LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

L'approximation stochastique en analyse des correspondances

Les cahiers de l'analyse des données, tome 7, n° 4 (1982), p. 387-394

<http://www.numdam.org/item?id=CAD_1982__7_4_387_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - *n°* 4 - *p.* 387-394

L'APPROXIMATION STOCHASTIQUE EN ANALYSE DES CORRESPONDANCES [APPR. CORR.]

par J. P. Benzécri (1)

1 *Histoire et principe de la méthode*

La présente leçon, dont la première rédaction remonte à 1967, propose un algorithme d'approximation stochastique qui, parce qu'il requiert une mémoire de calcul peu étendue, doit permettre d'analyser des tableaux de très grande dimension sur une machine dont la capacité de mémoire centrale n'excède pas 32 K.

Un exposé théorique a paru en 1969 (cf. J.-P. B. : Approximation Stochastique dans une algèbre normée non-commutative ; in *Bull. Soc. Math. France* ; T 97 ; pp 225-241 ; 1969). Les premiers résultats d'application pratique sont dans la thèse de J.-P. Fénelon (3e cycle, Paris ; 1973). En 1974, la méthode a été présentée par L. Lebart aux statisticiens du Comp. Stat. réunis à Vienne (cf. L. L. : On Benzécri's method for finding eigenvector by stochastic approximation (The case of binary data) ; *Proc. in Comput. Stat.* Physica Verlag ; Vienne ; pp 202-211 ; 1974). Depuis L. Lebart a écrit un programme, qu'il utilise couramment pour l'analyse des grandes enquêtes par questionnaire, comme il est expliqué au chapitre V du livre de L. Lebart, A. Morineau et N. Tabard : *Techniques de la description statistique* ; Dunod, 1977.

Quant au principe mathématique de la méthode, la leçon qui suit donne sous forme intuitive un aperçu suffisant de ce qui est démontré dans l'article de 1969. Mais pour comprendre le lien entre cette première conception et les applications des développements dus à L. Lebart, il nous paraît utile de remonter au principe philosophique.

Initialement, nous supposons que les individus i décrits par les lignes d'un tableau I × J, se présentent successivement, comme issus d'une source aléatoire, suivant une loi de probabilité déterminée (cf. § 1.1) : cette source aléatoire n'est que la formalisation mathématique de l'ensemble potentiellement infini I associé selon nous à une notion telle que celle d'espèce animale.

L'ensemble I étant potentiellement infini, l'analyse factorielle du tableau I × J, ne peut jamais s'achever : mais pratiquement on peut supposer (et démontrer, moyennant des hypothèses raisonnables sur la source aléatoire) que si l'on note If un ensemble fini de n individus, les résultats de l'analyse du tableau If × J (c'est-à-dire les facteurs sur J, ou les axes dans l'espace R_J) ne subissent plus, quand n tend vers l'infini, que des fluctuations évanescentes.

---

(1) Professeur de statistique. Université Pierre et Marie Curie.

Cependant, il n'est pas raisonnable d'effectuer une analyse factorielle complète chaque fois qu'on reçoit les données relatives à un nouvel individu. Il conviendrait plutôt de n'utiliser celui-ci que pour une simple retouche des facteurs ; à condition, cela s'entend, que ces retouches successives n'aient pas pour conséquence d'augmenter indûment l'amplitude des fluctuations des résultats. C'est ce calcul par retouches successives que réalise l'algorithme d'approximation stochastique; offrant une analogie frappante avec le jeu naturel de notre esprit ; lequel, par chaque fait nouveau, corrige la vision synthétique qu'il s'était formée d'un domaine. Ainsi l'approximation stochastique apparaît selon le mot de Henri Bergson (cf. [*BERGSON*] ; in C.A.D. : Vol VII n° 4 ; § 2 *in fine*) comme un "processus tout dynamique, assez analogue à la représentation purement qualitative qu'une enclume sensible aurait du nombre croissant des coups de marteau".

