TRIBHUVAN UNIVERSITY JOURNAL
VOL. XVI, JUNE 1993

Centering of data in Principal Component Analysis in Ecological Ordination*

[1] Khadga Basnet

ABSTRACT

In this paper the original data matrix is transformed in ecological ordination to make it more uniform. Centering is a form of transformation of a data matrix. An appropriate transformation depends upon the types of floral or faunal variation in the data. This short article reviews (1) non-centering of data, (2) various forms of centering of data, and (3) their effects on the results of ordination and interpretation when applied to data. there are advantages and disadvantages in both centering and non-centering of data. There is no universal rule that tells one to be used in the analysis. However, non-centered ordination has some important advantages in vegetation studies; (1) it allows an assessment of the number and sharpness of discontinuities in the sample, (2) in discontinuous data, it minimizes the interference between the variation on the two sides of the break. The final decision, however, depends upon the aims of user and the nature of data.

INTRODUCTION

Original data matrix is transformed in ecological ordination to make it more uniform. It may be called as 'grinding of data'. Centering is a form of transformation of a data matrix. This means, the scores are expressed as deviations from the mean of the variable (Noy-Meir 1973, Pielou 1984).

Austin and Greig-Smith (1968) studied various transformations on ordinations and classifications of vegetation of a rain forest. They showed that the appropriate transformation depended on the types of vegetational variation in the data. Conclusions in the same line were drawn in various studies (Orloci 1967,

---

* This paper is the extension of my project work and discussion in quantitative ecology at Rutgers University. Comments of P.J. Morin, J.M. Faceli, and anonymous reviewers were helpful in improving the paper.

1 *Dr. Basnet is a Lecturer in Zoology, TU, Patan Campus, Lalitpur.*

1975, Noy-Meir 1973 , Feoli 1977). For this reason, it is reasonable to state that there is no universal rule as such intransformation of data. If the data have to be transformed, an appropriate transformation has to be selected depending upon the type of data.

The objective of this short article is to review centering and uncentering of data matrix in ecological ordination. This is important because more and more studies have been using quantitative approaches to unfold complicated ecological problems (e.g. Basnet 1992, 1993). The following are the themes of the review: (1) non-centering of data, (2) various forms of centering of data, (3) their effects on the results of ordination and interpretation when applied to vegetation data.

NON-CENTERING OF DATA

Original raw data are used in Principal Component Analysis (PCA) in which the new axes coincides with the intersection of the old axes. That is to say the origin of axes does not change (Pielou 1984). The sum of the squares is the sum of squared distance of the points from the original data. Therefore, it has no relation with the variance. In PCA, all the non-centered transformation forms have their first component unipolar having only positive or negative values for all the sites and species in one of the disjoint groups or series, but zero values for the other series (Noy-Meir 1973). This is the main contrast with the centered principal component analysis (Pielou 1984). It is regarded as 'general component' of a sub-matrix assuming that the first unipolar component as uninformative (Orloci 1966, Gower 1966). The second or third component is a similar unipolar 'series' component for the second sub-matrix. All other components are bipolar 'intra-series-axes' representing the continuous internal variation within the series, independently of that in the other series.

CENTERING OF DATA

When the observed measurements on each species have been transformed to deviations from the respective species means, the data is said to be transformed. There are various forms of centering as following: (See Noy-Meir 1973)

a. Centering by species means ($Xik - \bar{X}k$)

b. Centering by site means ($Xik - \bar{X}i$)

c. Double centering by site means ($Xik - \bar{X}k - \bar{X}i + \bar{\bar{X}}$);

Where, i = site index, k = species index, X = species in site score. Centering and normalization are not discussed here.

Most ecological ordinations by using principal component analysis or related methods centering by species (e.g. Groenewoud 1965), Orloci 1966, Yarranton 1967). Centering by species mean transfers the reference point to a hypothetical 'average stand' and in order to get the information, stands have to deviate from this point. Thus a relative difference in composition between sites are obtained.

Centering by site mean changes the reference point to an 'average species'. It implies an interest in species only as far as their distribution differs from that of the total vegetation. Also it implies an interest in stands only as far as their composition varies from equal proportion of all species.

In a double centering, a species contributes to the analysis only to the extent that its variance differs from the variance of the total vegetation.

Centered PCA producing bipolar components explores the maximal lines of variation in a set of multivariate data (Carleton and Maycock 1980, Noy-Meir 1973, Pielou 1984). It involves the specification of the origin, which is the point of reference of the multivariate model. In ecological sense, this point is utilized for the description of vegetation. It is a 'null' point and bears no information. For an useful information of the data analyzed there should be a deviation from this point.

ADVANTAGE OF CENTERING

Centering is necessary for variables of the 'interval-scale' type on which the zero point is arbitrary like in pH, temperature etc. But the usual measurements taken in the vegetation data are ratio-scale (Noy-Meir 1973) like the relative density, relative basal area, presence or absence of a species in a site. This means that centering is not inherently necessitated by the nature of the vegetation data.

