STATISTIQUE ET ANALYSE DES DONNÉES

HENRI CAUSSINUS
LOUIS FERRE

Analyse en composantes principales d'individus définis par les paramètres d'un modèle

Statistique et analyse des données, tome 14, n° 3 (1989), p. 19-28.
<http://www.numdam.org/item?id=SAD_1989__14_3_19_0>

Statistique et Analyse des Données
1989 - Vol. 14 n°3 - p. 19 - 28

ANALYSE EN COMPOSANTES PRINCIPALES D'INDIVIDUS DEFINIS PAR LES PARAMETRES D'UN MODELE.

Henri CAUSSINUS et Louis FERRE

Laboratoire de Statistique et Probabilités
Université Paul Sabatier - U.R.A. C.N.R.S. D 745
118, route de Narbonne
31062 TOULOUSE CEDEX

Résumé - *On discute comment procéder à une Analyse en Composantes Principales cohérente lorsque les variables sont les estimations des p paramètres d'un modèle donné pour n individus différents. Un cas particulier important est celui des courbes de croissance ; dans ce cas, cette note arrive aux conclusions d'Houllier (1987) en plaçant l'étude sous un jour quelque peu différent qui permet de nouveaux développements. Nous en donnons deux : le choix de poids pour les individus et le choix du nombre de dimensions que l'analyse devrait retenir.*

Summary - *We discuss the proper way of performing Principal Component Analysis when the variates are estimates of the p parameters of a given model. The special case of growth curves is an important one. In this case we are led to the same conclusions as Houllier (1987) by setting the problem in a slightly different framework. This framework allows us to give further developments. One of them concerns the choice of suitable weights for the units. Another one concerns the optimal choice of dimensionality.*

Mots-clés : *Analyse en Composantes Principales, courbes de croissance, modèles, métriques, pondérations.*

Indices de classification STMA : 06.070, 04.160, 07.140.

Manuscrit reçu le 14 mars 1989, révisé le 12 février 1990

1 - INTRODUCTION

A l'origine des réflexions que nous présentons ici se trouve un article sur la comparaison de courbes et de modèles de croissance (Houllier, 1987). L'auteur y discute un choix naturel de métrique permettant de remplacer l'Analyse en Composantes Principales (ACP) des données brutes (mesures successives de la variable observée) par une ACP convenable des paramètres d'un modèle ajusté à ces données. Notre objectif est de montrer comment cette démarche s'intègre simplement dans le cadre du "modèle à effets fixes", cadre que nous avons préconisé pour son intérêt synthétique et son utilité dans certains développements pratiques de l'ACP : choix de métrique, problèmes de dimension (Caussinus, 1986a,b ; Besse et al., 1986, 1987,1988 ; Ferré, 1989,1990). Cela nous conduit à compléter la discussion d'Houllier (1987) sur quelques points. Le second paragraphe rappellera donc le modèle à effets fixes (ou fonctionnel), tandis que le troisième cherchera à discuter dans quelle mesure le problème considéré peut s'y rattacher. Nous pourrons alors envisager d'améliorer certaines analyses par des pondérations adéquates des individus (paragraphe 4). Nous montrerons ensuite que certains résultats sur le choix de dimension (Ferré, 1989) s'appliquent dans la situation présente (paragraphe 5).

2 - MODELE A EFFETS FIXES : RAPPEL ET NOTATIONS

Soit Y une matrice $n \times p$ d'éléments Y_i^j ($i=1,\ldots,n$; $j=1,\ldots,p$). Le vecteur colonne transposé de la ligne i de Y sera noté $Y_{i.}$. Le modèle à effets fixes correspond aux hypothèses suivantes :

(H)
(i) les Y_i ($i=1,\ldots,n$) sont des vecteurs aléatoires indépendants à valeurs dans $\mathbb{R}^p$;
(ii) Soit $\mathbb{E}(Y_i) = y_i$, il existe une variété linéaire F_{q_0} de dimension q_0 (q_0 donné) telle que $y_i \in F_{q_0}$ pour tout $i=1,\ldots,n$;
(iii) $\mathrm{Var}(Y_i) = \frac{\sigma^2}{w_i}\,\Gamma$ (les réels w_i et la matrice symétrique définie positive Γ sont supposés connus; σ est un réel éventuellement inconnu).

