Principal Component Analysis in the Presence of Group Structure

W. J. Krzanowski

Applied Statistics, Vol. 33, No. 2 (1984), 164-168.

Stable URL:
http://links.jstor.org/sici?sici=0035-9254%281984%2933%3A2%3C164%3APCAITP%3E2.0.CO%3B2-J

Applied Statistics is currently published by Royal Statistical Society.

Appl. Statist. (1984),
33, *No.* 2, *pp.* 164–168

Principal Component Analysis in the Presence of Group Structure

By W. J. KRZANOWSKI

University of Reading, UK

[Received September 1983. Revised December 1983]

SUMMARY

A nested series of hypotheses on dispersion structure is identified when observations are grouped in a multivariate sample. A simple method of estimation is suggested for one of these hypotheses, and results using this method are compared with those previously obtained by maximum likelihood methods. Using these hypotheses, an analogy may be drawn between comparison of principal components between groups and comparison of regressions between groups.

Keywords: Between-group analysis; Eigenvalues; Eigenvectors; Principal components

Suppose that observations are made of a p-variate random vector X in each of k populations. Many parametric techniques of multivariate analysis require normality assumptions, and a typical starting point for any analysis of such observations is that X has a $N_p(\mu_i, \Omega_i)$ distribution in population π_i $(i = 1, \ldots, k)$. Consider now a sample of size n_i taken from population π_i, and let $\bar{x}_i$, S_i denote the maximum likelihood (ML) estimates of μ_i, Ω_i.

A common hypothesis tested on such data is H_a: $\Omega_i = \Omega$ for all i. This can sometimes be a prelude to further analysis (e.g. canonical variate analysis, multivariate analysis of variance). The likelihood ratio test statistic for H_a is given (Mardia *et al.*, 1979, p. 140) by $-2 \log \lambda_a = \Sigma n_i \log |S_i^{-1} S|$, where $S = (\Sigma n_i S_i)/(\Sigma n_i)$ is the ML estimate of Ω. Under H_a, $-2 \log \lambda_a$ has an asymptotic χ^2 distribution with $\frac{1}{2} p(p+1)(k-1)$ degrees of freedom. An improved approximation to this distribution has been given by Box (1949), valid for each n_i as small as 20 provided that k and p are not large (< 6).

H_a may often have to be formally rejected when, nevertheless, it may be suspected that there is some similarity between the Ω_i which could be used to advantage, either in subsequent analysis or to improve the precision of other estimates. One possibility is that the Ω_i share common principal axes although the size of each axis, and its relative importance, may vary from population to population. This is equivalent to requiring the Ω_i to be simultaneously reducible to diagonal form by the same orthogonal matrix, i.e. H_b: $L^t \Omega_i L = \Lambda_i$ $(i = 1, \ldots, k)$, where L is an orthogonal $(p \times p)$ matrix and the Λ_i are all diagonal matrices. (Note, however, that there is no implication here that the rank order of elements of Λ_i is the same for each $i = 1, \ldots, k$.)

The relevance of H_b in practice may be seen more easily from the standpoint of principal component analysis. In this technique, the original vector of variables X is transformed to a new vector $Y = (Y_1, \ldots, Y_p)^t$. Each new variable is a linear function of the original variables, i.e. $Y_i = a_i^t X$, and the new variables are ranked according to their variances, i.e. var $(Y_1) \geqslant$ var $(Y_2) \geqslant \ldots \geqslant$ var (Y_p). Y_i is the ith principal component of the system, the elements of the vector a_i are the coefficients of the ith principal component, and the values of

Present address: Department of Applied Statistics, University of Reading, Whiteknights, Reading, Berks. RG6 2AN, UK.

 0035–9254/84/33164 $2.00

$Y_1, \ldots, Y_p$ corresponding to any individual in a sample are the principal component scores for that individual.

In practice, principal component analysis is often used to identify major sources of variation between individuals in a sample and to effect data reduction by indicating any minor, discardable, sources of variation. Interpretation of the weighted averages $a_i^t X$ in terms meaningful to the experimenter will additionally enable labels to be attached to these sources of variation, and may give further insight into the situation under study. Interest in the present note centres on the case where the same vector of measurements is made on individuals in a number of different groups. The most general situation is one in which the sources of variation between individuals, as well as their associated interpretations, may differ arbitrarily from group to group. In hypothesis H_b, on the other hand, it is postulated that the sources of variation are the same from group to group, but may be ranked differently in each group and may assume different levels of importance in each group. This may be a perfectly reasonable assertion to test in many practical situations. For example, if the same set of examinations is taken by each student in a number of different schools or colleges (as in the data set discussed by Krzanowski, 1979), it may be sensible to assume that the basic sources of variation between students are similar in each college, but that differences in teaching practice between colleges will place different emphases on these sources. Thus a college which specializes in teaching numerical skills may reduce the variability between students in arithmetic relative to the other colleges at the expense perhaps of the variability between students in other directions, such as language construction, say. Arithmetic and language construction may nevertheless be present as major sources of variability in all colleges. As further examples, one may reasonably postulate similar sources of variability between people's attitude to work, but differing in importance between the various socio-economic groupings; or between subjects' responses to psychological stimuli, with differing levels of importance between adults and children. Such differences in ranking may occur particularly among the less important components. If a principal component study can be deemed to be compatible with H_b, then clearly better estimation of parameters can be achieved under this hypothesis and a more concise data description effected.

