MATHÉMATIQUES ET SCIENCES HUMAINES

SIMON RÉGNIER

Non-fécondité du modèle statistique général de la classification automatique

Mathématiques et sciences humaines, tome 82 (1983), p. 67-74

<http://www.numdam.org/item?id=MSH_1983__82__67_0>

Math. Sci. hum. (21^{e} année, n°82, *1983*, p.67-74)

NON-FÉCONDITÉ DU MODÈLE STATISTIQUE GÉNÉRAL DE LA CLASSIFICATION AUTOMATIQUE *

Simon RÉGNIER
Centre de Calcul de la Maison des Sciences de l'Homme, Paris.

RÉSUMÉ

Dans une perspective analogue à celle, classique, du modèle linéaire général commun aux analyses de variance, de covariance, de régression, l'auteur analyse un modèle statistique qui lui semble assez général pour englober toutes les problématiques classificatoires. Ce modèle comprend tout naturellement comme cas particulier le modèle le plus général de l'analyse discrimante, qui correspond au cas de figure « agréable » où le contenu des classes que l'on cherche à remplir est quelque peu connu *a priori*.

A ce niveau de généralité, la seule méthode de traitement disponible est celle du maximum de vraisemblance. On montre alors que ce modèle conduit à des algorithmes très lourds mais pertinents dans des situations dites « paramétriques », où chaque classe à remplir est caractérisée par une loi de probabilité inconnue dépendant de quelques paramètres réels, dans une famille de lois *a priori* connue.

Mais dans la situation non-paramétrique (à notre avis la plus courante, spécialement quand le champ des observations possible est fini) celle où les classes à construire sont *a priori* totalement indéterminées, on montre que le même modèle traité par le même principe du maximum de vraisemblance conduit à une classe de classifications « les plus vraisemblables » parfaitement dépourvues d'intérêt physique, parce qu'en dehors de certain cas très particulier où l'on obtient des classifications très pertinentes mais *a priori* évidentes, on obtient en général une classe de classifications globalement invariante par permutation des objets. Bref, le modèle ainsi traité conduit à partitionner le cardinal de l'ensemble d'objets, et non cet ensemble lui-même.

ABSTRACT

In a perspective similar to the classical approach through the general linear model common to variance, covariance, and regression analysis, the A. studies a statistical model that seems general enough to encompass all classification problems. This model naturally includes as a special case the more general model of discriminant analysis, which corresponds to the "pleasant" case where something is known *a priori* about the content of the classes that are being looked for

At this level of abstraction, the only available method is that of maximum likelihood. The A. shows that this model leads to algorithms that are unwieldy, but operative in the so-called " parametric " situations, in which each of the classes to be found is characterized by an unknown law of probability — depending upon a few real parameters — among a set of laws known *a priori*.

But in the non parametric case (to us, the most common, especially when the range of possible observations is finite), viz. the case when the classes that are looked for are *a priori* totally unknown, the same model, handled according to the same principle of maximum likelihood, leads to a class of "most likely" classifications that are devoided of any physical interest. The reason is that except for some very special cases in which highly relevant, but *a priori* obvious classes are obtained, the result is generally a class of classifications that remain wholly unvariant when objects are permuted. In brief, the model, when handled in this way, leads to partition the cardinal of the set of objects, instead of the set itself.

---

* Extrait des Actes du Colloque "Archéologie et Calculateurs. Problèmes sémiologiques et mathématiques", Marseille, 7-12 Avril 1969, Paris, Editions du CNRS, 1970.

1. — Introduction

Le lecteur de l'ouvrage de M. LERMAN (voir Bibliographie) ressent probablement, comme l'auteur lui-même, une sorte de malaise : l'utilité physique, l'adéquation des outils proposés aux finalités des diverses sciences humaines susceptibles d'en faire usage est mal garantie, et l'on manque d'un point de vue général à ce niveau.

