STATISTIQUE ET ANALYSE DES DONNÉES

FRANK CRITCHLEY

Principal component analysis : some majorisation, perturbation and nonnegative matrix theory

Statistique et analyse des données, tome 13, n° 1 (1988), p. 8-14.

<http://www.numdam.org/item?id=SAD_1988__13_1_8_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 1 - p.8-14

PRINCIPAL COMPONENT ANALYSIS: SOME MAJORISATION, PERTURBATION AND NONNEGATIVE MATRIX THEORY

Frank CRITCHLEY

Department of Statistics
University of Warwick
Coventry CV4 7AL
England

Abstract: *Results from three branches of mathematics are drawn together in a study of principal component analysis. These results illumine the properties of the method and both explain and extend empirical findings with it. Finally, a new standardisation is briefly indicated.*

Résumé: *Quelques résultats de trois parties des mathématiques sont assemblés dans une étude de l'analyse en composantes principales. Ces résultats donnent de la lumière sur les propriétés de cette méthode et expliquent et étendent quelques phénomènes empiriques trouvés avec cette méthode. Finalement, on indique un nouveau choix de la normalisation des échelles des variables.*

Keywords: *eigenvalues; majorisation; nonnegative matrix; perturbation; principal component.*

STMA classification index: *06-070, 00-050.*

1. INTRODUCTION

We study principal component analysis from the viewpoint of three branches of mathematics. This provides new insights and properties of the method as well as both explaining and extending some well-known empirical results from using it. Finally, a new possibility is indicated for standardising the scales of the variables when they are positively correlated. This possibility will be explored further in a later paper.

The three branches of mathematics are majorisation (section 2), the perturbation theory of a simple eigenvalue (and its normalised eigenvector) of a real symmetric matrix (section 3) and nonnegative matrix theory (section 4).

Manuscrit reçu le 30.10.86, révisé le 4.1.88

2. MAJORISATION

Let H be a real, symmetric nonnegative definite $p \times p$ matrix. We think of H as being either the sample or the population covariance matrix of a set of p random variables. In principal component analysis, there is often interest in the k largest or the k smallest eigenvalues of H. We lose no generality in supposing that both the vector $\underline{\lambda} \equiv (\lambda_1, \ldots \lambda_p)^T$ of eigenvalues of H and the vector $\underline{\ell} \equiv (h_{11}, \ldots, h_{pp})^T$ of its diagonal elements lie in $D_+ = \{(x_1, x_2, \ldots, x_p)^T | x_1 \geq x_2 \geq \ldots \geq x_p \geq 0\}$. Then:

Theorem 1: (Schur (1923)): $\ell_1 + \ldots + \ell_p = \lambda_1 + \ldots + \lambda_p$ and $\forall k = 1, \ldots p-1$:

$$\ell_1 + \ldots + \ell_k \leq \lambda_1 + \ldots + \lambda_k; \text{ equivalently : } \ell_p + \ldots + \ell_{p-k+1} \geq \lambda_p + \ldots + \lambda_{p-k+1}.$$

We say $\underline{\ell}$ is majorised by $\underline{\lambda}$, written $\underline{\ell} \prec \underline{\lambda}$. This result actually holds for any Hermitian matrix as shown by Fan (1949) using extremal properties of $\lambda_1 + \ldots + \lambda_k$ and $\lambda_p + \ldots + \lambda_{p-k+1}$ given in Fan (1950). These twin sets of optimal properties have natural statistical interpretations in a principal component analysis, as noted in the case of the k largest eigenvalues by Rao (1973, p. 591). The importance of Theorem 1 was enhanced when Horn (1954) and Mirsky (1958) showed that no stronger ordering between $\underline{\ell}$ and $\underline{\lambda}$ is generally true. The excellent book by Marshall and Olkin (1979) expands the above remarks and gives an encyclopaedic account of the many inequalities flowing from $\underline{\ell} \prec \underline{\lambda}$. We note here some applications of interest for principal component analysis where the diagonal elements $\underline{\ell}$ have a natural interpretation and where H is known to be nonnegative definite.

