COENOSES 9(3): 155-158



# SWEEP-OUT COMPONENT ANALYSIS AS AN ORDINATION MODEL: AN ALTERNATIVE TO PRINCIPAL COMPONENT ANALYSIS

**M. Uddin[1], M. Atiqullah[2] and S. Shahid Shaukat[3]**

1 Department of Statistics, University of Karachi, Karachi - 75270
2 Institute of Statistical Science, Block - 5, Gulshan-e-Iqbal, Karachi - 75300
3 Department of Botany, University of Karachi, Karachi - 75270, Pakistan

**Keywords:** Principal component analysis, Sweep-out component analysis, Varimax rotation.

**Abstract.** The Sweep-Out Component analysis (SCA) is suggested as a method to develop environmental ordination. Comparison of SCA ordination (pertaining to environmental data set) and the unrotated PCA and PCA with varimax rotated ordinations revealed superiority of SCA over PCA in terms of greater mechanical validity, greater correlations with the original variables, greater parsimony (higher explained variance by the first three components) and a better correlation of the components with the corresponding vegetational ordination axes. The characteristics and the utility of the SCA model in ecological context are discussed.

## Introduction

A variety of formal and informal ordination methods have been developed in the last four decades and these have often been reviewed (Orlóci, 1978; Whittaker & Gauch, 1982; Greig-Smith, 1983; Jolliffe, 1986, Shaukat & Uddin, 1989). Among the ordination techniques that are frequently used, principal component analysis (PCA) has received considerable attention in ecological studies (Orlóci, 1966; Bouxin, 1975; Carleton, 1980, Miyata, 1983). Despite certain inherent weaknesses (Gauch *et al,* 1977; Clymo, 1980, del Moral, 1980) it has been shown (Feoli, 1977; Nichols, 1977) that PCA can provide unique, objective and parsimonious representations that are predictable and ecologically meaningful. The problems of non-linearity and discontinuity are largely circumvented when PCA or variants of factor analysis (FA) are employed for the purpose of constructing environmental ordinations (Shaukat & Uddin, 1989). This paper proposes 'sweep-out component analysis' (SCA) as an ordination method with particular reference to an environmental data set. The effectiveness of the sweep-out component analysis is tested against the more popular PCA technique.

## Methods

### *The sweep-out component analysis*

The proposed method is derived from 'sweep-out' estimation procedure described by Atiqullah (1968) and Atiqullah & Uddin (1993). Given a set of p-correlated variables ($X_1$, $X_2$,...,$X_p$), SCA derives p-uncorrelated variables ($Y_1$, $Y_2$,..,$Y_p$) by performing sweep-out operation on the dispersion matrix **S** of the data set **X**. The components of variation are extracted directly from the derived matrix **T** which is obtained as follows:

i) The first row of **S** is multiplied by the inverse of the first element $s_{11}$ and taking the resulting row containing unity as a first pivotal row.

ii) From every other non-pivotal row of **S** subtract a row obtained by multiplying the pivotal row by the first element of non-pivotal rows so that the first column consists of zeros except unity in the pivotal position

iii) Repeat steps (i) and (ii) on the successively reduced matrices. This would finally result in an upper trianqular matrix **T** of the form:

$$\mathbf{T} = \begin{bmatrix} 1 & t_{12} & t_{13} & \cdots & t_{1p} \\ & 1 & t_{23} & \cdots & t_{2p} \\ & & 1 & \cdots & t_{3p} \\ & & & & \vdots \\ & & & & t_{pp} \end{bmatrix}$$

The pivotal divisors of the matrix are:

$$\delta_1^2 = s_{11} \; ; \quad \delta_2^2 = s_{22} - \frac{s_{12}s_{21}}{s_{11}} \; ; \quad \text{etc.}$$

and the coefficients t's are:

$$t_{ij} = s_{ij} \,/\, \delta_1^2$$

$$t_{2j} = \left( s_{2j} - \frac{s_{21}s_{1j}}{s_{11}} \right) / \, \delta_2^2 \; ; \quad j = 1,2,....,p$$

etc.