Il est temps d'essayer de combler cette lacune. Nous allons formuler un modèle statistique général pour la classification automatique : modèle comparable au modèle linéaire classique qui sert de cadre général commun aux analyses de régressions, de variance, ou de covariance; modèle très proche aussi du modèle général de l'analyse discriminante décrit à propos d'une étude classificatoire sur des gouaches peintes par des malades mentaux (Régnier, 1966).

L'analyse discriminante effectue en effet la classification d'un échantillon fini, dans des populations statistiques prédéfinies chacunes par sa loi de répartition, ou au moins par un échantillon de cette loi. C'est dire que les classes finales sont relativement connues d'avance dans l'abstrait.

Le problème général de la classification est si l'on veut du même type, avec des classes non définies a priori. Tout au plus peut-on prévoir, en général, la qualité d'information visée au niveau de l'ensemble des classes (de l'ensemble quotient E/X en termes de théorie classique des ensembles). Ce sera une information d'ordre temporel (cf. La classification des tombes étrusques dans De la Genière et de De la Vega, 1969) — et l'ensemble E/X devra être ordonné — d'ordre psychopathologique (cf. La clasification des gouaches peintes par des malades mentaux. Régnier, 1966), d'ordre économique et géographique, etc. Et cette information concernant la structure et la signification de l'ensemble des classes E/X, on s'est souvent interdit de l'utiliser à priori, pour garder un moyen de contrôle de la validité, de la pertinence des classifications obtenues.

Par exemple, dans une étude relative à l'évolution des prix du blé en France au XIX^{e} siècle (S. Régnier, inédit), on savait a priori que les classes devaient être formées de départements contigus. Mais cette notion géographique n'est pas intervenue dans l'algorithme de calcul, et le fait d'obtenir des classes toutes connexes sauf une a été un important argument quant à la validité de la procédure employée.

2. — Modèle statistique

Les modèle que nous proposons maintenant va obéir au même principe.

Les n objets O_i qu'il s'agit de classer sont considérés comme n observations indépendantes, extraites de une ou plusieurs populations statistiques sur un même espace $\mathfrak{X}$, le nombre g de ces populations peut toujours être supposé borné :

$$g \leqslant k .$$

Cette contrainte n'en est pas une lorsque $k = n$. Les k populations possibles ont des répartitions inconnues P_j sur $\mathfrak{X}$. On peut tout au plus savoir a priori que chaque P_j appartient à une certaine classe Ω_j de lois de probabilité sur $\mathfrak{X}$.

Les paramètres inconnus dans ce modèles sont alors :

— d'une part, ceux des lois

$$P_j \in \Omega_j$$

— d'autre part, la fonction d'affectation :

$$a : i \rightarrow j = a(i)$$

qui indique de quelle population P_j l'objet i est tiré.

Si les lois P_j son discrètes, la vraisemblance prendra la forme :

$$V = \prod_{i=1}^{n} P_{j=a(i)}(O_i) = \prod_{j=1}^{k} \prod_{i \in a^{-1}(j)} P_j(O_i)$$

Si les lois P_j sont absolument continues par rapport à une même mesure, il faut remplacer le symbole P par les expressions des densités de probabilités aux points O_i.

3. — Traitement au maximum de vraisemblance

Une estimation des paramètres P_j et a est maintenant en principe possible par la méthode du maximum de vraisemblance.

Remarquons que si les classes Ω_j étaient réduites chacune à une seule loi, nous serions ramenés au problème de l'analyse discriminante et que cette méthode affecterait chaque objet O_i à la population où sa vraisemblance $P_j(O_j)$ est la plus grande. Telle est bien, à des détails près, la procédure habituellement employée en analyse discriminante.