Dans ce cas, l'ACP (de dimension q_0) de la matrice des données Y, avec les poids w_i et la métrique $M = \Gamma^{-1}$ sur l'espace $\mathbb{R}^p$ des individus, fournit une estimation de

moindres carrés de F_{q_0} et des paramètres y_i, estimation optimale en un certain sens (Besse et al., 1987, 1988).

3 - ACP D'INDIVIDUS DEFINIS PAR LES PARAMETRES D'UN MODELE

Supposons maintenant que les éléments de Y_i estiment les p paramètres d'un modèle pour l'individu i. Dans quelle mesure les hypothèses (H) sont-elles raisonnables ?

(i) La première hypothèse est évidemment réaliste si les essais ou observations ayant conduit aux n estimations sont indépendants, ce qui est le plus souvent réalisé.

(ii) La deuxième hypothèse est implicite, quelle que soit la présentation retenue, dès lors que les individus sont représentés en q_0 dimensions au moyen d'une ACP. On peut considérer que y_i est la vraie valeur du paramètre pour le $i^{ème}$ individu ; $\mathbb{E}(Y_i) = y_i$ est le cas de l'estimation sans biais (modèle linéaire en particulier), sinon cette relation reste en général valable avec une approximation raisonnable, et d'autant plus que les données initiales par individu sont nombreuses. Il est clair alors que l'objectif est d'approcher au mieux les y_i par une estimation permettant une représentation en un nombre réduit q de dimensions. Il a été démontré que la valeur optimale de q (celle qui permet la meilleure estimation des y_i au sens d'une erreur quadratique moyenne) n'est pas nécessairement la "vraie" dimension q_0 introduite dans (H) (ii) (en pratique d'ailleurs impossible à préciser) mais une dimension éventuellement plus faible (Besse et al., 1987, 1988). D'ailleurs, on peut même admettre que la vraie dimension q_0 est égale à p, ce qui est trivialement vérifié, dès lors que σ^2 est connu (ou estimable sans que q_0 soit strictement inférieur à p).

(iii) En général, on peut écrire ici $Var(Y_i) \simeq (1/m_i)\,\Gamma_i$ (ou $(1/(m_i-p))\,\Gamma_i$) où m_i est le nombre de données brutes pour le $i^{ème}$ individu. La matrice Γ_i est inconnue, mais estimable à partir des données brutes dans tous les cas usuels. Bien entendu, la qualité de cet estimateur dépendra du modèle, de la technique d'estimation utilisée (sans doute souvent le maximum de vraisemblance) et surtout de m_i, puisqu'il faut fréquemment se contenter d'estimateurs dont seules les qualités asymptotiques (m_i grand) sont avérées. D'autre part, Γ_i varie en général avec i. Si cette variation est petite, l'hypothèse (H) (iii) est raisonnable avec $w_i = m_i/m$, $\sigma^2 = 1/m$, et Γ remplacé par une estimation "moyenne",

où nous avons noté $m = \sum_{i=1}^{n} m_i$ le nombre total de données brutes (on a donc $\sum_{1}^{n} w_i = 1$ et, dans la pratique, σ^2 est petit). Cette discussion est très voisine de celle de Houllier (1987) sur ce point (§ 2.3 et 2.4) pour le cas des modèles de croissance. Mais on peut encore améliorer l'approximation des $(1/m_i)\,\Gamma_i$ par une expression de la forme $(\sigma^2/w_i)\,\Gamma$ en cherchant des poids "meilleurs" que m_i/m : c'est ce qui sera discuté au paragraphe 4 ci-dessous.

Remarques.

