LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Sur l'analyse d'un tableau dont l'une des colonnes à un poids prédominant

Les cahiers de l'analyse des données, tome 4, n° 4 (1979), p. 413-416

<http://www.numdam.org/item?id=CAD_1979__4_4_413_0>

Les Cahiers de l'Analyse des Données
Vol. IV 1979 n° 4 p. 413 416

SUR L'ANALYSE D'UN TABLEAU DONT L'UNE DES COLONNES A UN POIDS PRÉDOMINANT [COL. PRED.]

par J P Benzécri (1)

1 *Un exemple de données naturelles*

On trouve publiées au Journal Officiel (cf e.g. J.O. 10 avril 1926, annexe) des statistiques démographiques de la forme suivante : pour chaque département (ou arrondissement) sont donnés après sa population, le nombre pour une année, des naissances, des mariages, des décès, des divorces, des naissances d'enfants mort-nés et des décès d'enfants de moins d'un an. Comment analyser un tel tableau ? Signalons d'abord que si l'on prend pour base l'événement affectant un individu, il semble naturel de multiplier par deux les colonnes de mariage et divorce dont chaque élément concerne deux individus. Ainsi on aura dans le tableau trois colonnes à peu près égales en poids : naissance, mariés, décédés ; trois colonnes comparables entre elles et bien plus légères que les trois premières : mort-nés, morts en bas âge, divorcés ; et enfin une colonne population dont le poids est tout à fait prédominant, puisqu'il est à peu près le produit de la colonne naissance par l'espérance de vie comptée en années. On pourrait par des coefficients amener les sept colonnes à des poids semblables. Ici, on tentera, par une réflexion mathématique, d'éclairer l'analyse d'un tableau dont l'une des colonnes a un poids prédominant. L'article suivant offre un exemple de telle analyse, due à L. Lebart.

2 *Description mathématique de la situation typique*

Soit I un ensemble fini dont chaque élément i désigne une ligne des tableaux rectangulaires analysés ici ; soit J l'ensemble des colonnes, et $J^+ = J \cup \{0\}$, ce même ensemble complété d'une colonne exceptionnelle, d'indice 0, dont nous ferons tendre le poids vers l'infini. On note $\{k(i,j) \mid i \in I, j \in J\}$ un tableau de correspondance donné sur I x J ; et $\{k^1(i,0) \mid i \in I\}$ une colonne isolée dont on suppose que tous les éléments sont strictement positifs et ont pour somme 1. A partir de ces données, on construit une famille continue de tableaux sur I x J^+ dépendant d'un paramètre t qui varie dans R^+ de 0 à l'infini ; on pose :

$$\forall\, t \in R^+,\ \forall\, i \in I,\ j \in J \quad : \quad k^t(i,j) = k(i,j) \quad ;$$

$$\forall\, t \in R^+ \quad \forall\, i \in I \quad : \quad k^t(i,0) = t k^1(i,0) \quad .$$

Autrement dit, le tableau k^t est obtenu en bordant le tableau k d'une colonne $k^t(i,0)$, proportionnelle à $k^1(i,0)$ et de poids total t. D'après le comportement limite des résultats du tableau k^t, pour t tendant vers l'infini, nous serons à même d'interpréter les graphiques issus d'un tableau naturel, tel que celui considéré ci-dessus (au § 1).

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

3 *Distances distributionnelles limites*

La distance distributionnelle sur l'ensemble I est donnée par la formule :

$$d^2(i,i') = (k^t(i,0)/k^t(i)) - (k^t(\imath',0)/k^t(i'))^2 (k^t(0)/k^t) + \ldots$$

$$\ldots \quad (k(i,j)/k^t(i)) - (k(i',j)/k^t(i'))^2/(k(j)/k^t \; j \; J \; ;$$

où l'on a supprimé l'indice t chaque fois qu'il s'agit de termes ne dépendant pas effectivement de t (et provenant du tableau k sur I x J). On a les formules :

$$k^t(\imath) = k(i) + tk^1(\imath,0) \; ; \; k^t(0) = tk^1(0) = t, \text{ (cf hyp. n° 2).}$$

$$k^t = \quad k(i) \; i \quad I + t \quad k^1(i,0) \; i \quad I = k^0 + t \; ;$$

On en déduit pour $d^2(i,\imath')$ une expression asymptotique en t^{-1} valable pour t tendant vers l'infini. On a :

