COENOSES 2(1): 19-23, 1987

AN EXPERIMENTAL EVALUATION OF THE EFFECT OF DATA CENTERING, DATA STANDARDIZATION, AND OUTLYING OBSERVATIONS ON PRINCIPAL COMPONENT ANALYSIS

S.G. Hinch and K.M. Somers* Department of Zoology, University of Western Ontario, London, Ontario, Canada N6A 5B7

Keywords: Experiment, PCA, Centering, Standardization, Outliers, Correlation, Community

*Present address: Department of Zoology, University of Toronto, Toronto, Ontario, Canada M5S 1A1

Abstract: We used simulated data to evaluate the influence of data centering, data standardization, and outliers on R-type principal component analysis (PCA). First principal components from centered analyses extracted increasing amounts of variance as the correlation structure of the raw data increased, while non-centered analyses derived only one principal axis which was independent of data correlation. Data standardization resulted in a loss of information by equating variances among all variables. The percentage of the total variance incorporated in the first component was always slightly greater from centered analyses on covariance matrices, than on correlation matrices, and was independent of data correlation. Outliers showed greatest influence on PCA when the data were weakly correlated. Outliers affected the results from both types of resemblance matrices equally.

Introduction

Principal component analysis (PCA) is an ordination technique that summarizes patterns of covariation between variables. It is widely used in population and community research. Geometrically, PCA is a rotation of the original coordinate system to new orthogonal axes coinciding with directions of maximum linear variation in the data (Orlóci 1978; Pimentel 1979). These axes (principal components) are linear combinations of the variables, mathematically derived by an eigenanalysis of a resemblance matrix based on the original data. The first component summarizes the largest portion of the total variance. Successive components account for decreasing proportions while remaining linearly uncorrelated with previous components (Stauffer et al. 1985). Components may be computed on the basis of resemblance between relevés (Q-PCA) or resemblance between variables (R-PCA) (Orlóci 1967; Feoli 1977). Mathematically, these procedures are duals (Orlóci 1966, 1967; Gower 1967).

PCA can be modified in several ways. Data may be centered, whereby each data element is expressed as a deviation from the corresponding row and/or column mean. The axes are thus rescaled; their origin moved to the center of gravity of the data swarm. A great majority of ecological ordinations use centered data since ordinations of non-centered data produce a trivial «general component» which is often difficult to interpret (Orlóci 1966). However, Noy-Meir (1973) contends that non-centered ordinations of certain types of data produce interesting results and contain information which is lost by centered ordinations.

Data may be standardized, which is a method of weighting information contained in the raw data usually by a parameter of the distribution of each variable such as the standard deviation (Noy-Meir *et al.* 1975; Orlóci 1966, 1975, 1978). The method of standardization is critical since ordinations of the same data with different standardizations can produce strikingly different results (Noy-Meir 1971). This study uses simulated data to review the influence of data centering and data standardization on the results of PCA.

Multivariate outliers, those data points failing to maintain the general pattern of the variables (Campbell 1980) and consequently unduly influencing correlations (Tabachnick and Fidell 1983), can have pronounced effects on multivariate analyses (Gnanadesikan 1977, Harner and Whitmore 1981). PCA can be especially sensitive to outlying observations (Gnanadesikan 1977, Devlin *et al.* 1981). A second purpose of this study, therefore, is to use simulated data to examine the effects of a multivariate outlier on the results of PCA.

Methods

Data Simulation

For simplicity, the simulated data matrices were uniformly correlated, such that all inter-variable correlations approximate the grand mean correlation (see Morrison 1976). Three, 10-variable by 30-observation matrices were generated with grand mean correlations of 0.9, 0.5 and 0.2 (R.H. Green, Dept. of Zoology, University of Western Ontario, pers. comm.). All 10 variables were simulated with a mean of 10.0; the first 5 with a standard deviation of 1.0, the second 5 with a standard deviation of 1.5. Different standard deviations were used to demonstrate the effects of variation within variables on the results of PCA.