It is well known empirically that if a single variable has a much higher variance than the rest, then a single eigenvalue will dominate a principal component analysis. Theorem 1 shows why this is so, generalises this result from one to any number k of dominant variables and establishes the natural converse that if k variances are negligible with respect to the remainder then the same is true of the eigenvalues. An example of the practical utility of this result is given in Critchley (1983).

Recall that a real-valued function ϕ defined on a subset A of $\mathbf{R}^n$ is said to be Schur-convex on A if $\underline{x} \prec \underline{y}$ on A implies $\phi(\underline{x}) \leq \phi(\underline{y})$. Thus it follows at once from Theorem 1 and consideration of the set D_+ to which $\underline{\ell}$ and $\underline{\lambda}$ here belong that:

Theorem 2: Let ϕ be a Schur-convex function on D_+. Then:

$$\phi(\underline{\ell}) \leq \phi(\underline{\lambda}) \tag{1}$$

If also ϕ is non-decreasing (non-increasing) in each argument, then the stronger result holds that:

$$\phi(\ell_1, \ldots, \ell_k) \leq \phi(\lambda_1, \ldots, \lambda_k), k = 1, \ldots p \tag{2}$$

$$\text{or:} \phi(\ell_{p-k+1}, \ldots, \ell_p) \leq \phi(\lambda_{p-k+1}, \ldots, \lambda_p), k = 1, \ldots p \tag{3}$$

respectively.

Thus the whole set of variances, the k largest variances and the k smallest variances give a variety of information about the corresponding subsets of the spectrum of eigenvalues of H. Examples of this are, in the same numbering as the above theorem:

(1a) $\phi(x_1,\ldots,x_p) = \Sigma_{i=1}^{p} g(x_i)$ with $g(\cdot)$ convex. In particular, the variance of the eigenvalues $\lambda_1,\ldots,\lambda_p$ is never less than the variance of the variances $\ell_1,\ldots,\ell_p$.
(1b) $\phi(x_1,\ldots,x_p) = \underline{b}^T\underline{x}$ with $b_1 \geq \ldots \geq b_p$. This provides a host of inequalities between the same ordered contrast of $\underline{\ell}$ and of $\underline{\lambda}$. For example, taking $b_1 = 1$ and $b_i = -1$ for all $i > 1$, we find that if the largest variance exceeds the sum of all the other variances then the same is true of the eigenvalues.
(2a) $\phi(x_1,\ldots,x_k) = \Sigma_{i=1}^{k} g(x_i)$ with $g(\cdot)$ convex and increasing on $x \geq 0$. In particular, if x^+ denotes $\max(x,0)$:

$$\Sigma_{i=1}^{k}(\ell_i - a)^+ \leq \Sigma_{i=1}^{k}(\lambda_i - a)^+, k = 1,\ldots,p, a \in R.$$

(2b) $\phi(x_1,\ldots,x_k) = \underline{b}^T\underline{x}$ with $b_1 \geq \ldots \geq b_k \geq 0$.
(2c) $\phi(x_1,\ldots,x_k)$ any symmetric gauge function (Marshall and Olkin, 1979, page 96). In particular,

$$(\ell_l^r + \ldots + \ell_k^r)^{\frac{1}{r}} \leq (\lambda_1^r + \ldots + \lambda_k^r)^{\frac{1}{r}}, k = 1,\ldots,p, r \geq 1.$$

(3a) $\phi(x_{p-k+1},\ldots,x_p) = \Sigma_{i=p-k+1}^{p} g(x_i)$ convex and decreasing on $x \geq 0$. In particular, for H positive definite:

$$\ell_{p-k+1}^{-1} + \ldots + \ell_p^{-1} \leq \lambda_{p-k+1}^{-1} + \ldots + \lambda_p^{-1}, k = 1,\ldots p$$

(3b) $\phi(x_{p-k+1},\ldots,x_p) = \underline{b}^T\underline{x}$ with $0 \geq b_{p-k+1} \geq \ldots \geq b_p$.

On account of the great difficulty of the mathematics of other cases, the distribution theory and inference methods for principal component analysis are almost entirely limited to the case where H is a covariance matrix. However, because of the effects of different scales of measurement of the p variables alluded to above and discussed further in the next section it is common to standardise H in some way. Note that when a correlation matrix is used, or more generally when all the variances are equal, Theorem 1 gives no information about $\underline{\lambda}$ in terms of $\underline{\ell}$.

