REVUE DE STATISTIQUE APPLIQUÉE

P. BESSE
L. FERRE

Sur l'usage de la validation croisée en analyse en composantes principales

Revue de statistique appliquée, tome 41, n° 1 (1993), p. 71-76
<http://www.numdam.org/item?id=RSA_1993__41_1_71_0>

© Société française de statistique, 1993, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1993, **XLI** (1), 71-76

SUR L'USAGE DE LA VALIDATION CROISÉE EN ANALYSE EN COMPOSANTES PRINCIPALES

P. BESSE et L. FERRE

Laboratoire de Statistique et Probabilités, U.A. C.N.R.S. 745
Université Paul Sabatier, 31062 Toulouse Cedex France
E-Mail : besse@cix.cict.fr

RÉSUMÉ

De nombreux critères ont été proposés dans la littérature pour guider le choix de la dimension en Analyse en Composantes Principales (ACP). Un développement de Taylor du critère obtenu par Validation-Croisée (cf. Wold (1978), Eastment et Krzanowski (1982), Krzanowski (1983, 1987)), conséquence de la théorie des perturbations, montre, que malgré un coût en temps de calcul important, ce critère est équivalent à ceux prenant simplement en compte la part de variance expliquée.

Mots-clés : *Validation Croisée, Analyse en Composantes Principales, Choix de la dimension, Théorie des perturbations.*

SUMMARY

Many criteria have been proposed in literature in order to lead the choice of the number of dimensions in Principal Components Analysis. The perturbation theory enables to calculate a Taylor expansion of the criterion as performed as Cross-Validation. It shows that, in addition to being computationally costly, Cross-Validation offers little advantage over simpler methods which are based on the part of explained variance.

Key-words : *Cross-Validation, Principal Component Analysis, Dimension choice, Perturbation Theory.*

1. Introduction

L'Analyse en Composantes Principales recherche une approximation de la matrice initiale (n, p) des données par une matrice de rang inférieur q issue d'une décomposition en valeurs singulières. Elle fournit ainsi des représentations graphiques dans des sous-espaces de faible dimension. La question qui se pose alors, et qui a été largement débattue dans la littérature, concerne le choix du nombre de composantes ou de dimensions qui doivent être retenues. Différentes stratégies ont été proposées dont Jolliffe (1986) présente une revue. Certaines

sont des règles heuristiques considérant la part de variance expliquée ou le comportement de l'éboulis des valeurs propres (screegraph), d'autres proposent des tests, par exemple d'égalité des dernières valeurs propres, moyennant des hypothèses probabilistes classiques. Un dernier type d'approches, fondées sur des techniques de rééchantillonnage, ne nécessite pas d'hypothèses sur les distributions mais consomme beaucoup de temps de calcul.

La Validation Croisée a été introduite par Allen (1971) dans le but d'optimiser le choix des variables explicatives dans le modèle linéaire. Depuis Craven et Whaba (1979), elle est également communément utilisée en régression non-paramétrique pour optimiser le paramètre de lissage lors de modélisations par des fonctions splines et pour optimiser la taille de la fenêtre dans la méthode du noyau : Györfi *et al.* (1989). Ce sont Wold (1978), Eastment et Krzanowski (1982) puis Krzanowski (1983, 1987) qui ont proposé l'emploi de cette technique pour aider au choix de la dimension en ACP. Cet article se propose de montrer que, malgré un coût de calcul important, l'usage de la Validation Croisée en ACP n'apporte pas une règle de décision plus «objective» que les techniques usuelles heuristiques. Cette équivalence est obtenue par un développement de Taylor du critère de la validation croisée utilisant la théorie des perturbations.

2. Validation Croisée en ACP

L'ACP, qui peut être introduite de différentes façons, est ici présentée comme l'estimation d'un modèle à effet fixe (voir *e.g.* Caussinus (1986)). Soit $\{Y_i; i = 1, ..., n\}$ n variables indépendantes à valeurs dans R^p. On suppose que chacune des réalisations y_i observées se décompose sous la forme :

$$y_i = \mu + z_i + e_i$$

avec $\sum_{i=1}^{n} z_i = 0$ et où les e_i sont supposés *i.i.d.* avec $E(e_i) = 0$ et $\text{Var}(e_i) = \sigma^2 I_p$.

On pose alors pour hypothèse que les effets fixes z_i sont contenus dans un sous espaces F_q de dimension q inférieure à p ou, ce qui est équivalent, que la matrice $\mathbf{Z}(n \times p)$ centrée, qui admet les vecteurs z_i' comme lignes, est de rang q.