Hypothesis H_b was first proposed and investigated by Flury (1983a, b), who has given a full analysis, including the estimation of L and Λ_i by maximum likelihood and the testing of H_b by likelihood ratio methods. Given ML estimates $\hat{L}$ and $\hat{\Lambda}_i$, the L–R statistic for H_b is $-2 \log \lambda_b = \Sigma\, n_i \log |\, S_i^{-1}\, \hat{\Omega}_i \,|$ where $\hat{\Omega}_i = \hat{L} \hat{\Lambda}_i \hat{L}^t$. To obtain $\hat{L}$ and $\hat{\Lambda}_i$, however, a complicated iterative numerical algorithm is required. The purpose of this brief note is to show that simple estimates of these quantities are readily obtainable by use of standard computer package facilities for principal component analysis. Moreover, the reasonableness or otherwise of hypothesis H_b can be assessed informally by a simple method. Finally, an analogy can be drawn between the comparison of principal components using this analysis and familiar methods for comparison of regression lines.

If H_b is true, then $L^t\, \Omega_i\, L = \Lambda_i$ (diag) for $i = 1, \ldots, k$. Thus

$$\sum_{i=1}^{k} (L^t \Omega_i L) = \sum_{i=1}^{k} \Lambda_i = \Lambda \text{ (diag)}, \quad \text{i.e. } L^t \Psi L = \Lambda \text{ where } \Psi = \Omega_1 + \ldots + \Omega_k.$$

Hence the columns of L are the eigenvectors (i.e. principal component coefficients) of Ψ corresponding to the eigenvalues given by the diagonal elements of Λ. A simple estimate $\tilde{L}$ of L can thus be obtained from a principal component analysis of $T = S_1 + \ldots + S_k$. Let the jth column of $\tilde{L}$ be written $\tilde{l}_j$. Then on setting $\tilde{\lambda}_{ij} = \tilde{l}_j^t\, S_i\, \tilde{l}_j$, a simple estimate of Λ_i is given by $\tilde{\Lambda}_i = \text{diag}\,(\lambda_{i1}, \ldots, \lambda_{ip})$. Use of $\tilde{\Omega}_i = \tilde{L}\, \tilde{\Lambda}_i\, \tilde{L}^t$ in place of $\hat{\Omega}_i$ in $-2 \log \lambda_b$ will provide an approximation to the true L–R statistic, for testing H_b. It is surmised that the approximation will be very close, and this is illustrated below. However, there is an alternative way of informally assessing the reasonableness of H_b. This can be used if not all n_i are equal, and

is based on the following argument. If H_b is true, then $L^t \Omega_i L = \Lambda_i$ (diag) for $i = 1, \ldots, k$ so that $L^t(n_i/N)\Omega_i L = (n_i/N)\Lambda_i$ for $i = 1, \ldots, k$ where $N = \Sigma n_i$. Thus

$$\sum_{i=1}^{k} [L^t(n_i/N)\Omega_i L] = \sum_{i=1}^{k} (n_i/N)\Lambda_i = \Lambda_0 \text{ (diag)}$$

and the columns of L are therefore given by the eigenvectors of $\Psi_0 = (n_1\Omega_1 + \ldots + n_k\Omega_k)/N$ corresponding to the eigenvalues given by diagonal elements of Λ_0. Now Ψ_0 is estimated by $(n_1S_1 + \ldots n_kS_k)/N$, which is just the pooled estimate S defined earlier. Thus *if H_b is true*, principal component analyses of both S and T should yield (as component coefficients) estimates of the same quantities L. A comparison of the two sets of coefficients (either visually or by using the method of Krzanowski, 1979) should therefore show up the reasonableness of H_b. Similar sets of coefficients indicate that H_b is tenable, whereas very different sets indicate that it is not. If all n_i are equal, however, this method is not appropriate as $S = (1/k)T$ and the coefficients are bound to agree. A referee has pointed out that if H_b is true, then *all* weighted means of the individual S_i would have the same eigenvectors. This could be used to obtain an alternative informal test, or even to proceed more formally to some sort of union-intersection test, but will not be pursued further here.

Flury (1983b) applied his maximum-likelihood algorithm to four different data sets extracted from the multivariate literature: Iris Data (Fisher, 1936), Painted Turtle Data (Jolicoeur and Mosimann, 1960), North American Marten Data (Jolicoeur, 1963) and Real and Forged Bank Note Data (Flury and Riedwyl, 1983). His maximum likelihood estimates $\hat{L}$ for these data sets are reproduced in Table 1, and compared with them are the simple estimates $\tilde{L}$ obtained from a principal component analysis of T. For clarity of presentation, the coefficients are multiplied by 100 to yield integer values and only the differences between the simple estimates and the ML estimates are given. This shows up important differences more clearly. As can be seen, agreement between the two methods is remarkably close, only the Iris data exhibiting more than trivial differences. Qualitatively they can all be said to give identical results. The L–R statistic values $-2 \log \lambda_b$ given by Flury (1983b), and approximations to these obtained from the simple estimates as described above are given in Table 2. Again there is very close agreement between exact and approximate values, with no conflict in the conclusions that would be reached from the two sets of results. Note, however, that the approximate value is greater than the exact one in each case. It can be shown that this will always be so.