Inversement, pour une fonction a fixée, à valeurs dans $1, 2 \ldots k$ (il y a k^n telles fonctions), les n observations sont partagées en k sous-échantillons. Chaque sous-échantillon non vide permet d'estimer la population P_j correspondante au maximum de vraisemblance, et l'on peut contrôler l'hypothèse que ces estimations soient significativement différentes en formant le rapport l de la vraisemblance obtenue : $V(a)$, à celle obtenue à partir d'une fonction $a(i) =$ constante, qui affecte tous les objets O_i à la même population Po, Po appartenant à l'une quelconque des classes Ω_j :

$$P_o \in \bigcup_j \Omega_j$$

Ce rapport $l = V(a)/Vo$ va suivre en général, pour a fixé, une loi connue (la distribution asymptotique de deux logarithmes de l étant une loi du chi-carré) dans l'hypothèse où les k populations P_j sont en fait confondues. Et cette distribution fournit un test de cette hypothèse nulle, face à l'hypothèse alternative où au moins deux populations sont distinctes.

Pour estimer la fonction d'affectation a selon le même principe, il faut rechercher quel choix de a rend maxima la vraisemblance $V(a)$ déjà maximisée quant à l'estimation des populations P_j. On obtiendra par suite un rapport l le plus grand possible. Et l'on ne sait alors plus rien dire quant à la distribution de log (l) dans l'hypothèse nulle.

Dans le cas particulier où les classes Ω_j se réduisent chacunes à une seule loi, on doit remarquer que la procédure ci-dessus se ramène à celle de l'analyse discriminante. La première phase d'estimation des populations P_j est simplement supprimée.

***

Nous allons voir maintenant que cette procédure générale d'estimation au maximum de vraisemblance peut conduire à des résultats intéressants dans les modèles restrictifs, par exemple lorsque les classes Ω_j sont des familles de lois de probabilité dépendant de quelques paramètres réels, mais que dans le modèle non restrictif, où les lois P_j sont à priori quelconques, on est conduit à une classification stéréotypée et dépourvue d'intérêt.

4. — L'exemple des populations normales dans R^p

Considérons donc en premier lieu le cas où toutes les k classes de lois Ω_j sont identiques à l'ensemble de toutes les lois normales sur l'espace euclidien à p dimensions :

$$\mathfrak{X} = R^p$$

Chaque loi P_j dépend de $p(p+3)/2$ paramètres réels, à savoir les p coordonnées du vecteur espérance mathématique, et les $p(p+1)/2$ termes de la matrice symétrique des variances-covariances.

Pour a fixé, l'estimation des

$$k\,p\,\frac{p+3}{2}$$

paramètres est, dans son principe très simple. Chaque sous-échantillon de n_j observations définit une loi empirique dont on calculera le vecteur moyen et la matrice de variance. Ce sont les estimations au maximum de vraisemblance des paramètres correspondants dans la population P_j.

Le rapport de vraisemblance

$$l = \frac{V(a)}{V_o}$$

suit une loi connue; pour préciser :

2 log l suit la loi du chi-carré à

$$(k-1)\quad p\left(\frac{p+3}{2}\right)$$

degrés de liberté, dans l'hypothèse où les k populations normales sont en fait identiques (voir par exemple, Kendall, Vol. III, p. 266; The advanced theory of Statistics, Charles Griffin, 1958).

Mais en optimisant a, on obtient un maximum général $\hat{l}$ du rapport l, qui suit une loi très différente, et « plus à droite » : $\hat{l}$ a plus de chance d'être grand, toujours sous la même hypothèse nulle. C'est cette distribution, d'une étude difficile, qui permettrait seule de conclure si la classification obtenue finalement est significative, si les k populations estimées sont deux à deux significativement différentes.

5. — Le modèle non-restrictif dans les problèmes usuels

Les lois P_j sont totalement inconnues a priori.