Si $\mathbf{Y}$, de ligne courante y_i', désigne la matrice $(n \times p)$ des observations, l'estimation par les moindres carrés des paramètres de ce modèle conduit à estimer μ par la moyenne empirique $\overline{\mathbf{y}}$ et à rechercher l'approximation de la matrice centrée $\mathbf{X}$ déduite de $\mathbf{Y}$ par une matrice $\widehat{\mathbf{Z}}_q$ de rang q inférieur à p :

$$\widehat{\mathbf{Z}}_q = \arg\min\{\text{tr}(\mathbf{X} - \mathbf{Z})'(\mathbf{X} - \mathbf{Z}) \mid \mathbf{Z}(n \times p),\ \text{rang}(\mathbf{Z}) = q\}.$$

La solution est donnée par la décomposition en valeurs singulières de X :

$$\widehat{\mathbf{Z}}_q = \sum_{k=1}^{q} \lambda_k^{1/2} t_k u_k',$$

où $\{u_k; k = 1, \ldots, p\}$ sont les vecteurs propres orthonormés de la matrice des covariances empiriques ($\mathbf{V} = \mathbf{X}'\mathbf{X}/n$) rangés dans l'ordre décroissant des valeurs propres λ_k. Les vecteurs $t_k = \lambda_k^{-1/2}\mathbf{X}u_k$ sont vecteurs propres de la matrice $\mathbf{W} = \mathbf{X}\mathbf{X}'/n$. En notant $\mathbf{U}_q$ la matrice de ces q premiers vecteurs propres, on obtient : $\widehat{\mathbf{Z}}_q = \mathbf{X}\mathbf{U}_q\mathbf{U}_q'$.

La mise en œuvre de la validation croisée consiste à minimiser la fonction Press de Allen qui évalue la qualité prédictive d'un «modèle» :

$$\text{Press} = \frac{1}{np}\sum_{i=1}^{n} \|\widehat{\mathbf{Z}}_{q^i}^{(i)} - z_i\|^2,$$

où $\widehat{\mathbf{Z}}_{q^i}^{(i)}$ est une estimation ou prédiction de z_i construite sans tenir compte de l'observation y_i et $\|.\|$ représente la norme usuelle dans R^p. Dans le cas de l'ACP, comme $\widehat{\mathbf{Z}}_{q^i}^{(i)}$ est obtenu par la projection de x_i sur le sous-espace engendré par les q premiers vecteurs propres, Wold (1978), Eastment et Krzanowski (1982) puis Krzanowski (1987) utilisent une adaptation de la fonction Press :

$$\text{Press}(q) = \frac{1}{np}\sum_{i=1}^{n} \|\widehat{\mathbf{Z}}_{qi}^{(i,.)} - z_i\|^2$$

où $\widehat{\mathbf{Z}}_{q^i}^{(i,.)}$ est le vecteur de coordonnées :

$$\widehat{\mathbf{Z}}_{q^ij}^{(i,j)} = \sum_{k=1}^{q} t_{ik}^{(j)}(\lambda_k^{(j)})^{1/4}(\lambda_k^{(i)})^{1/4}u_{jk}^{(i)};\ \begin{matrix} j = 1, \ldots, p \\ i = 1, \ldots, n \end{matrix}.$$

Les termes indicés par (i), respectivement (j), sont obtenus par décomposition en valeurs singulières de $\mathbf{X}$ privée de la $i^{\text{ème}}$ ligne, respectivement $j^{\text{ème}}$ colonne. Ce sont respectivement les éléments propres des matrices $\mathbf{V}^{(i)}$ et $\mathbf{W}^{(j)}$. Ainsi, l'estimation de z_{ij} par $\widehat{\mathbf{Z}}_{q^ij}^{(i,j)}$ n'utilise pas l'élément x_{ij}.

3. Développement par la théorie des perturbations

De la même façon que Critchley (1985) pour l'approximation d'une fonction d'influence, l'élimination d'une ligne (respectivement d'une colonne) dans le tableau des données peut être considérée comme une perturbation de la matrice $\mathbf{V}$ (resp. $\mathbf{W}$). En effet, la suppression de la $i^{\text{ème}}$ ligne (resp. $j^{\text{ème}}$ colonne) conduit à une matrice $\mathbf{V}^{(i)}$ (resp. $\mathbf{W}^{(j)}$) à diagonaliser qui s'écrit :

$$\mathbf{V}^{(i)} = \mathbf{V} + (\mathbf{V} - x_i x_i')/(n-1) - x_i x_i'/(n-1)^2 (\text{resp. } \mathbf{W}^{(j)} = \mathbf{W} - x^j x^{j\prime}/n).$$