The main advantage of centering, however, is a more efficient concentration of information about the ordination. This is important particularly for the graphic display and interpretation (Noy-Meir 1973). A centered PCA is more useful in some of the following cases:

1. When the data exhibit little or no between axes heterogeneity and nearly all the heterogeneity in the data is within-axes heterogeneity.
2. When the data points have appreciable protection in all axes.

3. When the contrast among the quadrant is less pronounced and their contents differ in degree rather than in kind.

ADVANTAGES OF NON-CENTERING

In non-centered data, the point of reference is the all zero record. The multivariate analysis will use and describe all departures from this absolute zero. So, it will give an absolute result of species in a study site.

The main advantage of non-centered ordination is that they distinguish disjunctions from more differences in a 'continuum' or 'between-axes' from within-axes' heterogeneity of clusters (Dale 1964). The distinction between unipolar component in the first case and in several unipolar component in the second case is considerable interest. Each such component define and typify a vegetation series. If there is more than one, it suggests compositional disjunction in the sample. The bipolar components describe aspects of the continuous variation within series, each differentiating in a series two compositional phases as species and sites with extreme positive and negative values.

Non-centering is even more advantageous when there is disjunction in the data. It minimizes interference between variation on both sides of the break. It allows a clearer and simpler picture of the data structure and facilitates a phytosociological interpretation of the individual axes. This is improved when non-centered PCA is followed by rotation to simple structure (Noy-Meir 1970, 1971). Interpretation of individual axis allows the analysis to be carried into more dimensions whereas in centered PCA, the interpretation is limited to two or three dimensions. Non-centered PCA is more useful in the following cases:

1. When data exhibit between-axes heterogeneity.
2. When there are clusters of data points such that each cluster has zero projection on some subset of the axes, a different subset of axes for each cluster. In such case each of the first few principal axes passes through one of the qualitatively different clusters and these axes tend to be unipolar (Pielou 1984,Noy-Meir 1973).
3. When the quadrant belongs to groups having non-identical lists of common species.

CONCLUSION

The foregoing comparisons of centering and non-centering of data matrix is not complete. Clear and detailed discussions of the use of centered and uncentered ordinations on different data are given by Orloci (1967), Noy-Meir (1973), Feoli (1977). Non-centering is not uninformative all the time as it is supposed. It may be true in case of homogeneous sample (Noy-Meir 1973) which is rare in nature. The limitation of centered PCA to two or three dimensions can be overcomed by the interpretation of individual axis which allows the analysis to be carried into many dimensions.

There are deficiencies in both centering and non-centering of data. There is no universal rule that tells which one to use in the analysis. But certainly, non-centered ordination has some important advantages in vegetation studies; (1) it allows an assessment of the number and sharpness of discontinuities in the sample, (2) in discontinuous data, it minimizes the interference between the variation on the two sides of the break. The final decisions, however, depend upon the aims of user and the nature of data.

REFERENCES CITED

Basnet, K. (1993), Controls of environmental factors on pattern of Montana rain forest in Puerto Rico. Tropical Ecology, 34:1-15.

Basnet, (1992), Effect of topography on the pattern of trees in tabonuco (dacryodes excelsa) dominated rain forest of Puerto Rico. Biotropica, 24:31-42.

Austin, M.P. and Greig-Smith, P. (1968), The application of quantitative methods to vegetation survey. II. Some methodological problems of data from rain forest. J. Ecol. 56:827-844.

Carleton, T.J. and P.F.Maycock (1980), Vegetation of the Boreal Forest south of James Bay: Non-centered component analysis of the vascular flora. Ecology, 61:1199-1212.

Dale, M.B. (1964), The application of multivariate methods to heterogeneous data. Ph.D. thesis, University of Southampton.

Feoli, E. (1977), On the resolving power of principal component analysis in plant community ordination. Vegetation, 33:119-25.

Gower, J.C. (1966), Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53:325-338.

Groenewoud, H.van. (1965), Ordination and classification of Swiss and Canadian coniferous forests by various biometrics and other methods. Ber.geobot. Forsch Inst. Rubel, 36:28-102.

Lambert, J.M and M.B. Dale (1964), The use of statistics in phytosociology. Adv. Ecol. Res. 2:59-99.

Noy-Meir, I. (1970), Component analysis of the semi-arid vegetation in southeastern Australia. Ph.D. thesis, Australian National University.

_____, (1971), Multivariate analysis of the semi-arid vegetation in southeastern Australia. I. Nodal ordination by component analysis. Proc. ecol. soc. Aust., 6:159-193.

_____, (1973), Data transformation in ecological ordination. I. Some advantages of non-centering. Journal of Ecology, 61:329-341.

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

_____, (1967), Data centering: A review and evaluation with reference to component analysis. Syst. Zool., 16:2208-212.

_____, (1975), On information flow in ordination. Vegetation, 29:11-16.

Pielou, E.C. (1984), The interpretation of Ecological Data: A primer on classification and ordination. John Wiley and Sons Inc. p. 263.

Yarranton, G.A. (1967), Principal components analysis of data from saxicolous bryophyte vegetation at Steps Bridge, Devon. I.A. qualitative assessment of variation in the vegetation. Canadian Journal of Botany, 45:93-115.