In section 4 we consider the case where all the elements of H are positive. We have then the following result which is stronger than Theorem 1:
Theorem 3 (Berman and Plemmons, (1979), p. 97)

$$\ell_1 + \ldots + \ell_p = \lambda_1 + \ldots + \lambda_p, \ell_1 \leq \lambda_1 \text{ and for all } 1 \leq j < k \leq p:$$

$$(\ell_1 + \ldots + \ell_{j-1}) + \ell_{k-1} + \ell_k \leq (\lambda_1 + \ldots + \lambda_j) + \lambda_k.$$

3. PERTURBATION THEORY

Consider again the case where a small number of variances dominate the rest. The case of k negligible variances is analogous and is not discussed separately. Little is lost in assuming that the k dominant variables are not highly correlated with each other. Indeed, perhaps at the loss of some interpretability, these correlations are of course removable

entirely by a preliminary rotation in the corresponding k dimensional subspace and a possible subsequent reduction of k. Taking these correlations to be $0(\varepsilon)$, where ε denotes the order of the other $p-k$ standard deviations, and exploiting non-negative definiteness, we may regard H as a perturbed version of

$$B \equiv \text{ diag}(\ell_1, \ldots, \ell_k, 0, \ldots, 0) \equiv \Sigma_{i=1}^{k} \ell_i \underline{e}_i \underline{e}_i^T$$

where $\underline{e}_i$ denotes the i^{th} unit coordinate vector is $\mathbb{R}^p$. Specifically, we have:

$$H = B + \varepsilon C + \frac{1}{2}\varepsilon^2 D + 0(\varepsilon^3)$$

where $c_{ij} = 0$ for $i = j \leq k$ and for $\min(i,j) > k$
and: $d_{ij} = 0$ for $\min\ (i,j) \leq k$.

Under the condition that $\ell_1, \ldots, \ell_k$ are distinct, we may use standard perturbation theory for a simple eigenvalue of a real symmetric matrix. For each $i = 1, \ldots k$, assume that the corresponding perturbations of $\ell_i, \underline{e}_i$ are:

$$\ell_i(\varepsilon) = \ell_i + \varepsilon\mu_i + \frac{1}{2}\varepsilon^2 v_i + 0(\varepsilon^3)$$

and:

$$\underline{e}_i(\varepsilon) = \underline{e}_i + \varepsilon \underline{f}_i + \frac{1}{2}\varepsilon^2 \underline{g}_i + 0(\varepsilon^3).$$

Then applying Lemma 2.1 of Sibson (1979) we have at once:
Theorem 4: $\mu_i = 0, v_i = 2\{\Sigma_{1 \leq j \neq i \leq k} c_{ij}^2/(\ell_i - \ell_j) + \ell_i^{-1}\Sigma_{j>k} c_{ij}^2\}$

$$\text{and } \underline{f}_i = \begin{pmatrix} c_{i1}/(\ell_i - \ell_1) \\ \vdots \\ c_{i,i-1}/(\ell_i - \ell_{i-1}) \\ 0 \\ c_{i,i+1}/(\ell_i - \ell_{i+1}) \\ \vdots \\ c_{ik}/(\ell_i - \ell_k) \\ c_{i,k+1}/\ell_i \\ \vdots \\ c_{ip}/\ell_i \end{pmatrix}$$

Corollary 5: If the k dominant variables are uncorrelated with each other,

$$v_i = 2\ell_i^{-1}\{\Sigma_{j>k} c_{ij}^2\} \text{ and } \underline{f}_i = \ell_i^{-1}(0, \ldots 0, c_{i,k+1}, \ldots, c_{ip})^T.$$

In particular, the eigenvalue ℓ_i and the only non-zero component of $\underline{e}_i$ are unchanged to first order. In the case of a single dominant variable, and also when the dominant variables are uncorrelated with each other, the second order change in ℓ_i is non-negative, which refines the corresponding statement in Theorem 1. In this latter case, *all* components of $\underline{e}_i$ associated with the dominant variables are unchanged to first order.