1 - Si l'étude comparative de plusieurs courbes de croissance "modélisées" comme dans Houllier (1987) entre évidemment dans le cadre ci-dessus, et fournit un exemple particulièrement intéressant, d'autres analyses plus classiques peuvent, entre autres, être présentées de cette façon : c'est fondamentalement ce qui est fait pour l'Analyse des Correspondances dans Caussinus (1986 a,b) ; c'est ce qu'on peut faire aussi pour l'Analyse Factorielle Discriminante dans laquelle les centres de gravité de chaque groupe estiment bien le vecteur moyenne de chaque sous-population (on notera dans ce cas que la méthode du paragraphe 4 ci-dessous peut permettre une analyse plus correcte que l'analyse classique faisant intervenir la matrice de variances et covariances intra-classes usuelle).

2 - Nos arguments pour le choix de la métrique à partir du modèle à effets fixes s'appuient sur la structure de covariance du bruit (i.e. sur Γ), l'objectif étant de représenter au mieux la partie "pertinente" des individus (i.e. les y_i). Or, la structure de covariance de l'estimateur d'un paramètre multidimensionnel se trouve en étudiant comment se transfère sur cet estimateur la structure de covariance des données. Bien que se plaçant dans une optique différente, c'est un calcul de même nature que fait Houllier (1987) pour analyser le passage de l'ACP "ordinaire" des données brutes (il suppose la première ACP faite avec la métrique identité (§ 2.2)) à l'ACP des paramètres estimés. Dans notre optique, ce choix de la métrique identité au départ est bien équivalent à supposer les bruits sur chaque mesure indépendants et de même variance, hypothèse reprise en général dans l'estimation des Γ_i (notons cependant que, dans les deux optiques, on peut partir de bruits ayant n'importe quelle structure de covariance pourvu qu'elle soit connue à un facteur près).

4 - CHOIX DE LA METRIQUE ET DES POIDS

Nous développons ici la question soulevée au paragraphe 3 (iii). On estime, pour chaque i=1,...,n, la variance $(1/m_i)\,\Gamma_i$ de Y_i par une matrice qu'on notera C_i. Il faut approcher les C_i par des matrices de la forme $\alpha_i\, C$ où les α_i sont des réels positifs et C une matrice symétrique définie positive, afin de retrouver une situation analogue à celle du modèle H (§ 2(iii)).

Dans un premier temps, cherchons à approcher chaque C_i par $\alpha_i\, C$ sans introduire les contraintes $\alpha_i > 0$ et C définie positive. Les C_i et C peuvent être considérées comme des vecteurs de $\mathbb{R}^{(p^2)}$. Si la qualité de l'approximation est mesurée par la distance euclidienne canonique de cet espace le problème se réduit à l'ACP non centrée ordinaire (métrique identité, poids égaux) des unités C_i, i=1,...,n, laquelle consiste bien à chercher C et les scalaires α_i qui minimisent $\sum_{i=1}^{n} \left\| C_i - \alpha_i\, C \right\|^2$ où $\|.\|$ est la norme canonique de $\mathbb{R}^{(p^2)}$. Dans ce qui suit, le produit scalaire correspondant est noté <.,.>. Par ailleurs, lorsqu'il sera utile de reconfigurer C en une matrice colonne $p^2 \times 1$, celle-ci sera notée vec(C).

Montrons maintenant que les α_i obtenus sont toujours positifs et C est symétrique, définie positive dès que les C_i le sont.

La symétrie de C est immédiate, car C est une combinaison linéaire des C_i, éléments du sous-espace des matrices symétriques. Considérons l'ensemble $\mathcal{C}$ des matrices semi-définies positives ; $\mathcal{C}$ est un cône convexe fermé, pointé, dont l'intérieur est égal à l'ensemble des matrices définies positives. De plus, on montre que, quels que soient M_1 et M_2 appartenant à $\mathcal{C}$, on a $\langle M_1, M_2 \rangle \geq 0$. En effet, soit $\sum_{i=1}^{p} \lambda_i\, x_i\, {}^t x_i$ la décomposition spectrale de M_2, on a

$$\langle M_1, M_2 \rangle = \mathrm{tr}\,(M_1\, M_2) = \sum_{i=1}^{p} \lambda_i\ \mathrm{tr}\,(M_1\, x_i\, {}^t x_i) = \sum_{i=1}^{p} \lambda_i\ {}^t x_i\, M_1\, x_i \geq 0.$$

Soit, maintenant, $V_c = \sum_{i=1}^{n} \mathrm{vec}(C_i)\, {}^t[\mathrm{vec}(C_i)]$.