Dans sa conception originelle qui est celle de la présente leçon, la méthode participe donc de trois ordres de pensée différents.

(a) Elle est l'analogie mathématique précise d'une synthèse psychologique se déroulant dans le temps.

(b) Elle étudie la convergence des facteurs calculés sur un échantillon d'effectif croissant issu d'une source aléatoire ; (les conditions de convergence données dans l'article de 1969 offrant d' ailleurs matière à des spéculations mathématiques ultérieures).

(c) Elle offre au programmeur un algorithme de base très simple, et requérant une place de mémoire minima ; donc tout désigné pour l'analyse de grands tableaux.

C'est de ce dernier point de vue que L. Lebart en a repris l'étude. Or dans la pratique, même si une source potentielle infinie est à l'origine de la plupart des analyses (ensemble des malades traités pour telle affection ; des sujets répondant à tel questionnaire etc.),un ensemble important de données actuellement disponibles est la matière de toute étude concrète (les données nouvelles parvenues en cours d'étude, étant simplement projetées sur les axes comme des éléments supplémentaires). L'ensemble I n'est pas ouvert, mais clos. Et dès lors qu'on ne pense plus à (a) ni (b), l'algorithme lui-même peut être modifié du point de vue de (c).

Ainsi L. Lebart remarque que si (cf. § 1.2) au lieu de multiplier les vecteurs axiaux factoriels par $1 + (d(i)/i)$ ($d(i)$ étant en bref la matrice d'inertie de rang 1 du i-ème individu) on fait la somme des produits par $d(i)$ d'un état des vecteurs axiaux (initialement choisis, ou issus d'une approximation antérieure), on se trouve tout simplement avoir multiplié les vecteurs axiaux initiaux par la matrice d'inertie globale du nuage N(I) : et cela sans avoir calculé explicitement cette matrice, donc sans encombrer la mémoire de l'ordinateur plus que ne le fait l'algorithme de base présenté dans la présente leçon. L. Lebart a donc tiré de l'algorithme d'approximation stochastique une nouvelle forme de calcul de l'algorithme de la puissance itérée (appliqué croyons-nous depuis Hotelling en analyse factorielle). Il importe de souligner que pour nous l'introduction du diviseur i dans $1 + (d(i)/i)$ est essentielle, parce qu'elle assure que toute l'amplitude des retouches tend vers zéro, quand i tend vers l'infini ; question qui ne se pose pas pour un ensemnle I clos.

L. Lebart n'abandonne pas cependant l'approximation stochastique: il lit par exemple deux fois l'ensemble I des n individus disponibles en appliquant l'appr. stoch. comme si ces 2n individus provenaient d'une source ouverte ; puis il effectue quelques autres lectures de I (par exemple trois) pour multiplier par une puissance de la matrice d'inertie (ici la puissance troisième) les vecteurs axiaux

fournis par l'appr. stoch. . L'intérêt de ce procédé mixte est que l'appr. stoch. fournit rapidement des axes grossièrement approchés ; tandis que la puissance itérée se stabilise ensuite plus vite. De plus L. Lebart projette le nuage sur le sous-espace engendré par les axes approchés ainsi calculés en deux étapes : et effectue pour ce nuage projeté une analyse factorielle usuelle exacte, qui améliore l'orientation des premiers axes. Sans prétendre exposer ici, les remarquables travaux de L. Lebart (cf. *op. laud.* ch V) nous signalerons encore que dans l'étape d'approximation stochastique il trouve très avantageux de lire deux fois de suite l'ensemble I : une fois dans un ordre , puis une fois dans l'ordre opposé (1, 2,..., n ; puis n, (n-1),...1).