$$k^t(i,0)/k^t(i) = k^1(\imath,0)/(k^1(\imath,0) + k(i)/t) \quad 1 - (k(i/k^1(i,0))t^{-1};$$

$$k^t(\imath,0)/k^t(i)) - k^t(\imath',0)/k^t(i'))^2/(k^t(0)/k^t) = 0(k^t) = 0(t^{-2};$$

$$d^2(\imath,i') \quad t^{-1} \quad (k(\imath,j)/k^1(i,0)) - (k(i',j)/k^1(\imath,0))^2/k(j) \; j \quad J \quad ,$$

formule où l'on reconnaît l'expression usuelle de la distance distributionnelle pour la correspondance k'i,j) sur I x J, à ceci près qu'aux sommes marginales $k(\imath)$ on a substitué $k^1(\imath,0)\; t^{1/2}$. Cette formule laisse attendre que, comme on le démontrera au § 4, les valeurs propres sont des $\grave{a}(t^{-1})$; les facteurs considérés comme des fonctions φ^i sur I de moyenne nulle et de variance 1, (chaque i ayant pour masse limite $k^1(i,0))$, tendent vers une limite quand t croît indéfiniment ; tandis que les fonctions F(i) (coordonnées des points du nuage rapporté aux axes factoriels, sont des $0(t^{-1/2})$ (comme les distances, dont elles sont les composantes principales).

Considérons maintenant les distances distributionnelles sur l'ensemble J^+. On a quels que soient j et j' dans J :

$$d^2(\jmath,\jmath') = \quad (k(i,j)/k(j)) - (k(i,j')/k(j'))^2/(k^t(i)/k^t)_{\; \imath \in I} \; ;$$

$$d^2(\jmath,0) \quad = \quad (k(i,j)/k(\jmath)) - (k^t(\imath,0)/k^t(0))^2/(k^t(i)/k^t)_{i \in I} \; .$$

Ainsi sur l'ensemble J^+, les distances ont des limites finies quand t tend vers l'infini. Nous savons déjà, d'après l'étude de l'ensemble I, que les valeurs propres sont des $0(t^{-1})$. Quant aux facteurs leur étude est compliquée par le fait que, contrairement à ce qui est sur l'ensemble I, la loi marginale sur J^+ a une limite singulière : la masse p^t_0 a pour limite 1, et les p^t_j (masses relatives des j) sont des $0(t^{-1})$. Le point 0 tend donc à être au centre de gravité, g, du nuage J^+, et la distance 0g est un $0(t^{-1})$ (puisqu'elle est de l'ordre de 0j = 0(1) multiplié par la masse relative des j qui est un $0(t^{-1})$) . On a donc après projection sur les axes factoriels :

$$G(0) = 0(t^{-1}) \; ; \; G(j) = 0(1) \; ;$$

et pour les facteurs normalisés φ^j :

$$\varphi^0 = 0(t^{-1/2}) \; ; \quad j = 0(t^{1/2}).$$

Sur la représentation simultanée usuelle (par F et G) des deux ensembles I et J^+, il faut donc s'attendre à ce que le point 0 soit très proche de l'origine (t^{-1}) ; les points j soient les plus éloignés (t^0) et les points i à une distance intermédiaire ($t^{-1/2}$).

4 Equations limites des facteurs

On sait qu'un facteur φ relatif à la valeur propre λ satisfait à l'équation :

$$\forall\, i \in I : \varphi^i = \lambda \Sigma\{s^i_{i'} \varphi^{i'} \,|\, i \in I\},$$

dont les coefficients $s^i_{i'}$, s'expriment comme suit en fonction de t :

$$s^i_{i'}(t) = (1/K^t(i))\,[t\,k^1(i,0)k^1(i',0) = \Sigma\{k(i,j)k(i',j)/k(j) \,|\, j \in J\}].$$

(Dans cette formule on a tenu compte de l'égalité :

$$k^t(i,0)k^t(i',0)k^t(0) = tk^1(i,0)k^1(i',0)\ .)$$

On munit R^I du produit scalaire :

$$\langle \varphi^I, \psi^I \rangle = \Sigma\{\varphi^i \psi^i \; p_i(t) \,|\, i \in I\} \; ; \text{ où :}$$

$$p_i(t) = k^t(i)/k^t = (k(i) + tk^1(i,0))/(t + k^0).$$

Les facteurs relatifs aux diverses valeurs propres sont alors orthogonaux, car relativement à ce produit scalaire $s^i_{i'}(t)$ est symétrique. En particulier tout facteur non trivial φ^I est orthogonal au facteur trivial δ^I, (constant égal à un, relatif à la valeur propre 1) et est donc contenu dans l'hyperplan H^I de R^I :

$$H^I = \{u^I \,|\, u^I \in R^I \; ; \; \langle u^I, \delta^I \rangle = 0 \quad ;$$

L'hyperplan dépend de t, mais tend vers une limite H^I_{inf} quand t tend vers l'infini.