On a pour tout M_1 de C : $V_c\,\mathrm{vec}(M_1) = \sum_{i=1}^{n} \langle M_1, C_i \rangle\, \mathrm{vec}(C_i)$. Or $\langle M_1, C_i \rangle \geq 0$ $\forall i \in \{1,\ldots,n\}$, et donc $V_c\, M_1 \in C$.

Enfin, si $\forall i \in \{1,\ldots,n\}$, C_i est inversible, c'est-à-dire $C_i \in \overset{o}{C}$, et si $M_1 \in C \setminus \{0\}$, alors $\langle M_1, C_i \rangle > 0$, et donc $V_c\,\mathrm{vec}(M_1) \in \overset{o}{C}$, car $\overset{o}{C}$ est encore une cône convexe.

L'ACP des C_i conduit à vec(C) vecteur propre normé de V_c associé à la plus grande valeur propre, et $\alpha_i = \langle C_i, C \rangle$. Les résultats annoncés se déduisent alors de la proposition suivante, corollaire des théorèmes (3.2) p.6 et (3.26) p. 13 de Berman et Plemmons (1979) :

Soit E un espace vectoriel et C un cône convexe fermé, d'intérieur non vide, pointé. Soit A un endomorphisme de E tel que $AC \subset C$. Alors, le vecteur propre associé à la plus grande valeur propre de A appartient à C.

De plus, si $\forall M_1 \in C \setminus \{0\}$, $AM_1 \in \overset{o}{C}$, alors le vecteur propre en question appartient à $\overset{o}{C}$.

Remarques :

1. Le principe de la méthode proposée ci-dessus est tout à fait équivalent à celui du "compromis" introduit par Escoufier (1980) dans un contexte différent.

2. On peut changer la métrique sur l'espace des matrices $p \times p$. Il peut aussi, et surtout, paraître naturel de considérer seulement les triangles supérieurs des matrices symétriques C_i et C afin de ne pas doubler l'importance des covariances. On travaille alors dans $\mathbb{R}^{p(p+1)/2}$ au lieu de $\mathbb{R}^{(p^2)}$; on peut vérifier que les α_i et le C ainsi obtenus conservent les propriétés requises.

EXEMPLE.

Nous avons repris les données de Bailly (1986) concernant la croissance de plantations forestières. Comme dans Bailly (1986) et Houllier (1987), le modèle utilisé pour la croissance de la plantation i est :

$$(1) \qquad Z_{it} = y_i^{1} \left(1 - y_i^{4} \exp((y_i^{2} - t)/y_i^{3})\right)^{-(1/y_i^{4})} + E_{it}$$

où Z_{it} est la différence entre la circonférence moyenne au temps $t+t_0$ (en années) et au temps t_0, origine de l'étude. Les "erreurs" E_{it} sont supposées indépendantes et de même variance. L'estimation du paramètre à 4 dimensions $y_i = [\overset{1}{y_i}, \overset{2}{y_i}, \overset{3}{y_i}, \overset{4}{y_i}]$ est faite au moyen des observations pour $t = t_0+1$ à $t = t_0+t_1$ (t_1 variant de 15 à 20 selon le cas) sauf éventuellement celles pour lesquelles l'expression (1) de Z_{it} n'aurait pas de sens (cette possibilité est d'ailleurs un des inconvénients du modèle adopté, mais ce n'est pas notre propos de discuter ce point). Ce faisant, m_i varie entre 13 et 19. Une autre difficulté vient de la grande instabilité apparente des estimations, puisque différents logiciels ont donné des estimations différentes des paramètres et surtout des matrices de variances et covariances. Ces différences ne semblent pas néanmoins beaucoup perturber la suite de l'analyse que nous proposons de faire. Nous avons poursuivi avec les estimations obtenues qui donnaient la plus petite somme des carrés résiduelle. Si l'on se réfère aux graphiques donnés par Bailly (1986), on voit que l'approximation des divers C_i par une matrice α_i C ne semble pas irréaliste, mais que les α_i peuvent être très différents d'une plantation à l'autre. C'est ce que l'on obtient avec la méthode ci-dessus, la liste des α_i étant :