Only one of the data sets (American Marten) had groups of appreciably differing size, so an additional principal component analysis of S was worthwhile only for this data set. Results are also displayed in Table 1, and they can be seen to be very similar to those from principal component analysis of T. This suggests that H_b is a reasonable hypothesis, which is supported by the L–R statistic value of 8.34 on 6 degrees of freedom from Table 2. Note, incidentally, that doubt is cast on the appropriateness of H_b in all the other data sets considered.

Finally, it can be seen that the three hypotheses, H_a: $\Omega_i = \Omega$ all i, H_b: $L^t \Omega_i L = \Lambda_i$ (diag) and H_c: Ω_i all distinct, provide a nested system (Flury, 1983a). This can be likened to the situation often encountered in regression analysis. Suppose that observations are taken on a series of dependent and regressor variables in a number of different groups or populations. A common mode of analysis is to fit separate regression equations in each group, and then to fit a series of constrained regressions in which the constraints imposed become progressively more severe. By this means it is possible to test hypotheses about parallelism or coincidence of regression lines across groups. If we now have multivariate observations split up into groups, and we wish to investigate the principal component structure across the groups, we can use the methods described earlier to do so. First we can obtain separate principal component analyses of each S_i, then a principal component analysis of T and finally one of S. Testing H_b tests the hypothesis that the k dispersion structures have a common underlying set of principal components. This can be likened to the case of parallel regression lines. Testing H_a then tests the further hypothesis of coincident dispersion structures.

TABLE 1

Estimated principal component coefficients under the common component hypothesis. M = Maximum likelihood estimates; T = components of T; S = components of S

Data set	*Component*	*Method*	*Coefficients* (× 100)			
A. (Iris)	1	M	74	25	60	18
$p = 4, k = 3$		T-M	0	7	−3	−2
$n_1 = n_2 = n_3 = 50$	2	M	−65	47	50	33
		T-M	2	−29	8	14
	3	M	−16	−83	52	6
		T-M	−22	−4	−6	−21
	4	M	11	−16	−33	92
		T-M	12	−16	−2	−7
B. (Painted Turtles)	1	M	64	49	59	
$p = 3, k = 2$		T-M	0	0	0	
$n_1 = n_2 = 24$	2	M	−66	74	11	
		T-M	1	1	−3	
	3	M	−38	−46	80	
		T-M	−2	2	0	
C. (American Marten)	1	M	73	−14	−66	9
$p = 4, k = 2$		T-M	0	3	−1	−1
$n_1 = 92, n_2 = 47$		S-M	−1	−12	4	7
	2	M	39	57	39	61
		T-M	−1	0	1	0
		S-M	2	1	0	−2
	3	M	49	−58	63	−19
		T-M	2	2	−1	−1
		S-M	−2	10	5	−11
	4	M	−28	−57	−8	77
		T-M	4	−3	0	−1
		S-M	−2	−4	18	−3
D. (Bank Notes)	1	M	77	−63	−9	−9
$p = 4, k = 2$		T-M	−1	0	0	0
$n_1 = 100, n_2 = 85$	2	M	31	54	−51	−59
		T-M	−2	−3	−1	−4
	3	M	4	3	78	−63
		T-M	−1	−2	0	1
	4	M	56	56	35	50
		T-M	2	2	−2	−4

TABLE 2

Comparison of exact and approximate L-R statistics

Data set	*Degrees of freedom*	*Exact L-R statistic*	*Approximation*
Iris	12	63.91	88.38
Painted Turtles	3	7.93	8.31
American Marten	6	8.34	9.39
Bank Notes	6	12.04	13.08

References

Box, G. E. P. (1949) A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317–346.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

Flury, B. N. (1983a) Some relations between the comparison of covariance matrices and principal component analysis. *Computat. Statist. and Data Anal.*, **1**, 97–109.

——— (1983b) A generalisation of principal component analysis to k groups. Technical Report 83-14, Department of Statistics, Purdue University.

Flury, B. N. and Riedwyl, H. (1983) *Angewandte Multivariate Statistik*. Stuttgart and New York: Verlag Gustav Fischer.

Jolicoeur, P. (1963) The degree of generality of robustness in *Martes Americana*. *Growth*, **27**, 1–17.

Jolicoeur, P. and Mosimann, J. E. (1960) Size and shape variation in the painted turtle: a principal component analysis. *Growth*, **24**, 339–354.

Krzanowski, W. J. (1979) Between-group comparison of principal components. *J. Amer. Statist. Ass.*, **74**, 703–707. Corrigenda in *J. Amer. Statist. Ass.*, **76**, 1022.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.