Reste la question de l'efficacité comparée des diverses méthodes disponibles par l'analyse factorielle des très grands tableaux de correspondance. La méthode de L. Lebart couramment employée par lui dans les dépouillements d'enquêtes est parfaitement au point. Cependant Brigitte Escofier explore dans sa thèse une toute autre voie :en bref si l'ensemble J des variables est d'un cardinal trop élevé, elle analyse des sous-tableaux $I \times Js$ (où Js est une partie de J) puis combine les résultats de ces analyses pour retrouver une approximation des facteurs afférents à $I \times J$ tout entier. Cette méthode est reprise dans [*BANDE BURT*] (C.A.D. Vol VII n° 1 ; 1982) d'un point de vue un peu différent. En bref,en adjoignant à $I \times Js$ une colonne "reste", cumulant l'ensemble J-Js des colonnes omises, on a du nuage N(J) une représentation euclidienne qui est celle fournie par l'analyse de $I \times J$, mais rapportée à des axes ajustés au sous-nuage N(Js). D'une part cette propriété facilite quelque peu la recherche des facteurs issus de $I \times J$, suivant une combinaison analogue à celle de B. Escofier. D'autre part les analyses partielles sont dans une certaine mesure plus intéressantes que l'analyse globale elle-même : celle-ci étant souvent perturbée par des modalités singulières des variables, notamment les omissions ; tandis que les vues de N(J) ajustées à des sous-nuages N(Js) bien choisis montrent le mieux la structure globale.

Sans prétendre offrir au lecteur de conclusion décisive, alors que les progrès de l'informatique bouleversent sans cesse les calculs de coût et de faisabilité, nous soulignerons au terme de cette introduction que des idées diverses, suggérées à plusieurs auteurs par des préoccupations souvent très éloignées les unes des autres, contribuent à nous donner plus de prise sur les très grands ensembles de données.

2 *Calcul des facteurs par approximation stochastique*

2.1 Soit $\{k(i,j)\}$ un tableau de nombres indicé par les ensembles finis I et J. Supposons que I soit un ensemble d'individus dont le choix dépend de diverses contingences d'échantillonnage, tandis que J est un ensemble bien déterminé de variables. Il est alors légitime d'assigner pour objectif principal à l'analyse factorielle le calcul des valeurs propres λ_α et des facteurs $G_\alpha^J = \{G_\alpha(j) \mid j \in J\}$, sur J ; le calcul des valeurs $F_\alpha(i)$ se faisant, pour les individus auxquels on s'intéresse, par la formule de transition. Les facteurs G_α^J sont eux-mêmes définis comme vecteurs propres de l'application linéaire :

$$G^J \to G^J \circ f_J^I \circ f_I^J = G^J \circ \pi_J^J \qquad (1)$$

(cf TII B n° 5, §§ 3 & 5 et TII B n° 7, § 1.3.1). Les composantes de π_J^J sont données par les formules :

$$\pi_{j'}^{j} = \Sigma\{(k(i,j')/k(i))\ (k(i,j)/k(j)) \mid i \in I\} \quad ; \qquad (2)$$

$$k(i) = \Sigma\{k(i,j) \mid j \in J\}$$

$$k(j) = \Sigma\{k(i,j) \mid i \in I\} \quad ;$$

et le facteur G_α satisfait à l'équation :

$$G_\alpha(j) = \lambda_\alpha^{-1}\Sigma\{G_\alpha(j')\ \pi_{j'}^{j} \mid j' \in J\}. \qquad (1')$$

De cette formule, il est intéressant de passer au cas limite d'un ensemble I potentiellement infini, ou source aléatoire infinie d'individus. On supposera donc que les lignes (ou individus indicés par i) du tableau à analyser sont un échantillon de points de R_J (espace vectoriel des mesures sur l'ensemble fini J) tirés au sort suivant une loi $f(z_J)\ d\ z_J$. On définira alors π par l'équation :

$$\pi_{j'}^{j} = \int\{(z_{j'}/k(z_J))(z_j/E_j)\ f(z_J)\ dz_J \mid z_J \in R_J\} \quad ; \qquad (3)$$

où on a posé :

$$E_j = \int\{z_j\ f(z_J)\ dz_J \mid z_J \in R_J\}\ ,$$

$$k(z_J) = \Sigma\{z_j \mid j \in J\}\ ;$$

(avec pour $\int_X f(x)\ dx$ la notation sans indices : $\int\{f(x)\ dx \mid x \in X\}$).