Pour n suffisamment grand, il est possible d'utiliser la théorie des perturbations des opérateurs linéaires (Kato, 1966), pour obtenir les développements des

éléments propres de $\mathbf{V}^{(i)}$ (resp. $\mathbf{W}^{(j)}$) en fonction de ceux de $\mathbf{V}$ (resp. $\mathbf{W}$). Ces développements sont valides sous l'hypothèse :

$$(n-1)^{-1} < \inf(\lambda_k - \lambda_{k+1})/2K$$

où K est un majorant de la plus grande valeur des normes de $(\mathbf{V}^{(i)} - \mathbf{V})$ et $(\mathbf{W}^{(j)} - \mathbf{W})$, pour $i = 1, ..., n$ et $j = 1, ..., p$.

A l'ordre 2, on obtient pour $k = 1, ..., p$:

$$\lambda_k^{(i)} = \lambda_k + (n-1)^{-1}\lambda_k(1 - t_{ik}^2) + O((n-1)^{-2});$$
$$\lambda_k^{(i)1/4} = \lambda_k^{1/4} + (n-1)^{-1}\lambda_k^{1/4}(1 - t_{ik}^2)/4 + O((n-1)^{-2});$$
$$\lambda_k^{(j)} = \lambda_k - n^{-1}\lambda_k(u_{jk}^2) + O((n-1)^{-2});$$
$$\lambda_k^{(j)1/4} = \lambda_k^{1/4} - n^{-1}\lambda_k^{1/4}(u_{jk}^2)/4 + O((n-1)^{-2});$$
$$u_{jk}^{(i)} = u_{jk} + (n-1)^{-1}\lambda_k^{1/2}\sum_{l\neq k}^{p}\frac{\lambda_l^{1/2}}{(\lambda_l - \lambda_k)}t_{il}t_{ik}u_{jl} + O((n-1)^{-2});$$
$$t_{ik}^{(j)} = t_{ik} + n^{-1}\lambda_k^{1/2}\sum_{l\neq k}^{p}\frac{\lambda_l^{1/2}}{(\lambda_l - \lambda_k)}u_{jl}u_{jk}t_{il} + O((n-1)^{-2}).$$

On déduit de ces expressions [1] le développement à l'ordre 2 de $\widehat{\mathbf{Z}}_{q^{ij}}^{(i,j)}$:

$$\widehat{\mathbf{Z}}_{q^{ij}}^{(i,j)} = \sum_{k=1}^{q}\left\{\lambda_k^{1/2}t_{ik}u_{jk} + \frac{\lambda_k^{1/2}}{n-1}\left\{\left[(1-t_{ik}^2) - \frac{n-1}{n}u_{jk}^2\right]u_{jk}t_{jk}/4 + \right.\right.$$
$$\left.\left. + \sum_{l\neq k}^{p}\frac{\lambda_l^{1/2}\lambda_k^{1/2}}{(\lambda_l - \lambda_k)}t_{il}u_{jl}\left[t_{ik}^2 + \frac{n-1}{n}u_{jk}^2\right]\right\}\right\} + O((n-1)^{-2}).$$

De cette expression on tire le développement de la fonction Press à l'ordre 2 ci-dessous en remarquant que : $\sum_{j=1}^{p} u_{jl}u_{jk} = \frac{1}{n}\sum_{i=1}^{n} t_{il}t_{ik} = \begin{cases} 1 \text{ si } l = k, \\ 0 \text{ sinon.} \end{cases}$

$$\text{Press}(q) = (1/p)\sum_{k=q+1}^{p}\lambda_k -$$
$$-\frac{2}{np(n-1)}\left\{\sum_{l=1}^{q}\sum_{k=q+1}^{p}\frac{\lambda_l\lambda_k}{(\lambda_k - \lambda_l)}\left[\sum_{i=1}^{n}t_{il}^2t_{ik}^2 + (n-1)\sum_{j=1}^{p}u_{jl}^2u_{jk}^2\right]\right\}$$
$$+O((n-1)^{-2}).$$

[1] Nous remercions l'un des relecteurs pour la rectification apportée à ce calcul.

Cette expression, facilement calculable à partir des résultats de l'ACP, permet d'obtenir une valeur approchée de la fonction Press sans utilisation intensive d'un ordinateur et d'économiser ainsi du temps de calcul. Cependant, on notera que pour n grand, la fonction Press devient très voisine, au facteur $1/p$ près, de la somme des valeurs propres négligées : elle est alors décroissante en q et atteint son minimum pour $q = p$!