where $\delta_1^2$, $\delta_2^2$,...., $\delta_p^2$ are the pivotal divisors in the sweep-out operation of rows performed on **S**. The pivotal divisors determine the variances of the derived variables $y_j$ and the summation of **X** variables is related to the weighted summation of y-variables by the equation:

$$\sum_{i=1}^{p} x_j = \sum_{j=1}^{p} W_j y_j$$

where $W_j$ represents the jth row sum of the matrix **T**. The sum of variances and covariances contained in matrix **S** equals the sum of the weighted squares variation in the derived variables, as follows:

$$\sum_{i=1}^{p} \sum_{j=1}^{p} s_{ij} = \sum_{j=1}^{p} W_j^2 \delta_j^2$$

The $W_j^2\delta_j^2$ are the jth sweep-out components and the proportion of the variation explained by jth component is obtained as

$$W_j^2 \delta_j^2 / \sum W_j^2 \delta_j^2$$

Among p! sets of $(X_1, X_2, \ldots, X_p)$, each set generating a matrix **T**, select the combination which yields the largest sweep-out components. In practice, the computational effort is considerably reduced by rearranging the successive pairs of variables and testing for the rank order of the components. The program SOCA performs the computations automatically.

*The data set and its characteristics*

The SCA and PCA ordinations were performed using the environmental data gathered from 22 stands in Gadap area, Southern Sind, Pakistan (Shaukat *et al.*, 1980). This data set consists of 11 soil variables as follows: 1, soil depth (cm); 2, soil pH; 3, organic matter (%); 4, $CaCO_3$; 5, exchangeable sodium (ppm); 6, exchangeable potassium (ppm); 7, maximum water holding capacity (%); 8, coarse sand (%); 9, fine sand (%); 10, silt (%); and 11, clay (%). Environmental variables were used because these are monotonic (James, 1971) and the data matrix does not contain excessive zero entries. As opposed to vegetation variable (species or other structural attributes), the environmental variables, to a great extent overcome the problem of non-linearity inherent in PCA or other linear models. The variables were suitably transformed (Shaukat & Uddin, 1989). The corresponding vegetation data of the 22 stands were used to correlate the environmental axes (gradients) derived from FA and PCA. This data set was restricted to the importance value index (Curtis & McIntosh 1951) of 17 well-represented species to avoid the problem of exececssive zero values (Austin 1976). Furthermore, the data set was standardized to standard scores (Noy-Meir *et al.*, 1975) for use in PCA.

## Results and discussion

Fig. 1 shows two-dimensional environmental ordinations based on SCA, PCA and PCA with varimax rotation. The ordinations show continuity in soil characteristics and discrete groups cannot be recognized. Despite the difference in the shape of the two configurations, there is an overall similarity in the ordinations. The correlation coefficient $r(D_1,D_2)$ between the corresponding stand distances of SCA and PCA in the three dimensions ordination space was found to be 0.7230 and the Euclidean distance $(D_1,D_2)$ was 335.82 indicating close similarity between the ordinations. Table 1 sets out the percentage of total variance explained by the first three components of SCA, PCA and PCA with varimax rotation. The highest proportion of total variance was explained by the first component of SCA (74.45%) followed by PCA (37.51%). The cumulative explained variance for the first

**Table 1. Percentage of total variance explained by the first three components of ordinations based on SCA and PCA.**

| Ordination method | Percentage of total variance explained | | | |
|---|---|---|---|---|
| | Comp. 1 | Comp. 2 | Comp. 3 | Cumulative |
| Sweep-out component analysis | 74.45 | 12.23 | 6.37 | 93.05 |
| Principal component analysis | 37.51 | 18.84 | 15.60 | 71.75 |
| Principal component analysis (varimax rotation) | 23.63 | 35.53 | 20.79 | 79.95 |