A randomly selected observation was deleted from each of the 3 simulated matrices and replaced by an outlying

observation. Outliers were derived by multiplying the standard deviation of each variable by a constant and adding this value to the mean of that variable. To examine the effects of increasing outlier weight on PCA performance, 5 matrices with a single outlier were created with standard deviations multiplied by 1, 2, 4, 6 and 8 respectively. In total, 15 10-variable by 30-observation matrices were created.

Analyses

To examine the influence of data centering and standardization on the results of PCA, each of the 3 uniformly correlated matrices was analyzed with an R-type PCA (Ryan *et al.* 1981). During each analysis, the raw data matrices were summarized in the following fashion: (a) non-centered and non-standardized; the resemblance matrix being an analogue of the covariance matrix, (b) non-centered and standardized; the resemblance matrix being an analogue of the correlation matrix, (c) centered (by variable) and non-standardized; the resemblance matrix being the covariance matrix, (d) centered (by variable) and standardized; the resemblance matrix being the correlation matrix.

To examine the influence of an outlier, centered R-type PCA (Orlóci 1978) was performed on each 10-by-30 matrix containing single outliers of different weights. Both the covariance and the correlation matrices were used in these analyses to determine whether both resemblance matrices were equally sensitive to outliers.

Results

Data Centering and Standardization

A comparison of centered and non-centered PCA indicated that eigenvalues of the first component from non-centered analyses were, as expected, nearly 2 orders of magnitude greater than their centered counterparts (Table 1). Eigenvalues from successive components were not as extreme. In centered PCA, the first axis maximizes the sum of squared deviations, thus the eigenvalue is proportional to the variance of that axis. Therefore comparisons of axis variation from centered analyses can be made simply by comparing eigenvalues. However, in non-centered PCA, the first axis joins the zero origin to the centroid of the data swarm. Consequently, the sum of squares of the observations along this axis is not proportional to the variance. Eigenvalues from non-centered analyses merely reflect the distance between the data swarm and the origin (Fig.1). In terms of eigenvalue size, centered and non-centered PCA are not comparable.

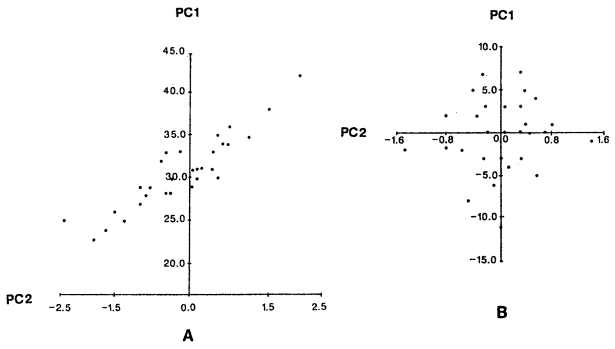

Fig. 1. Projections of scores derived from R-type PCA of the covariance matrix from non-centered raw data (A) and centered raw data (B) in two-dimensional component space. Raw data matrices have a grand mean correlation of 0.9.

As the grand mean correlation of the raw data increased, the first component from centered PCA extracted more of the total variance, while variation extracted by subsequent components decreased (Table 2). This contrasts with non-centered PCA which derived only one major axis for each data set, independent of the correlation structure (i.e. the grand mean correlation).

The amount of variation summarized by the first several eigenvalues from centered PCA of covariance matrices was always somewhat greater than from correlation matrices. Thus, some information was lost (presumably in the form of variation within variables) as a result of data

Table 1. Eigenvalues of the first four components derived from R-type centered and non-centered PCA on covariance and correlation matrices created from simulated multivariate data with uniform correlation.

