Je crains que la réponse soit simplement l'absence actuelle de programme. C'est une lacune des méthodes factorielles telles qu'elles sont largement pratiquées. Nous nous sommes endormis dans le confort des solutions en vecteurs propres et valeurs propres des problèmes aux moindres carrés sans contraintes.

Je terminerai par une remarque de vocabulaire concernant l'article de M. AITKIN *et al.* Pour moi, l'algorithme décrit en 4.2. est un algorithme de nuées dynamiques; ce que les auteurs appellent Etape E et Etape M correspond aux fonctions d'affectation et aux fonctions de représentations du livre de E. DIDAY

et al. Ce que les auteurs appellent Nuées Dynamiques n'est que le cas très particulier dit des moyennes mobiles. Ceci dit je m'interroge sur la signification des résultats de la dernière table du paragraphe 5. Comment un même profil peut-il donner des individus dans des classes différentes ?

Références

N. LAURO, L. D'AMBRA (1983). — L'analyse non symétrique des correspondances. 3e Journées Internationales d'Analyse de Données et Informatique, Versailles. Paru dans *Data Analysis and Informatics 3,* North-Holland, Amsterdam, 433-446.