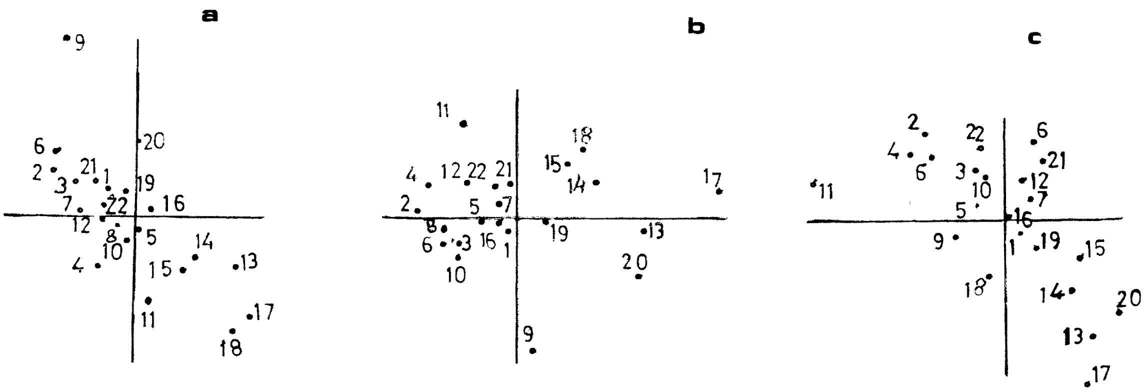


**Figure 1. Two-dimensional environmental ordination of 22 stands derived from (a) Sweep-out component analysis, (b) principal component analysis, and (c) principal component analysis with varimax rotation.**

**Table 2. Correlation coefficients between the eleven environmental (soil) variables with the first three components of SCA, PCA and PCA with varimax rotation.**

| Ordination method | Variables | Components of the environmental ordinations | | |
|---|---|---|---|---|
| | | Component 1 | Component 2 | Component 3 |
| Sweep-out component analysis | Soil depth | 0.3897 | -0.2154 | 0.3273 |
| | Soil pH | 0.7882 | 0.4877 | 0.3948 |
| | Organic matter | 0.1353 | -0.2937 | 0.1462 |
| | $CaCO_3$ | 0.7570 | -0.5566 | 0.4716 |
| | Exchangeable Na | 0.0187 | -0.0201 | 0.0374 |
| | Exchangeable K | 0.9999 | -0.7909 | 0.7008 |
| | Max. water holding capacity | 0.6437 | -0.3646 | 0.9620 |
| | Coarse sand (%) | 0.2750 | 0.3696 | 0.2899 |
| | Fine sand (%) | -0.0389 | -0.5364 | -0.2694 |
| | Silt (%) | 0.2986 | 0.0892 | 0.1010 |
| | Clay (%) | 0.0849 | 0.4025 | 0.5889 |
| Principal component analysis | Soil depth | 0.3133 | -0.1120 | 0.3748 |
| | Soil pH | 0.8699 | 0.3794 | -0.9066 |
| | Organic matter | -0.1100 | 0.3405 | 0.4720 |
| | $CaCO_3$ | 0.7974 | 0.4585 | -0.1776 |
| | Exchangeable Na | 0.7767 | 0.9423 | 0.0609 |
| | Exchangeable K | 0.8049 | 0.2996 | -0.1733 |
| | Max. water holding capacity | 0.7055 | -0.6324 | -0.0801 |
| | Coarse sand (%) | -0.4631 | 0.4797 | -0.6893 |
| | Fine sand (%) | 0.0084 | 0.2374 | 0.8869 |
| | Silt (%) | 0.6783 | -0.2931 | 0.0521 |
| | Clay (%) | 0.4476 | -0.7512 | 0.1242 |
| Principal component analysis with varimax rotation | Soil depth | 0.2255 | -0.5161 | -0.4752 |
| | Soil pH | 0.4221 | -0.7676 | -0.6487 |
| | Organic matter | 0.1142 | 0.0599 | -0.2169 |
| | $CaCO_3$ | 0.3035 | -0.6968 | -0.6072 |
| | Exchangeable Na | 0.3956 | -0.7931 | -0.9379 |
| | Exchangeable K | 0.3997 | -0.7755 | -0.5054 |
| | Max. water holding capacity | 0.5815 | -0.6929 | -0.1635 |
| | Coarse sand (%) | -0.9336 | 0.4789 | 0.1101 |
| | Fine sand (%) | 0.5706 | -0.0369 | -0.2133 |
| | Silt (%) | 0.4206 | -0.6049 | -0.0478 |
| | Clay (%) | 0.9875 | -0.5068 | 0.0571 |