La première formule donnée pour $\pi_{j'}^{j}$ devient un cas particulier de la seconde si on prend pour mesure $f(z_J)\ dz_J$ (loi de probabilité des individus) la mesure constituée par le système fini des masses ponctuelles égales à 1/Card I, placées aux points suivants z_{iJ} de R_J, qui ne sont autres que les lignes du tableau k_{IJ} :

$$z_{iJ} = \{k(i,j) \mid j \in J\} \quad ;$$

la formule intégrale devient alors en effet :

$$\pi_{j'}^{j} = \Sigma\{(k(i,j')/k(i))\ (k(i,j)/E_j)/\text{Card } I \mid i \in I\}\ , \qquad (4)$$

où : $E_j = \Sigma\{k(i,j) \mid i \in I\}/\text{Card } I = k(j)/\text{Card}I$.

.2.2 Cependant, considérer les lignes du tableau, comme provenant d'une source aléatoire, suggère d'appliquer à l'analyse factorielle des correspondances, le calcul par approximation stochastique. Le schéma général d'un tel calcul est en bref le suivant. Interviennent deux espaces D et X, l'espace des données et l'espace de l'inconnue: l'inconnue est un point $x \in X$, la donnée est une suite (que l'on suppose infinie) de points d(i) de D. L'algorithme d'approximation est une fonction Φ de X x D à valeur dans X; on pose :

$x(1)$ = un point initial quelconque

$x(2) = \Phi(x(1), d(1))$

$\vdots$

$$x(i+1) = \Phi(x(i), d(i)) \qquad (5)$$

L'algorithme est théoriquement satisfaisant si, les d(i) étant aléatoires indépendants, de loi une mesure μ_D^* (qui est la donnée cachée du problème), les x(i) ont une distribution qui pout i suffisamment grand est arbitrairement concentrée au voisinage de la solution d'un problème (qui, répétons-le, concerne μ_D^*). Pratiquement les d(i) ne sont qu une suite finie, d'où une mesure μ, système fini de masses ponctuelles sur D : mais on peut faire défiler plusieurs fois de suite les mêmes données... . Evidemment les espaces D et X sont des espaces vectoriels de dimension finie ; si on opère sur une machine, il suffit de réserver en mémoire la place pour un vecteur $d \in D$ et un vecteur $x \in X$; les données d(i) sont introduites successivement ; d(i) et x(i) sont effacés dès que x(i + 1) est calculé ; on peut ainsi loger celui-ci à la place de x(i), et une nouvelle donnée d(i + 1) à la place de d(i). Ces méthodes d'approximation stochastique sont couramment appliquées, par exemple, en théorie de la régression, où elles donnent des résultats aussi sûrs que les méthodes algébriques tout en économisant mémoire et aussi calculs arithmétiques.

Donnons maintenant un schéma de recherche de valeurs propres par approximation stochastique : nous verrons ensuite que ce schéma semble bien adapté à l'analyse des correspondances. Soit {d(i)} une suite infinie de matrices n x n aléatoires indépendantes, tirées au sort suivant la même loi de probabilité ; notons d_M la matrice moyenne, (ou espérance mathématique), que nous supposerons symétrique.