Pour a fixée, chaque sous-échantillon de n_j objets définit une loi empirique $\hat{P}_j$ qui constitue l'estimation au maximum de vraisemblance de la loi P_j correspondante. Cette vraisemblance atteint alors la valeur

$$\prod_{j=1}^{k}\left(\frac{1}{n_j}\right)^{n_j}$$

si les n_j objets sont deux à deux distincts; et plus généralement, si n_j^t objets sont confondus aux points $t = 1, 2 \ldots s$

$$\begin{cases} \hat{P}_j(t) = \dfrac{n_j^t}{n_j} \\ V(\hat{P}_j) = \displaystyle\prod_{1}^{s}\left(\frac{n_j^t}{n_j}\right)^{n_j^t} \end{cases}$$

La vraisemblance globale est alors,

$$V(a) = \prod_j V(P_j) = \prod_{j=1}^{k} \prod_{t=1}^{s} \left(\frac{n_j^t}{n_j}\right)^{n_j^t}$$

En admettant que les n objets se répartissent en paquets de r_t aux points $t = 1, 2 \ldots s$, on a :

$$\left.\begin{array}{l} \sum_t n_j^t = n_j \\ \sum_j n_j^t = r_t \end{array}\right\} \quad \text{et} \sum_j n_j = \sum_t r_t = n$$

et l'on peut également écrire

$$\log V(a) = \sum_{j=1}^{k} \sum_{t=1}^{s} n_j^t \log \frac{n_j^t}{n_j}$$

6. — **Maximisation**

Introduisons les lois empiriques : $\hat{P}_j$

$$\hat{P}_j(t) = \frac{n_j^t}{n_j} \quad \text{soit} \quad a_j^t$$

et leurs entropies

$$E_j = - \sum_{t=1}^{s} a_j^t \log a_j^t$$

alors

$$\log V(a) = \sum_j n_j \sum_t a_j^t \log a_j^t$$

et

$$\log V(a) = - \sum n_j E_j$$

— *log V(a) est donc la moyenne des entropies E_j pondérées par les effectifs n_j.* Chacune d'elles est minima lorsque la loi a_j est la plus concentrée possible. $E_j = 0$ si et seulement si la loi P_j est concentrée en un seul point t_j. Pour que toutes les entropies E_j puissent être nulles il faut donc que l'étendue s occupée par les n objets soit au plus égale au nombre k des populations hypothétiques.

On obtiendra alors une vraisemblance maxima $V(a) = 1$, en utilisant une fonction a qui concentre chaque échantillon n_j en un seul point t_j. Il y aura au moins s échantillons non vides, au plus k, et t contraintes :

$$r_t = \sum_{\substack{j \text{ tel que} \\ t_j = t}} n_j$$

Au plus k — s points t porteront plus d'un échantillon.

On notera que parmi toutes ces solutions mathématiquement équivalentes, la seule physique-

ment satisfaisante est celle qui répartit les n objets en $g = s$ populations, chacune concentrée en un seul point t, et laisse indéterminé les $k - s$ populations restantes du modèle.

En général s dépasse k. L'optimisation de $V(a)$ est alors très difficile. On peut noter cependant que, comme cette vraisemblance ne dépend de s que par l'intermédiaire des effectifs :

$$n_j^t = \begin{cases} \text{nombre d'objets affectés à } P_j \text{ parmi} \\ \text{les } r_t \text{ objets confondus au point } t, \end{cases}$$

La recherche du maximum détermine tout au plus ces effectifs, et laisse la fonction a proprement dite largement indéterminée.

Dans le cas particulier le plus fréquent, où il n'y a pas d'objets confondus :

$$r_t = 1 \text{ ou } 0$$

Alors $\quad n_j^t = 0$ ou 1 et comme $0 \log 0 = 0$

$$\sum_1^s n_j^t \log \frac{n_j^t}{n_j} = \sum_1^{n_j} \log \frac{1}{n_j} = n_j \log \frac{1}{n_j}$$

$$V(P_j) = \left(\frac{1}{n_j}\right)^{n_j} \quad \text{comme déjà vu et}$$

$$V(a) = \prod_{j=1}^{k} \left(\frac{1}{n_j}\right)^{n_j}$$

$$\text{Log } V(a) = -\sum n_j \log n_j$$

Cette quantité est maxima en même temps que l'entropie :

$$-\sum_1^k \frac{n_j}{n} \log \frac{n_j}{n}$$

quand la distribution des n_j est la plus uniforme possible. Si $n = k\,q + r$ avec $r < k$, on prendra pour cela $k - r$ effectifs n_j égaux au quotient q et les r restant égaux à $q + 1$.