three axes was also highest for SCA (93.05%) followed by PCA with varimax rotation (79.95%). To determine the mechanical validity of the ordinations, correlation coefficients were computed between the distance matrix of the environmental (soil) data set and the distance matrices of the ordination confiqurations r(D,D*). For SCA the value of r(D, D*) was 0.921 while that for PCA (unrotated) was 0.776 and for PCA with varimax rotation was 0.888. Thus, the SCA ordination shows the highest mechanical validity.

**Table 3. Correlation coefficients between the first three components of PCA vegetational ordination and the first three components of SCA and PCA environmental ordinations.**

| Ordination method | Component of environmental ordination | Components of vegetational ordinations | | |
|---|---|---|---|---|
| | | Component 1 | Component 2 | Component 3 |
| Sweep-out component analysis | 1 | 0.5723 | 0.1001 | 0.2327 |
| | 2 | -0.2668 | -0.1003 | -0.2355 |
| | 3 | 0.3153 | 0.2918 | 0.0482 |
| Principal component analysis | 1 | 0.6202 | -0.0428 | 0.2353 |
| | 2 | 0.0926 | -0.0779 | 0.2091 |
| | 3 | -0.2119 | 0.2595 | 0.4265 |
| Principal component analysis with varimax rotation | 1 | 0.2468 | -0.0331 | 0.2961 |
| | 2 | -0.5593 | -0.1279 | -0.2051 |
| | 3 | -0.4592 | -0.1781 | -0.1765 |

Correlation coefficients between the 11 environmental variables and the first three components of SCA, PCA and PCA with varimax rotation ordinations are given in Table 2. The first component of SCA and PCA ordinations showed highest correlation with exchangeable K and maximum water holding capacity and the low correlations with fine sand and organic matter while PCA with varimax rotation showed slightly different results showing highest correlation with coarse sand but again showed lowest correlation with organic matter. With respect to the second component, correlations of most of the environmental variables were negative in both SCA and PCA varimax rotated ordinations.

Correlation coefficients between the first 3 components of SCA, PCA and PCA (varimax rotation) environmental ordinations with the corresponding PCA vegetational ordinations are presented in Table 3. The first and the third components of SCA and PCA environmental ordinations yielded high correlations with the first component of PCA vegetational ordination while in case of PCA rotated solution second and third components showed high correlations with the first component of PCA vegetational ordination. The second component of PCA vegetational ordination in general, exhibited low (non-significant) correlations with the components of all the three environmental ordinations.

Both PCA and SCA are based on linear models. The performance of PCA ordinations is known to be affected by the curvilinearities inherent in vegetation (species abundance) data sets because of non-linear, presumably Gaussian response of species along environmental gradients and the non-linear change of sample similarity with increasing distance between samples (Gauch *et al.*, 1977; Digby & Kempton, 1987). Similar drawback is apparently associated with SCA. On the other hand, the environmental data sets comprise continuous uninterrupted variables that are mostly linearly correlated. Thus the problems of non-linearity and discontinuity are largely circumvented when environmental data set is subjected to either PCA or SCA.