Posons (cf équation 5) :

$x(1)$ = un vecteur quelconque de R^n,

$x(2) = x(1) + d(1)\, x(1)$

$\vdots$

$$x(i+1) = x(i) + (1/i)\, d(i)\, x(i) \ ; \qquad (6)$$

Pour i_1 et i_2 assez grands on a approximativement, (moyennant des hypothèses convenables que nous ne tenterons pas de préciser dans la présente leçon, cf op. laud. in Bull. Soc. Math. France)

$$x(i_2) \approx \exp(\mathrm{Log}(i_2/i_1).\ d_M).\ x(i_1) \ ; \qquad (7)$$

la justification heuristique de cette formule étant l'équivalence approximative :

$$\Pi\{(1 + (d/i)) \mid i_1 < i \leq i_2\} \approx \exp(\langle\langle\ \Sigma_i\ i^{-1}).\ d) \quad ; \qquad (8)$$

où l'on remplace les termes tous égaux à d, par des termes d(i) dont l'espérance mathématique est d_M. Ainsi, le vecteur x(i) construit suivant le processus (6) est une approximation du vecteur propre de d_M relatif à sa plus grande valeur propre λ_1 ; tandis que le rapport de $x(i_2)$ à $x(i_1)$, (qui lui est à peu près parallèle) nous donne $\exp(\lambda_1 . \log(i_2/i_1))$.

Pour approcher les p premières valeurs propres $\lambda_1 \cdots \lambda_h \cdots \lambda_p$, et les vecteurs propres correspondants de d_M, on construira une suite de p-uples de vecteurs $\{x_h(i)\}$, définie par l'algorithme suivant :

$$\forall\, h \in \{1,\ldots, p\} : y_h(i+1) = x_h(i) + \frac{1}{i}\, d(i)\, x(i)$$

$$\{x_h(i+1) \mid h = 1,\ldots, p\} = \mathrm{Orth}\{y_h(i+1) \mid h = 1,\ldots, p\}, \qquad (9)$$

où le symbole Orth désigne l'orthogonalisation de Schmidt sans normalisation ; i. e. on a :

$$x_h(i) = y_h(i) + \Sigma_{h'<h}\, t_h^{h'}\, y_{h'}(i) \qquad (10)$$

les coefficients $t_h^{h'}$ (nuls si $h' \geq h$) étant choisis tels que les $x_h(i)$ soient orthogonaux deux à deux. Ainsi, $x_h(i)$ donnera le h-ème, vecteur propre ; et le rapport de $x_h(i_2)$ à $x_h(i_1)$ (qui lui est à peu près parallèle) donnera $\exp(\lambda_h . \log(i_2/i_1))$, où λ_h est h-ème valeur propre de d_M. (Il s'agit évidemment de valeurs approchées dont la convergence n'est pas étudiée ici...)

2.3 Revenons à l'analyse factorielle des correspondances. Le coefficient de matrice $\pi_j^{j'}$, donné par l'équation (3) peut s'interpréter comme le quotient par Ej de l'espérance mathématique du quotient :

$$(z_j \cdot z_{j'}) / \Sigma\{z_j \mid j \in J\} \quad ; \qquad (11)$$

Ceci va nous permettre d'utiliser l'algorithme (9), à des modifications près dues à ce que le coefficient Ej est lui-même une espérance mathématique qui doit être estimée. Pour donner l'algorithme d'analyse factorielle, précisons nos notations. La donnée est une suite, indicée par i de vecteurs dont les composantes sont indicées par $j \in J$. On suppose que les vecteurs, notés $\{K_i(j) \mid j \in J\}$, sont introduits successivement, un à un dans la mémoire de calcul (quoique, en fait, il convienne d'introduire les vecteurs par paquets, aussi importants que le permet la mémoire) ; ainsi l'indice j est un indice de tableau, et c'est pourquoi on le note fonctionnellement ; l'indice i au contraire ne peut être présent dans la machine que sur un compteur auxiliaire, on figure donc i comme un indice usuel. On note K_i la somme $\Sigma_j K_i(j)$; on garde de plus en mémoire le tableau (à une dimension) des $C(j)$ dont les valeurs à l'étape i, sont :

$$C_i(j) = \Sigma \{ K_{i'}(j) \mid i' \leq i \} \quad ; \qquad (12)$$

Les inconnues sont les p premiers facteurs ; on a donc un tableau d'inconnues G(j ; h) où j ∈ J, et h varie de 1 à p ; à la i-ème itération (à l'introduction de la i-ème donnée) ce tableau a pour valeur $G_i(j ; h)$.