Naturellement. cette solution détermine seulement le système d'effectifs n_j et laisse à cela près la fonction d'affectation indéterminée. Cette procédure n'a donc aucune signification physique appropriée aux divers problèmes qui peuvent relever de ce modèle trop général. Elle aboutit à partitionner l'effectif n et non pas l'ensemble d'objets proprement dit.

BIBLIOGRAPHIE

De la Genière (M^me J.) et De la Vega (W. F.), 1968. — Analyse quantitative du mobilier funéraire de la fouille de Sala Consilina. *Calcul et Formalisation dans les Sciences de l'Homme*, C.N.R.S., Paris.

Kendall (M. G.), 1958 et seq. — *The Advanced Theory of Statistics*, 3 vol., Charles Griffin & Co., London.

LERMAN (I. C.), 1970. — *Les bases de la classification automatique.* Gauthier-Villars, Collection « Programmation », Paris.

RÉGNIER (S.), 1966. — *Classification et analyse des expressions plastiques non figuratives de malades mentaux.* Actes du Colloque International sur l'Informatique, Toulouse.

Note

Dans la présentation orale du texte qui précède, l'auteur a trouvé nécessaire d'introduire deux observations complémentaires :

1°) La démarche classificatoire envisagée ici ne correspond qu'à une partie de la problématique typologique envisagée dans la plupart des autres exposés. On peut distinguer, semble-t-il, deux types de visée : l'une, plus utilitaire, cherche à structurer un ensemble d'objets selon une ou plusieurs partitions emboîtées dans le but de rendre l'information assez considérable mise en jeu plus maniable. L'autre cherche à découvrir des classes ayant une certaine objectivité physique. C'est dans cette perspective que l'on s'est placé.

2°) Le modèle formulé ci-dessus n'a pas toute la généralité souhaitable. Si l'on se réfère aux situations exemplaires indiquées dans l'introduction, on voit que l'information a priori concernant les lois de probabilité P_j caractéristiques de chaque population peut être une relation entre toutes ces lois, par exemple une relation d'ordre, et non pas le k-uplet des relations d'appartenance indépendantes :

$$P_j \in \Omega_j$$

Pour avoir un modèle suffisamment général, il faut donc écrire que le k-uplet des k lois P_j appartient à une certaine classe Ω.

En notant LP ($\mathfrak{X}$) l'ensemble des lois de probabilité sur $\mathfrak{X}$, Ω sera une partie de [LP ($\mathfrak{X}$)] puissance k :

$$\Omega \subset [\mathrm{LP}\,(\mathfrak{X})]^k$$

et le modèle indiqué au paragraphe 2 correspond au cas de figure assez particulier où ce graphe Ω serait lui aussi un produit de k ensembles facteurs $\Omega_j \in \mathrm{LP}\,(\mathfrak{X})$:

$$\Omega = \prod_1^k \Omega_j$$

DISCUSSION

M^me^ K. SPARK-JONES. — Combien d'objets ont été classés dans l'expérience du prix du blé ? Est-ce la seule expérience réalisée ?

M. S. RÉGNIER. — Il y avait 89 objets, lesquels ont permis de tester un certain programme; le traitement dure 3 minutes sur l'IBM 704. Le même programme a été utilisé pour une expérience réalisée sur 200 malades mentaux. Ce programme a été écrit pour classer 500 objets, avec une restriction liée à l'encombrement-mémoire de la machine (Nombre max. de classes × Nombre max. d'objets < 10 000).