The superiority of SCA over PCA was depicted in many respects: a) The first component and the first three components cumulatively explained remarkably greater percentage of total variance than either unrotated or varimax rotated PCA. b) The SCA ordination exhibited greater mechanical validity than did PCA ordinations. c) Higher levels of correlations obtained for individual variables with the first components of SCA ordination compared with PCA ordinations. d) Greater correlation exhibited between the SCA environmental ordination axes and the vegetational ordination axes. The principal reason for these results in that the greater proportion of total variation in the data set is channelised into first few components of SCA over that of PCA, i.e., SCA gives rise to more parsimonious ordination.

In view of the above findings it is suggested that SCA provides a preferable alternative model to PCA for the ordination of environmental or other linear ecological data sets. The method can also be used for vegetational ordination if the data set comes from a narrow ecological gradient.

## References


Atiqullah, M. 1968. On estimation by the sweep-out method. Biometrika 55: 305-311.

Atiqullah, M. and M. Uddin. 1993. Sweep-out components analysis. J. Appl. Statist. Sci., 1: 67-79.

Austin, M. P. 1976. On non-linear species response models in ordination. Vegetatio 33: 33-41.

Bouxin, G. 1975. Ordination and classification in the savanna vegetation of Akagera park (Rwanda, Central, Africa). Vegetatio 29: 155-167.

Carleton, T. J. 1980. Non-centred component analysis of vegetation data: a comparison of orthogonal and oblique rotation. Vegetatio 92: 59-66.

Clymo, R. S. 1980. Preliminary survey of the peat-bog Hummel Knowe Moss using various numerical methods. Vegetatio 42: 129-198.

Curtis, J. T. and R. P. McIntosh. 1951. An upland forest continuum in the prairie-forest border region Wisconsin. Ecology 32: 476-496.

Digby, P.G.N. and R.A. Kempton. 1987. Multivariate analysis of ecological communities. Chapman & Hall, London

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

Gauch, H. G. 1982. Multivariate Analysis in Community Ecology. Cambridge University Press, London, New York.

Gauch, H. G., R. H. Whittaker and T. R. Wentworth. 1977. A comparative study of reciprocal averaging and other ordination techniques. J. Ecol. 65: 157-174.

Greig-Smith, P. 1983. Quantitative Plant Ecology, 3rd ed. University of California Press, Berkeley, Los Angeles.

James, F. C. 1971. Ordinations of habital relationships among breeding birds. Wilson Bull. 83: 215-236.

Jolliffe, I. T. 1986. Principal Component Analysis. Springer, New York.

Miyata, I. 1983. Influence of vegetation structure of the tree layer on development of the herb layer in a secondary forest. Jap. J. Ecol. 33: 71-18.

Moral, R. del. 1980. On selecting indirect ordination methods. Vegetatio 42: 75-84.

Nichols, S. 1977. On the interpretation of principal component analysis in ecological contexts. Vegetatio 34: 191-197.

Noy-Meir, I., D. Walker and W. T. Williams. 1975. Data transformations in ecological ordinations. 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. 1978. Multivariate Analysis in Vegetation Research. 2nd ed. Dr. W. Junk, The Hague.

Shaukat, S. S., M. A. Khairi, D. Khan and J. A. Qureshi. 1980. Multivariate approaches to the analysis of the vegetation of Gadap area, Southern Sind, Pakistan. Trop. Ecol. 21: 81-102.

Shaukat, S. S. and M. Uddin. 1989. A comparison of principal component and factor analysis as ordination model with reference to a desert ecosystem. Coenoses 4: 15-28.

Whittaker, R. H. and H. G. Gauch. 1978. Evaluation of ordination techniques. In: R. H. Whittaker (ed.) Ordination of Communities. Dr. W. Junk, The Hague.


*Manuscript received: June 1994*