L'opérateur π de projection orthogonale de R^I sur la droite (de vecteur directeur δ^I) des fonctions constantes, a pour composantes :

$$\pi^i_{i'}(t) = p_{i'}(t) = (k(i') + tk^1(i',0))/k^0 + t) \; ;$$

le projecteur $\pi(t)$ envoie sur 0 l'hyperplan $H^I(t)$ orthogonal à δ^I. De l'opérateur s(t) retranchons $\pi(t)$: il reste un opérateur r(t), (endomorphisme de R^I comme s et π) qui admet δ^I comme vecteur propre relatif à la valeur propre 0, laisse invariant l'hyperplan $H^I(t)$ et y a les mêmes vecteurs propres que s(t), relatifs aux mêmes valeurs propres : le reste r(t) nous donne donc les facteurs non-triviaux. Or on a :

$$r^1_{i'}(t) = (1/k^T(i))[\Sigma\{k(i,j)\; k(i',j)/k(j) \,|\, j \in J\} + \ldots$$
$$\ldots\; tk^1(i,0)k^1(i',0) - (k^t(i)k^t(i')/k^t)\,].$$

Le premier facteur $(1/k^t(i))$ est un $0(t^{-1})$; la somme ne dépend pas de t ; les derniers termes du crochet ont, quand t tend vers l'infini, une limite finie car :

$$(t+k^0)^{-1}\,[(t+k^0)\,tk^1(i,0)k^1(i',0) - (k(i)+tk^1(i,0))(k(i')+tk^1(i',0))] =$$
$$k^0k^1(i,0)k^1(i',0) - k(i)k^1(i',0) - k(i')k^1(i,0) + 0(t^{-1}) .$$

Ainsi la matrice r(t) est un $0(t^{-1})$ ce qui nous confirme la propriété annoncée au n° 3, que les valeurs propres (correspondant aux facteurs non-triviaux) sont des $0(t^{-1})$. De plus, les facteurs φ^i et leurs valeurs propres λ, multipliées par t, s'obtiennent comme vecteurs propres et valeurs propres de la transformation :

$$w^i_{i'}(t) = \pi^i_{i'}(t) + t\,r^i_{i'}(t).$$

Cette transformation w a de plus le vecteur propre δ^I relatif à la valeur propre 1. Quand t tend vers l'infini, w(t) tend vers une application linéaire non-dégénérée de R^I dans lui-même, symétrique relativement au produit scalaire $\Sigma\{\varphi^i\psi^ik^1(i,0)\}$; les vecteurs propres de $w(\infty)$ sont les facteurs limites : ses valeurs propres sont les produits par t des valeurs propres limites (sauf pour la valeur propre $\lambda = 1$ qui subsiste avec δ^1). on notera que l'on a :

$$w^i_{i'}(\infty) = k^1(i',0) + (1/k^1(i,0))[\Sigma\{k(i,j)k(i',j)/k(j) \mid j \in J\} + \ldots$$
$$\ldots\ k^0k^1(i,0)k^1(i',0) - k(i)k^1(i',0) - k(i')k^1(i,0)] .$$

Quant aux facteurs φ^j on peut les obtenir à partir des facteurs φ^i par l'équation usuelle :

$$\varphi^0\ \lambda^{1/2} = G(0) = \Sigma\{k^1(i,0)\varphi^i \mid i \in I\}$$
$$\varphi^j\ \lambda^{1/2} = G(j) = \Sigma\{k(i,j)\varphi^i/k(j) \mid i \in I\}.$$

On retrouve, comme annoncé au § 3 que $G(j) = 0(1)$; mais $G(0)$ est un $0(t^{-1})$ parce que φ^i est de moyenne nulle pour la mesure $p_i(t)$ qui ne diffère de $k^1(i,0)$ que par un $0(t^{-1})$.

Signalons sans démonstration que si les deux colonnes $k^1(i,0)$ et $k(i)$ sont proportionnelles (si on adjoint au tableau k une colonne 0 proportionnelle à sa marge) les facteurs φ ne dépendent pas de t, on a alors $\varphi^0 = 0$.