12.495	11.975	82.312	52.348
42.447	329.439	6399.421	12.460
107.577	32.613	81.197	43.693
85.048	382.315	31.106	14.707
72.569	1.361	2.729	6.204
190.370	31.317	7.456	1.933
.914	15.698	101.827	140.425
21.546	13.594	15.092	21.527

et la matrice C étant :

9.986E-01	1.447E-02	1.133E-01	-2.185E-03
1.144E-02	3.498E-04	1.490E-02	-2.496E-05
1.133E-01	1.490E-03	1.327E-02	-2.599E-04
-2.185E-03	-2.496E-05	-2.599E-04	5.346E-06

Par ailleurs la proportion d'inertie expliquée par le premier facteur de l'ACP des C_i est 99.96% ce qui prouve bien que, dans cet exemple, le modèle adopté s'adapte bien aux données grâce aux pondérations effectuées.

L'ACP des paramètres estimés (p=4) pour les n=32 plantations, avec les poids (α/α_i) (où $\alpha = \left(\sum_{i=1}^{p} 1/\alpha_i \right)^{-1}$) et le produit scalaire $M=C^{-1}$ où C est la matrice "moyenne" trouvée ci-dessus, donne des résultats assez sensiblement différents de ceux d'Houllier, sans doute essentiellement à cause des pondérations (ici, certains essais se comportent pratiquement comme des individus supplémentaires, tellement leur poids dans l'analyse est faible).

Remarque. Ainsi, cet exemple justifie bien le choix de la métrique C^{-1} sur l'espace des individus, à partir des divers C_i. Cependant, dans d'autres cas, les C_i pourraient être bien plus dissemblables (même à un facteur réel près) et donc leur remplacement par α_i C discutable. Dans ces conditions, le modèle H est mal adapté à cause de l'hypothèse (iii). Il peut toujours être intéressant de faire l'ACP des individus Y_i, estimations respectives des y_i, mais les arguments développés pour le choix optimal de la métrique ne tiennent plus. En fait, il semble qu'il n'y ait pas de choix optimal global puisque celui-ci dépend des individus ; de ce point de vue, le fait même d'envisager une ACP pourrait être contestable. On peut dire, surtout en situation exploratoire, que la simplicité de l'ACP la rend encore compétitive et que bien des arguments "empiriques" la soutiennent (notons à ce sujet que l'Analyse des Correspondances tombe sous le coup de cette même difficulté puisqu'elle entre dans le cadre décrit ici : cf § 3, Remarque 1). On pourrait aussi recourir à une analyse de nature plus complexe, qui pourrait être à l'ACP ce qu'est le modèle de Goodman (1986) à l'Analyse des Correspondances, c'est-à-dire abandonner les méthodes "euclidiennes" (du second ordre) pour une "modélisation" plus sophistiquée.

5 - CHOIX DE LA DIMENSION

Lorsqu'on présente l'ACP de paramètres estimés d'un modèle dans le cadre développé ci-dessus, au moyen des approximations explicitées au paragraphe 4, on se trouve dans un cas particulier où la structure de covariance est complètement estimée, y compris la valeur de σ^2 qui est estimée par $\hat{\sigma}^2 = \alpha$. C'est une situation dans laquelle peuvent s'appliquer les propositions de Ferré (1989) pour le choix optimal de la dimension. Ainsi, on mettra en évidence le nombre de réels indépendants effectivement utiles à la modélisation.