These results then explain theoretically, refine and extend what is commonly observed in practice: dominant variances are associated with dominant eigenvalues and have corresponding eigenvectors close to the relevant unit coordinate vectors. For an example, see Tables 2.2 and 2.4 of Kendall (1975).

Theorem 4 also shows that if two or more of the dominant variances should happen to be close when the associated variables are not uncorrelated then the corresponding elements of the corresponding eigenvectors can be volatile. For example if $\ell_1 \approx \ell_2$ while $c_{12} \neq 0$ then the second component of $\underline{f}_1$ and the first components of $\underline{f}_2$ can be large as their denominator is in modulus $(\ell_1 - \ell_2)$. This is natural in the sense that in the limit $\ell_1 \to \ell_2$, the first two elements of corresponding eigenvectors are indeterminate. This same perturbation theory underlies the general warning against interpreting too closely the elements of eigenvectors corresponding to eigenvalues that are very close: such elements are highly sensitive to small changes in the date. The analogous effects of multicollinearity in regression where several eigenvalues are near zero are well documented. An extreme form occurs in the principal component context when analysing a non-diagonal correlation matrix in which the correlations vary little if at all, as in Example 8.3 of Morrison (1976). For some further discussion see also Critchley (1985).

4. POSITIVELY CORRELATED VARIABLES

It frequently happens that (nearly) all the correlations in a principal component analysis are positive. In such cases it is common empirical experience that the dominant eigenvector has all its elements positive. Kendall (1975, p.24-5) comments to this effect in a worked example. It is easy to see theoretically why this is so. Suppose now that $H > 0$ where we interpret matrix and vector inequalities such as this elementwise. The classical work of Perron (1907) and Frobenius (1908, 1909) using only this positivity property of H gives:

Theorem 6:

(i) The largest eigenvalue $\lambda \equiv \lambda_1$ of H is positive and simple

(ii) H has an eigenvector $\underline{q} > \underline{0}$ corresponding to λ.

By orthogonality of eigenvectors corresponding to distinct eigenvalues, we have:

Corollary 7: In a principal component analysis of a positive covariance matrix all components beyond the first strictly contrast the variables. That is, the associated eigenvectors have elements of both signs.

The minimum and maximum row sums of H provide bounds for λ and also for γ, the ratio of the maximal and minimal elements of the corresponding positive eigenvector $\underline{q}$. Denoting these extreme row sums by a and b respectively we have the standard results:

Theorem 8: (Berman and Plemmons, (1979), Theorem 2.35) (i) $a \leq \lambda \leq b$; (ii) $\gamma^2 \geq b/a$. Equality holds in (i) and (ii) if and only if $a = b$.

Part (ii) shows that it is sufficient for $\underline{q}$ to be far from $\underline{u} \equiv (1/\sqrt{p})(1, 1, \ldots, 1)^T$ that b/a be much greater than one while part (i) shows that $b/a = 1$ is necessary and sufficient for

$\underline{q} = \underline{u}$. This gives some theoretical insight into another empirical finding. With positively correlated data, stabilising the variance is likely to reduce b/a and so is often associated with $\underline{q}$ moving nearer $\underline{u}$. Comparing Examples 8.1 and 8.3 of Morrison (1976) illustrates this well.

The elements of H also provide bounds on the extent to which the Perron-Frobenius eigenvalues dominates the others. We have:

Theorem 9. Let $K = \max_{i,j,k,\ell}(h_{ij}h_{k\ell}h_{i\ell}^{-1}h_{kj}^{-1})^{\frac{1}{2}}$ and $m = \min_{i,j}(h_{ij})$. Then:

$$(\ell_1 + \ell_2 - b)^+/b \leq \lambda_2/\lambda_1 \leq (K-1)/(K+1) \leq (\ell_1 - m)/(\ell_1 + m).$$

Proof. The first inequality follows from Theorems 1 and 8(i). The second and third follow from the work of Hoph (1963) and Ostrowski (1963) on positive matrices and the fact that H is non negative definite.

The lower bound on λ_2/λ_1 may be trivial as $b > \ell_1$. The upper bounds are more interesting. They are increasing functions of K and ℓ_1/m respectively and demonstrate that it is sufficient that the elements of H do not vary greatly for λ_1 to strongly dominate λ_2, and hence all the other eigenvalues. For a correlation matrix they show that: $\lambda_2/\lambda_1 \leq (1-\rho)/(1+\rho)$ where ρ is the minimal correlation. Thus, for example, if all correlations exceed 0.5, λ_1 is certain to be at least three times bigger than λ_2, while if $\rho \geq 0.8, \lambda_1 \geq 9\lambda_2$.

