Shape and Size-Constrained Principal Components Analysis

Per Sundberg

Systematic Zoology, Vol. 38, No. 2. (Jun., 1989), pp. 166-168.

Stable URL:
http://links.jstor.org/sici?sici=0039-7989%28198906%2938%3A2%3C166%3ASASPCA%3E2.0.CO%3B2-F

Systematic Zoology is currently published by Society of Systematic Biologists.

Points of View

Syst. Zool., 38(2):166–168, 1989

Shape and Size-Constrained Principal Components Analysis

PER SUNDBERG

Department of Zoology, University of Göteborg,
P.O. Box 250 59, S-400 31 Göteborg, Sweden

Natural selection may cause subpopulations to diverge in shape as well as in size. Traditionally, shape has been considered the most important element in the study of intraspecific variation since "shape tends to provide more reliable indications than size on the internal constitution of organisms" (Jolicoeur and Mosimann, 1960). Many authors have regarded size as random noise and have sought ways of removing it from their analyses although the rationale behind this has been criticized and deemed conceptually and mathematically flawed (e.g., Oxnard, 1978; Atchley, 1983; Jungers, 1984; Shea, 1985). There are, however, many situations where it is important to consider shape and size independently, without discarding information about the latter, and others where size has to be properly accounted for in studies of size-correlated shape changes. There are also situations where one wishes to standardize for size, not in order to remove it but to allow comparison of samples including juveniles and adults, or where sexual dimorphism in size may confound the analysis.

Several approaches have been suggested for ways to remove size, or to standardize for it. One immediate problem is that there are no generally agreed-upon definitions of size and shape. Both terms are often used, but seldom are authors explicit about what they mean by them. Three current approaches to a size definition can be recognized, each leading to different definitions of shape (Bookstein et al., 1985; Bookstein, 1989). The oldest and simplest concept of size is to use one variable, for example, height of a snail, as size, and relate other variables to this "size" by forming a ratio with "size" as the denominator. These ratios are then interpreted as shape variables. Such ratios are probably still the most common approach to quantifying shape despite the statistical (Pearson, 1897; Atchley and Anderson, 1978, and references therein) and conceptual ("size" cannot be imagined as just one particular single linear dimension [Humphries et al., 1981]) arguments against its use. A further objection is that biometric analysis should be based on several characters simultaneously, since selection presumably acts upon whole phenotypes and not on single characters.

Another way of looking at size is to assume it corresponds to general factors, i.e., linear combinations of appropriate suites of variables (see Mosimann and Malley, 1979; Bookstein et al., 1985). There are many ways in which such a general-size factor could be defined. It is most commonly taken as the first eigenvector obtained from a principal component analysis based on a variance-covariance, or correlation, matrix of log-transformed morphometric variables (Jolicoeur, 1963; Reyment et al., 1984). This has a long history in morphometry and its use can be traced back to Tessier

(1938). Jolicoeur and Mosimann (1960) argued that the first principal component can be viewed as size if all coefficients of the first eigenvector are about equal and of the same sign. The remaining eigenvectors are then interpreted as shape if they have coefficients of mixed signs and values. The orthogonality imposed by principal component analysis makes all vectors independent and if the first vector is taken as size, the remaining vector can then be called "size-free" shape. Thus, the lack of correlation between vectors is an artifact built into the method itself and nothing inherent in the data. On the contrary, from a biological point of view, it is more realistic to expect size-correlated shape changes, i.e., allometry. So, even if the first component can be interpreted as size, the other components are more realistically interpreted as mixtures of size and shape (Mosimann, 1970; Sprent, 1972; Humphries et al., 1981). Sprent (1972) further makes the point that since allometry implies differences in shape with size, it is unclear why any models are designed to partition out the factor of growth in multivariate analysis of shape.

Still, correction for size effect is of interest for example when comparing samples that vary in size because they are taken at different times of the year, or for studies concerned with patterns of covariation among variables. Various ways have been suggested to improve principal component analysis to better deal with size correction. Humphries et al. (1981) supplemented the technique with regression to find shape discriminators, independent of size, between populations. This so-called sheared principal component analysis is described in Rohlf and Bookstein (1987). Somers (1986) has recently suggested another modification of principal component analysis that aims to partition out a size axis in a population. Somers' approach aims to extract "a size component equivalent to an isometric size vector (. . .). The remaining information (i.e., by definition shape and random variation) is factored into components of shape covariation." The isometric size vector is computed as a function of the number of characters, as shown by Jolicoeur (1963), and is removed from the correlation matrix prior to the component analysis.

TABLE 1. Pearson's product-moment correlation coefficients (absolute values) between specimen scores along the size-constrained (Somers, 1986) first principal component and the subsequent components for three data sets: A. Measurements of thirteen morphological characters of the rough periwinkle (*Littorina saxatilis*) ($n = 120$), from Sundberg (1988). B. Three characters from female painted turtle (*Chrysemys picta marginata* ($n = 24$), from Jolicoeur and Mosimann (1960). C. Five characters from sparrows ($n = 49$), from Manly (1986).