Ceci posé l'algorithme s'écrit (en omettant l'orthogonalisation) :

$$G_{i+1}(j ; h) = G_i(j ; h) + (K_i(j)/C_i(j)) \Sigma\{(K_i(j')/K_i)\ G_i(j' ; h) \mid j' \in J\} \ ; \qquad (13)$$

Pour voir le lien de cet algorithme avec les formules (9) et (11) reprenons les notations de la formule (4). On peut alors récrire le second membre de (13), (en omettant l'indice h) :

$$G_i(j) + (k(i,j)/(iE_{ij})) \Sigma\{(k(i,j')/k(i))\ G_i(j') \mid j' \in J\} \ ; \qquad (14)$$

dans cette dernière expression, on a noté :

$$C_i(j) = i\ E_{ij}$$

parce que $C_i(j)/i$, qui est la moyenne des i premiers coefficients k(i,j) connus, est l'estimation, disponible au pas i, de l'espérance mathématique E_j des k(i',j) (ceux-ci considérés comme j-ème composante d'un vecteur aléatoire, dont chaque ligne du tableau de données est une réalisation).

On remarquera que le calcul est simplifié parce que la somme à droite de l'équation(13), se retrouve égale à elle-même dans toutes les composantes d'un facteur de rang h donné : géométriquement ceci correspond au fait qu'ici, pour reprendre les notations de la formule (6), les applications linéaires d(i) sont toutes de rang 1 (les d(i) sont en bref les contributions des individus i successifs à la matrice d'inertie du nuage $\mathcal{N}(J)$ dans R_J).

Pour avoir un algorithme complet, il reste à introduire l'orthogonalisation des facteurs ; mais il est inutile d'orthogonaliser à chaque pas, on ne le fera que si i est, e.g., un multiple de 20, voire moins souvent... Remarquons que l'on doit orthogonaliser par rapport au produit scalaire :

$$< X, Y > = \Sigma\{X(j)\ Y(j)\ k(j) \mid j \in J\} \ ;$$

les k(j) étant inconnues, on leur substitue les $C_i(j)$; comme on sait qu'il y a un vecteur propre égal à 1 on peut réduire les vecteurs à lui être orthogonal. Dans la pratique, certains préfèrent voir ce vecteur propre trivial apparaître graduellement au cours des itérations, et juger ainsi de la convergence du processus d'approximation stochastique. Mais nous estimons qu'orthogonaliser à 1 peut accélérer la convergence.

Nous n'avons rien dit non plus du calcul des valeurs propres : on sortira deux fois les facteurs, e.g. pour i = 100 et i = 200, (ou 1000 et 2000!), et on fera les quotients approchés :

$$G_{200}(j ; h)/G_{100}(j ; h), \ ; \qquad (15)$$

valeurs estimées des $\exp(\lambda_h \ .\ \mathrm{Log}\ 2)$

Notons pour conclure ce paragraphe que la méthode proposée permettrait de traiter des tableaux dont la plus petite dimension est de l'ordre de 1000 ; avec une méthode algébrique non stochastique on devrait calculer sur une matrice d'inertie 1000 x 1000, matrice dont les 10^6 coefficients ne tiennent, croyons-nous, dans la mémoire de calcul d'aucune machine usuelle (mais qu'il est toutefois possible d'appeler d'une mémoire auxiliaire). Et les essais d'application rapportés par J.-P. Fénelon dans sa thèse (3° cycle Paris 1973) encouragent à poursuivre dans la voie de l'approximation stochastique.

P.S. : depuis la rédaction de cette leçon, l'expérimentation de la méthode d'approximation stochastique s'est poursuivie, comme on l'explique au § 0.