Grand mean correlation	Raw data modification	Similarity matrix	PC1	PC2	PC3	PC4
0.9	centered	covariance	18.3	0.4	0.2	0.2
0.9	centered	correlation	9.4	0.2	0.1	0.1
0.9	non-centered	covariance	1054.7	1.0	0.3	0.2
0.9	non-centered	correlation	665.4	0.5	0.1	0.1
0.5	centered	covariance	7.9	1.8	1.3	1.1
0.5	centered	correlation	5.1	1.1	0.9	0.7
0.5	non-centered	covariance	1007.2	1.8	1.4	1.3
0.5	non-centered	correlation	825.5	1.1	1.0	0.7
0.2	centered	covariance	7.4	3.3	2.9	2.2
0.2	centered	correlation	3.6	1.4	1.2	1.1
0.2	non-centered	covariance	1013.8	3.3	3.0	2.2
0.2	non-centered	correlation	660.1	1.4	1.2	1.1

Grand mean correlation	Similarity matrix	PC1	PC2	PC3	PC4
0.9	covariance	93.8	1.9	1.1	1.0
0.9	correlation	93.7	1.6	1.0	1.0
0.5	covariance	53.0	11.8	8.7	7.3
0.5	correlation	50.7	11.2	9.2	6.7
0.2	covariance	36.2	16.2	14.1	10.7
0.2	correlation	35.8	14.3	12.3	10.6

Table 2. Percent of the total variance explained by the first four components derived from R-type centered PCA using non-standardized (e.g. covariance) and standardized (e.g. correlation) similarity matrices created from simulated multivariate data with uniform correlation.

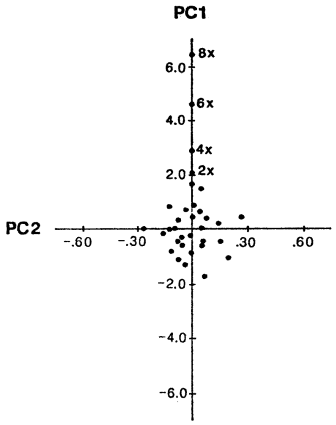

Fig. 2. Projections of scores derived from R-type PCA of the covariance matrix of centered data with overlays of the outlying observations of differing weight. Points labelled with 2X, 4X, 6X and 8X are simulated outliers described in Table 3.

Table 3. Eigenvalues with the percent of the total variance for the first four components derived from R-type centered PCA on simulated data with outliers. The grand mean correlation (G.M.C.) and similarity matrix type (S.M.) are listed. Outlier position refers to the weight of the outlying multivariate observation (i.e. none = no outlier, 2X = two times the standard deviation of each variable added to the mean of that variable, 4X = four times etc.).