One last comment on standardisation. Consider a re-scaling $H \rightarrow G = D^{-1}HD^{-1}$ where now D is a diagonal matrix with $d_{ii} > 0, i = 1, \ldots p$. A common choice is of course $d_{ii} = \sqrt{\ell_i}$, the correlation matrix G being advocated for use in principal component analysis in that constant diagonal elements puts the variables on an equal footing in a sense. Now it is intuitive, and can be shown formally, that D can be chosen so that the row sums of G take *any* preassigned positive values. This fact and Theorem 8 (ii) reinforce the familiar warning against trying to interpret the relative sizes of coefficients in the dominant eigenvector $\underline{q}$ in terms of the relative "importance" of the corresponding variables to the underlying "common factor". These relative sizes depend strongly on an arbitrary if conventional choice of scaling. Only the common-signedness of the elements of $\underline{q}$ has empirical support, reflecting in a natural way the positivity of the correlations among the variables. Indeed we may use the above fact to suggest another convention for standardisation: choose D so that G is doubly-stochastic. The variables are then all on the same footing in the sense that they contribute equally to the dominant eigenvector $\underline{u}$. In this case, the dominant eigenvalue is, of course, 1 and we note that, as is easily shown:

Theorem 10 (Berman and Plemmons, (1979), p.51): Any other eigenvalue of G is at most $\min\{1 - (\Sigma_j \min_i g_{ij}), (\Sigma_j \max_i g_{ij}) - 1\}$.

These ideas on standardisation are currently being pursued and we hope to publish the results shortly.

REFERENCES

[1] BERMAN, A. and PLEMMONS, R.J., (1979), Non-negative matrices in the mathematical sciences, London: Academic Press.

[2] CRITCHLEY, F., (1983), The Euclidean structure of a dendrogram, Warwick Statistics Research Report No. 48.

[3] CRITCHLEY, F., (1985), Influence in principal components analysis, Biometrika, Vol. 72, pp. 627–636.

[4] FAN, K., (1949) and (1950), On a theorem of Weyl concerning eigenvalues of linear transformations I and II, Proc. Nat. Acad. Sci. U.S.A., Vol. 35, pp. 652–655 and Vol. 36, pp.31-35.

[5] FROBENIUS, G., (1908) and (1909), Über Matrizen aus positiven Elementen I and II, S.-B. Preuss. Akad. Wiss. (Berlin), pp. 471–476 and 514–518.

[6] HOPH, E., (1963), An inequality for positive integral linear operators, J. Math. and Mech., Vol. 12, pp. 683-692.

[7] HORN, A., (1954), Doubly stochastic matrices and the diagonal of a rotation matrix, Amer. J. Math., Vol. 76, pp. 620–630.

[8] KENDALL, M.G. (1975), Multivariate analysis, London: Griffin.

[9] MARSHALL, A.W. and OLKIN, I., (1979), Inequalities: theory of majorization and its applications, London: Academic Press.

[10] MIRSKY, L., (1958), Matrices with prescribed characteristic roots and diagonal elements, J. London Math. Soc., Vol. 33, pp.14-21.

[11] MORRISON, D.F., (1976), Multivariate statistical methods, London: McGraw-Hill.

[12] OSTROWSKI, A.M. (1963), On positive matrices, Math. Annalen. Vol. 150, pp. 276–284.

[13] PERRON, O. (1907), Zur Theorie der Über Matrizen, Math. Ann., Vol. 64, pp. 248–263.

[14] RAO, C.R., (1973), Linear statistical inference and its applications, London: John Wiley.

[15] SCHUR, I., (1923), Über eine Klasse von Mittelbildungen mit Anwendungen die Determinanten, Theorie Sitzungsber. Berlin. Math. Gesellschaft, Vol. 22, pp.9–20.

[16] SIBSON, R. (1979), Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling, J. Royal Stat. Soc. B, Vol. 41, pp. 217–229.