Data set:	A	B	C
Correlation coefficient (r) between size-constrained first component and component			
2	0.98	0.10	0.65
3	0.99	0.84	0.06
4	0.73		0.28
5	0.60		0.76
6	0.83		
7	0.84		
8	0.88		
9	0.55		
10	0.67		
11	0.64		
12	0.19		
13	0.63		
Critical two-tailed r-values ($P = 0.05$):	0.18	0.40	0.28

The method is available in the form of a BASIC program from Somers. It extracts an isometric size vector from a correlation matrix based on log-transformed variables, extracts eigenvectors from the manipulated matrix, and then computes the component scores. When this program was used for studying shape differentiation between subpopulations of the periwinkle *Littorina saxatilis* using 13 shell characters (Sundberg, 1988), it was found that most of the 12 subsequent component scores were correlated with those along the first component to such a degree that the eigenvectors were suspected not to be orthogonal. To check whether this result was a consequence of high intra-character correlations in my data set, or a general feature of the program and the approach, the size-constrained PCA was applied to two other data sets: female painted turtles (from Jolicoeur and Mosimann, 1960), and Bumpus' sparrow data (in Manly, 1986). The results (Table 1) show that all of these sets display significant correlations between the first (the size) and subsequent components. The component scores are obtained by multiplying the data matrix by the eigenvectors; thus, non-zero correlations are expected even if the eigenvectors are orthogonal, and, furthermore, correlations arise from allometric shape changes. Truly orthogonal eigenvectors should still not lead to such high correlations as many of those reported in Table 1 if the original character values are properly scaled. This indicates that there are some problems in either the actual program, or in the approach in itself. Rohlf and Bookstein (1987) have also identified potential problems with Somers' approach and show that it will only lead to orthogonal vectors under restricted conditions. They suggest a simple matrix manipulation which will yield orthogonal vectors and make Somers' (1986) approach similar in spirit to Burnaby's (1966) method.

REFERENCES

ATCHLEY, W. R. 1983. Some genetic aspects of morphometric variation. Pages 346-363 *in* Numerical Taxonomy (J. Felsenstein, ed.). NATO ASI Series, Series G, Ecological Sciences, No. 1. Springer-Verlag, Berlin.

ATCHLEY, W. R., AND D. ANDERSON. 1978. Ratios and the statistical analysis of biological data. Syst. Zool., 27:71-78.

BOOKSTEIN, F. L., B. C. CHERNOFF, R. L. ELDER, J. M. HUMPHRIES, G. R. SMITH, AND R. E. STRAUSS. 1985. Morphometrics in evolutionary biology. Academy of Natural Sciences of Philadelphia Special Publ. 15. 277 pp.

BOOKSTEIN, F. L. 1989. "Size and shape": A comment on semantics. Syst. Zool., 38:173-181.

BURNABY, T. P. 1966. Growth-invariant discriminant functions and generalized distances. Biometrics, 22: 96-110.

HUMPHRIES, J. M., F. L. BOOKSTEIN, B. CHERNOFF, G. R. SMITH, R. L. ELDER, AND S. G. POSS. 1981. Multivariate discrimination by shape in relation to size. Syst. Zool., 30:291-308.

JOLICOEUR, P. 1963. The multivariate generalization of the allometry equation. Biometrics, 19:497-499.

JOLICOEUR, P., AND J. E. MOSIMANN. 1960. Size and shape variation in the painted turtle, a principal component analysis. Growth, 24:339-354.

JUNGERS, W. L. 1984. Aspects of size and scaling in primate biology with special reference to the locomotor skeleton. Yearb. Phys. Anthropol., 27:73-97.

MANLY, B. F. J. 1986. Multivariate statistical methods. A primer. Chapman and Hall, London.

MOSIMANN, J. E. 1970. Size allometry: Size and shape variables with characterizations of the lognormal and generalized gamma distributions. J. Am. Stat. Assoc., 65:930-945.

MOSIMANN, J. E., AND J. D. MALLEY. 1979. Size and shape variables. Pages 175-189 *in* Multivariate methods in ecological work (L. Orlóci, C. R. Rao, and W. M. Stiteler, eds.). International Co-operative Publishing House, Fairland, Maryland.

OXNARD, C. E. 1978. One biologist's view of morphometrics. Ann. Rev. Ecol. Syst., 9:219-241.

PEARSON, K., 1897. On a form of spurious correlation which may arise when indices are used in the measurements of organs. Proc. Royal Soc., London, 60: 489-502.

REYMENT, R. A., R. E. BLACKITH, AND N. A. CAMPBELL. 1984. Multivariate morphometrics. Second edition. Academic Press, London.

ROHLF, F. J., AND F. L. BOOKSTEIN. 1987. A comment on shearing as a method for "size correction." Syst. Zool., 36:356-367.

SHEA, B. T. 1985. Bivariate and multivariate growth allometry: Statistical and biological considerations. J. Zool., Lond., 206:567-590.

SOMERS, K. M. 1986. Multivariate allometry and removal of size with principal components analysis. Syst. Zool., 35:359-368.

SPRENT, P. 1972. The mathematics of size and shape. Biometrics, 28:23-37.

SUNDBERG, P. 1988. Microgeographic variation in shell characters of *Littorina saxatilis* Olivi. A question mainly of size? Biol. J. Linn. Soc., 35:169-184.

TESSIER, G. 1938. Un essai d'analyse factorielle. Les variants sexuels de *Maia squinada*. Biotypologie, 7: 73-96.

Received 26 January 1988; accepted 19 October 1988