			PC1		PC2		PC3		PC4	
G.M.C.	S.M.	outlier position	eigen value	% tot. var.	eigen value	% tot. var.	eigen value	% tot. var.	eigen value	% tot. var.
0.9	cov.	none	18.262	93.8	0.373	1.9	0.208	1.1	0.193	1.0
0.9	cov.	2X	19.983	94.6	0.368	1.7	0.201	1.0	0.182	0.9
0.9	cov.	4X	27.659	96.0	0.369	1.3	0.201	0.7	0.182	0.6
0.9	cov.	6X	40.500	97.2	0.369	1.0	0.201	0.5	0.173	0.4
0.9	cov.	8X	58.560	98.1	0.369	0.6	0.201	0.3	0.162	0.3
0.9	corr.	none	9.373	93.7	0.164	1.6	0.101	1.0	0.100	1.0
0.9	corr.	2X	9.448	94.5	0.151	1.5	0.091	0.9	0.087	0.9
0.9	corr.	4X	9.595	95.9	0.111	1.1	0.067	0.7	0.064	0.6
0.9	corr.	6X	9.720	97.2	0.077	0.8	0.046	0.5	0.043	0.4
0.9	corr.	8X	9.804	98.0	0.054	0.5	0.032	0.3	0.031	0.3
0.5	cov.	none	7.884	53.0	1.757	11.8	1.295	8.7	1.081	7.3
0.5	cov.	2X	9.181	57.4	1.688	10.6	1.276	8.0	1.069	6.7
0.5	cov.	4X	15.025	68.7	1.694	7.7	1.277	5.8	1.078	4.9
0.5	cov.	6X	24.849	78.4	1.697	5.4	1.277	4.0	1.083	3.4
0.5	cov.	8X	38.649	84.9	1.699	3.7	1.277	2.8	1.083	2.4
0.5	corr.	none	5.070	50.7	1.116	11.2	0.915	9.2	0.671	6.7
0.5	corr.	2X	5.486	54.9	1.063	10.6	0.858	8.6	0.608	6.1
0.5	corr.	4X	6.699	67.0	0.772	7.7	0.629	6.3	0.447	4.5
0.5	corr.	6X	7.725	77.3	0.529	5.3	0.434	4.3	0.310	3.1
0.5	corr.	8X	8.416	84.2	0.367	3.7	0.302	3.0	0.217	2.2
0.2	cov.	none	7.361	36.2	3.300	16.2	2.859	14.1	2.171	10.7
0.2	cov.	2X	10.010	44.0	3.219	14.1	2.836	12.5	2.174	9.6
0.2	cov.	4X	18.088	58.6	3.220	10.4	2.865	9.3	2.177	7.0
0.2	cov.	6X	31.630	71.1	3.220	7.2	2.877	6.5	2.179	4.9
0.2	cov.	8X	50.883	79.7	3.219	5.0	2.878	4.5	2.180	3.4
0.2	corr.	none	3.584	35.8	1.431	14.3	1.225	12.3	1.058	10.6
0.2	corr.	2X	4.357	43.6	1.288	12.9	1.048	10.5	0.945	9.5
0.2	corr.	4X	5.819	58.2	0.949	9.5	0.774	7.7	0.701	7.0
0.2	corr.	6X	7.087	70.9	0.659	6.6	0.539	5.4	0.488	4.9
0.2	corr.	8X	7.952	79.5	0.446	4.5	0.385	3.8	0.346	3.5

standardization. This pattern was independent of the grand mean correlation.

Outliers

Only centered PCA was used to compare eigenvalues (component variation) between analyses of resemblance matrices containing outliers with different weights (see Fig. 2). As outlier influence increased, the percentage of the total variance explained by the first component also increased (Table 3). This pattern was independent of the type of similarity matrix used. Correspondingly, information was lost from the second component, presumably being transferred to the first. This was most evident from analyses of the data matrices with 0.5 and 0.2 grand mean correlations where some information from the third and fourth components was also transferred to the first component. Outliers had less effect when the original data were highly correlated (i.e. grand mean correlation of 0.9 versus 0.2).

In general, first-component eigenvalues derived from covariance matrices were more sensitive to outlying observations, whereas those derived from correlation matrices were more robust, resisting change as the weight of the outlier increased. However, the percentage of total variance summarized by first-component eigenvalues was less affected by the type of resemblance matrix.

Discussion

Each non-centered PCA yielded one major component regardless of the data-matrix correlation. In contrast, centered PCA extracted a single major component, but the total amount of variance incorporated by this component depended on the grand mean correlation. Non-centered PCA will derive only one major component when the initial raw data have homogeneous correlation (Noy-Meir 1971, 1973; Pielou 1984). The length of this axis is determined by the distance from the origin to the data swarm, thus, non-centered PCA derives eigenvalues which do not represent maximal orthogonal variation. Non centered PCA will always be less efficient than centered PCA in summarizing variable configurations (Orlóci 1966, 1967); in this sense the usefulness of non-centered PCA for ordinations must be questioned. Feoli (1977) noted that although non-centered PCA does not intercept axes of maximal variation, it does intercept clusters in decreasing order of importance according to cluster size and homogeneity. Feoli suggests that non-centered PCA may be appropriate when the initial data is heterogeneous and cluster seeking is desired. Consequently, non-centered PCA is best used for data exploration to ascertain discrete patterns of observations, particularly when no *a priori* group structure is proposed (see Noy-Meir 1971, 1973). Alternatively, centered PCA summarizes patterns of covariation between variables, assuming that homogeneity (i.e. only a single group of observations) exists. Heterogeneous patterns among observations will confound the results of centered PCA (Thorpe 1976; Pimentel 1979; Shea 1985).