Dès lors, la fonction Press seule se révèle inefficace pour déterminer la dimension. Cela a conduit Wold (1978), Eastment et Krzanowski (1982) à considérer des fonctions (différentes) de Press pour sélectioner une dimension. Par exemple, ces deux derniers ont proposé de comparer la fonction $g(q) = [\text{Press}(q) - \text{Press}(q-1)]/\text{Press}(q)$ à des valeurs arbitraires. A partir des développements ci-dessus, on obtient aisément celui de cette fonction g à l'ordre 1 :

$$g(q) = \frac{\lambda_q}{\sum_{k=q+1}^{p} \lambda_k} + O(1/n).$$

Ainsi, pour n grand, la règle de décision se ramène à la comparaison de $\lambda_q / \sum_{k=q+1}^{p} \lambda_k$ à une valeur arbitraire, ce qui montre que la technique proposée, malgré une lourde mise en œuvre informatique, est semblable aux techniques de type «screegraph» ou «part de variance expliquée».

En conclusion, il semble que la technique de Validation Croisée, si efficace dans de nombreux domaines de la Statistique, se heurte ici aux spécificités de l'ACP : estimateurs biaisés, modèles emboîtés,... dès lors que l'on s'intéresse à l'estimation des y_i. Ces difficultés peuvent cependant être contournées, au prix d'une perte d'information, en considérant la stabilité des estimateurs des sous-espaces F_q (Daudin *et al.*, 1989, Besse, 1992) qui seuls sont consistants.

Annexe

Rappels sur la théorie des perturbations :

En utilisant les résultats de Kato (1966) et sous les conditions mentionnées dans la section 3, on obtient les développements suivants pour la $i^{ème}$ valeur propre simple $\lambda_i(\varepsilon)$ et le vecteur propre associé $u_i(\varepsilon)$ d'une matrice $V(\varepsilon)$ s'écrivant sous la forme $V(\varepsilon) = V + \varepsilon U + O(\varepsilon^2)$:

$$\lambda_i(\varepsilon) = \lambda_i + \varepsilon \text{tr}(P_i U) + O(\varepsilon^2)) \text{ et } u_i(\varepsilon) = u_i - \varepsilon S_i U u_i + O(\varepsilon^2)$$

où $P_i = u_i u_i'$ et $S_i = \sum_{l \neq i} u_l u_l' / (\lambda_l - \lambda_i)$.

Les expressions présentées en section 3 sont obtenues par application de ces résultats aux matrices $\mathbf{V}^{(i)}$ et $\mathbf{W}^{(j)}$ et en utilisant le fait que $t_k = \lambda_k^{-1/2} \mathbf{X} u_k$.

Références

ALLEN D.M. (1971). The prediction sum of squares as a criterion for selecting variables. Tech. Report n° 23, Dpt of Statistics, Univ. of Kentucky.

BESSE Ph. (1992). PCA Stability and Choice of Dimensionality. *Statistics & Probability Letters*, 13, 405-410.

CAUSSINUS H. (1986). Models and uses of principal components analysis. *Multidimensional Data Analysis*, J. de Leeuw et al. (eds.), DSWO Press, Leiden, 149-170.

CRAVEN P., WAHBA G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377-403.

CRITCHLEY F. (1985). Influence in principal components analysis, *Biometrika*, 72, 627-636.

DAUDIN J.J., DUBY C., TRÉCOURT P. (1989). P.C.A. stability studied by the bootstrap and the infinitesimal jackknife method. *Statistics*, 20, 255-270.

EASTMENT M.T., KRZANOWSKI W.J. (1982). Cross-validatory choice of the number of components from a principal components analysis. *Technometrics*, 24, 73-77.

GYÖRFI L., HÄRDLE W., SARDA P., VIEU P. (1989). Non parametric curve estimation from time series. *Lecture Note in Statistics*, 60, Springer Verlag, Heidelberg.

JOLLIFFE I.T. (1986). Principal component analysis, *Springer Verlag*, New-York.

KATO T . (1966). Perturbation theory for linear operators. *Springer Verlag*, New-York.

KRZANOWSKI W.J. (1983). Cross-validation choice in principal component analysis : some sampling results. *Journal of Statistical Computational Simulation*, 18.

KRZANOWSKI W.J. (1987). Cross-validation in principal component analysis. *Biometrics*, 43, 575-584.

WOLD S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20, 397-405.