L'utilisation de $\hat{f}_q$, estimateur de l'erreur quadratique moyenne dans l'estimation des vecteurs y_i (Ferré, 1989), conduit pour l'exemple ci-dessus ($\hat{\sigma}^2$ étant égal à .26, valeur très faible par rapport aux valeurs propres de l'ACP effectuée) aux résultats suivants :

$$\hat{f}_1 = 403789 ; \quad \hat{f}_2 = 132872 ; \quad \hat{f}_3 = 405 ; \quad \hat{f}_4 = 33 .$$

La dimension à retenir est $q^* = \arg\min \hat{f}_q$ et vaut 4 dans cet exemple. Il n'est donc pas envisageable de considérer une nouvelle modélisation à l'aide d'un nombre réduit de paramètres sans perdre quelque information. En particulier, bien que les deux premiers axes reproduisent respectivement 77.5% et 15% de l'inertie totale, il apparaît inacceptable, au vu des valeurs du critère, de retenir une solution uni ou bi-dimensionnelle ; on notera cependant que l'écart $\hat{f}_3 - \hat{f}_4$ est sans commune mesure avec $\hat{f}_2 - \hat{f}_4$.

Une étude identique, menée sur une population plus homogène extraite de l'exemple précédent (sur les seuls 9 essais de Tencin) conduit à une valeur minimale du critère pour la dimension 2. On a en effet :

$$\hat{f}_1 = 731 ; \quad \hat{f}_2 = 102 ; \quad \hat{f}_3 = 142 ; \quad \hat{f}_4 = 179 .$$

Ainsi, lorsque les seuls essais de Tencin sont pris en compte, une reparamétrisation du modèle à l'aide de seulement deux réels est alors souhaitable.

BIBLIOGRAPHIE

BAILLY A. (1986) : Modélisation de la croissance en circonférence du peuplier : premiers résultats. AFOCEL 1985, Annales de recherches sylvicoles, 359-394.

BERMAN A., PLEMMONS R.J. (1979) : Nonnegative matrices in the Mathematical Sciences. Academic Press, New York.

BESSE P., CAUSSINUS H., FERRE L., FINE J. (1986) : Some guidelines for Principal Component Analysis. In COMPSTAT 86, 23-30, Physica-Verlag Heidelberg for IASC.

BESSE P., CAUSSINUS H., FERRE L., FINE J. (1987) : Sur l'utilisation optimale de l'Analyse en Composantes Principales. C.R. Acad. Sc. Paris, t. 304, I,15, 459-462.

BESSE P., CAUSSINUS H., FERRE L., FINE J. (1988) : Principal Component Analysis and optimization of graphical displays. Statistics, 19,2,301-312.

CAUSSINUS H. (1986 a) : Models and uses of Principal Component Analysis. In Multidimensional Analysis, J. de Leeuw et al. eds, 149-170, DSWO Press, Leiden.

CAUSSINUS H. (1986 b) : Quelques réflexions sur la part des modèles probabilistes en Analyse des Données. In Data Analysis and Informatics, IV, E. Diday et coll. éd. 151-165, Nord-Holland, Amsterdam.

ESCOUFIER Y. (1980) : L'analyse conjointe de plusieurs matrices de données. Biométrie et Temps, E. Jolivet et al., éd.,59-76.

FERRE L. (1989) : Choix de la dimension optimale pour certains types d'analyses en composantes principales. C.R. Acad. Sc. Paris, t. 309, I, 959-964.

FERRE L. (1990) : A mean square error criterion to determine the number of components in Generalized Principal Component Analysis. Preprint, Laboratoire de Statistique et Probabilités, Toulouse.

GOODMAN L.A. (1986) : Some useful extensions of the usual Correspondence Analysis approach and the usual log-linear models approach in the analysis of contigency tables (with discussion). Intern. Statist. Rev.,54,3,243-309.

HOULLIER F. (1987) : Comparaison de courbes et de modèles de croissance ; choix d'une distance entre individus. Statistique et Analyse des Données,12,3,17-36.