Data standardization in PCA, via the correlation matrix, may be selected if the analyst does not want variables with large variances to dominate the analysis, or if independence of measurement scale is desired. However, some information, in the form of variation within variables, is lost as a result of standardization. Alternatively, the covariance matrix maintains the original variation of each variable, which will vary according to measurement scale. Reyment *et al.* (1984) suggest that the correlation matrix is useful for exploratory work, but emphasize that the distribution theory based on the covariance matrix can, only in a few cases, be applied to the correlation matrix. Consequently, tests of significance and confidence intervals for eigenvalues are usually not available for the correlation matrix (see Morrison 1976 for further details).

In general, the effects of an outlier on PCA depends on its position relative to the ellipsoidal swarm of points in variable space. The major effect is to shift and stretch the ellipsoid in the direction of the outlier (Harner and Whitmore 1981). Where an outlier is situated along the major axis of variation, the ellipsoid is elongated disproportionately and the first component accounts for greater portions of the total variance than if the outlier was absent. Alternatively, if an outlier is orthogonal to the major axis of variation, the ellipsoid is contracted and the second component extracts a larger portion of the total variance. Outliers in this study were simulated to fall along the major axis of variation (Fig. 2) and thus, the percent of the total variance summarized by the first principal component increased with increasing outlier weight while the second component showed opposite trends.

Data correlation structure also affects the influence of an outlier. Atypical observations have less effect on variation summarized by the first component when data exhibit a uniform, strong correlation pattern. Single outliers in weakly correlated data have greater influence on the results of PCA.

Since the correlation matrix standardizes variation, we initially suspected that the influence of an outlier would be reduced in analyses using this matrix. Such was not the case, however, as outliers affected analyses using either the correlation or the covariance matrix in a similar manner. Therefore, to reduce the effects of outliers, we suggest careful scrutiny of the data (*e.g.* Gnanadesikan and Kettenring 1972; Rohlf 1975), although this is often difficult especially in high-dimensional space (Devlin *et al.* 1981; Tabachnick and Fidell 1983). Alternatively, if covariance estimation is required and the probability of outliers is high, the use of a robust or resistant resemblance matrix is advocated (e.g. Campbell 1980; Harner and Whitmore 1981).

A major problem with PCA, and one not specifically addressed in this study, is that PCA assumes a linear data structure, relying on this linear condition for independence between subsequent components (Orlóci 1979). Quite often, especially with community data, intervariable relationships are non-linear. Therefore, before choosing between centered or non-centered analysis, and between

the covariance or correlation matrix, one must establish whether the data pattern is linear. Where non-linear patterns are evident, and data transformations are ineffective, alternate procedures such as multidimensional scaling or detrended correspondence analysis may be appropriate (*e.g.* Fasham 1977; Hill and Gauch 1980; Orlóci *et al.* 1984).

Acknowledgment. We would like to thank Prof. L. Orlóci for his advice and encouragement. Financial support was generously provided by Prof. R. H. Green through an NSERC Operating Grant.

REFERENCES

CAMPBELL, N. A. 1980. Robust procedures in multivariate analysis. I. Robust covariance estimation. Appl. Statist. 29: 231-237.

DEVLIN, S. J., R. GNANADESIKAN and J. R. KETTENRING. 1981. Robust estimation of dispersion matrices and principal components. J. Am. Statist. Ass. 76: 354-362.

FASHAM, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology 58: 551-561.