M. M. BORILLO. — Ce modèle est-il applicable quand la définition de l'objet est liée à un référant qui varie, c'est-à-dire par exemple lorsqu'il faut introduire en plus des critères « présence » et « absence » le critère « impossibilité d'affecter la présence ou l'absence » ?

M. S. RÉGNIER. — J'aborde le problème de la Classification en aval du problème « Comment représente-t-on les objets ? ». Mais ceci ne présente pas de difficultés. Si les « Items » prennent 3 valeurs — par exemple : présence, absence et non pertinence — x ne prendra plus la valeur 2^{52} (il y a 52 Items) mais 3^{52} où 3 représente un ensemble à 3 éléments qui prennent les valeurs 0,1 et 1/2 respectivement pour l'absence, la présence et la non pertinence.

M. I. C. LERMAN. — Je ne suis pas très assuré de la relative fécondité dans le cas que vous avez signalé. Prenons un exemple particulier dans le cadre de ce modèle. Supposons qu'on ait un ensemble de points dans un plan qui se répartissent entre deux modes de probabilités normales. Si on connaît les paramètres de ces deux lois de probabilités, le problème me paraît être un problème d'analyse discriminante. Si on ne les connaît pas, il faut les estimer. Comment décide-t-on de les estimer ? Peut-on faire une discrimination *a priori* ?

M. S. RÉGNIER. — Cet exemple a l'avantage de bien préciser la situation. On a un grand nuage de points dont certains sont issus d'une première loi normale et les autres d'une deuxième loi normale. On serait dans une situation discriminante si on connaissait *a priori* les paramètres des deux lois de probabilité. Les situations de classification proprement dites sont celles où l'on ne connaît rien *a priori*. Dans ces conditions, comment opérer ? Je vais envisager tous les partages des 200 points en deux effectifs *a priori* indéterminés — cela peut être 1 et 199, 100 et 100, ... —. Pour un partage donné j'ai un moyen d'estimer les deux populations; j'évalue ensuite si ces deux populations sont significativement différentes. Le partage retenu sera celui qui correspond aux deux populations les plus significativement différentes.

M. M. DAVY de VIRVILLE. — Il me semble que la non fécondité du modèle provient de l'absence de structure sur l'espace $\mathfrak{X}$. Quand on a deux objets différents, la seule chose que l'on puisse affirmer c'est qu'ils sont différents; on ne peut pas préciser cette différence.

M. S. RÉGNIER. — En fait il y a une structure sur $\mathfrak{X}$; mais elle n'intervient pas au niveau de w. Dans ces conditions, vous avez raison.

M. R. E. TOMASSONE. — Je voudrais faire une remarque d'ordre pratique concernant toutes les partitions possibles en deux groupes que vous envisagez de faire pour discriminer les deux populations les plus significativement différentes. Le nombre d'opérations intervenant dans le calcul serait tellement grand que le traitement serait irréalisable même sur les ordinateurs les plus puissants.

M. S. RÉGNIER. — Je partage votre opinion.

En effet le nombre de partition de 200 objets en deux groupes est environ de 2^{200}, soit 1024^{20}; les ordinateurs les plus rapides seraient donc incapables de réaliser un si grand nombre d'opérations. Néanmoins certaines méthodes de calcul, comme par exemple la méthode du simplexe en programmation linéaire, permettent d'obtenir en un temps machine raisonnable l'optimum absolu même si le polyèdre a un très grand nombre de sommets, car ils ne sont pas tous explorés.

Pour revenir au problème des classifications, on procède de la même manière : on part d'une partition donnée et on chemine en transférant un objet d'une classe à l'autre en cherchant toujours le meilleur progrès possible. Cependant dans ce cas, on n'est pas sûr d'obtenir l'optimum absolu.