FEOLI, E. 1977. On the resolving power of principal component analysis in plant community ordination. Vegetatio 33: 119-125.

GNANADESIKAN, R. 1977. *Methods for statistical data analysis of multivariate observations.* John Wiley and Sons. New York. 311 pp.

GNANADESIKAN, R., and J. R. KETTENRING. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28: 81-124.

GOWER, J. C. 1967. Multivariate analysis and multivariate geometry. The Statistician 17: 13-28.

HARNER, E. J. and R. C. WHITMORE. 1981. Robust principal component and discriminant analysis of two grassland bird species habitats. In: D. E. CAPEN [ed.], *The use of multivariate statistics in studies of wildlife habitat*, pp. 209-221. U.S.D.A. Forest Service General Technical Report, RM-87.

HILL, M. O., and H. G. GAUCH. 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42: 47-58.

MORRISON, D. F. 1976. *Multivariate statistical methods.* Second edition. McGraw-Hill Book Co., New York. 415 pp.

NOY-MEIR, I. 1971. Multivariate analysis of the semi-arid vegetation in south-eastern Australia: Nodal ordination by component analysis. Proc. Ecol. Soc. Aust. 6: 159-193.

NOY-MEIR, I. 1973. Data transformations in ecological ordination. I. Some advantages of non-centering. J. Ecol. 61: 329-341.

NOY-MEIR, I., D. WALKER and W. T. WILLIAMS. 1975. Data transformations in ecological ordination. II. On the meaning of data standardization. J. Ecol. 63: 779-800.

ORLÓCI, L. 1966. Geometric models in ecology I. The theory and application of some ordination methods. J. Ecol. 54: 193-215.

ORLÓCI, L. 1967. Data centering: a review and evaluation with reference to component analysis. Syst. Zool. 16: 208-212.

ORLÓCI, L. 1975. Measurement of redundancy in species collections. Vegetatio 31:65-67.

ORLÓCI, L. 1978. *Multivariate analysis in vegetation research.* 2nd ed. Dr. W. Junk B. V. Publishers. The Hague. 451 pp.

ORLÓCI, L. 1979. Non-linear data structures and their description. In: L. ORLÓCI, C. R. RAO and W. M. STITELER [eds], *Multivariate methods in ecological work*, pp. 191-202. International Co-operative Publishing House, Fairland, Maryland.

ORLÓCI, L., N. C. KENKEL, and P. H. FEWSTER. 1984. Probing simulated vegetation data for complex trends by linear and nonlinear ordination methods. Abstracta Botanica 8: 163-172.

PIELOU, E. C. 1984. *The interpretation of ecological data. A primer on classification and ordination.* John Wiley and Sons, New York. 263 pp.

PIMENTEL, R. A. 1979. *Morphometrics: The multivariate analysis of biological data.* Kendall/Hunt Publishing Co. Iowa. 276 pp.

REYMENT, R. A., R. E. BLACKITH and N. A. CAMPBELL. 1984. *Multivariate morphometrics.* 2nd ed. Academic Press, London. 233 pp.

ROHLF, F. J. 1975. Generalization of the gap test for the detection of multivariate outliers. Biometrics 31: 93-101.

RYAN, T. A. Jr., B. L. JOINER and B. F. RYAN. 1981. *Minitab reference manual.* Duxbury Press, Boston. 154 pp.

SHEA, B. T. 1985. Bivariate and multivariate growth allometry: Statistical and biological considerations. J. Zool., Lond. 206: 367-390.

STAUFFER, D. F., E. O. GARTON and R. K. STEINHORST. 1985. A comparison of principal components from real and random data. Ecology. 66: 1693-1698.

THORPE, R. S. 1976. Biometric analysis of geographic variation and racial affinities. Biol. Rev. 51: 407-452.

TABACHNICK, B. G. and L. S. FIDELL. 1983. *Using multivariate statistics.* Harper and Row Publishers, New York.

Manuscript received: July 1986