TRANSFORMATION OF SEQUENTIAL QUANTITATIVE CHARACTERS

R. L. THOMAS, J. E. GRAFIUS and S. K. HAHN*
Michigan State University

Received 1.v.70

1. Introduction

Any particular observed developmental event may be explicitly or implicitly visualised as one of a chain of connected occurrences within the organism. Such events are in general influenced by the magnitude and direction of previous or concomitant ones. For events or traits which may be defined as quantitative measures the association between characters may be described by their covariance. Under certain circumstances it would be advantageous and interesting to examine a trait in isolation or divorced from previous developmental events and thus the influence of known correlations with previous causative events should be removed. There should, of course, be some sound reason to suspect that these correlations do imply causation and undoubtedly all causative prior events will not have been measured. The mechanism of causation which might imply linkage, pleiotropy or any other genetic or physiological influence does not directly concern us at this stage. What we do suggest is that a character be examined in genetic analysis in isolation from as well as in connection with previous developmental characters.

In the present paper a method is suggested for isolation of a sequential character from its association with previous characters in the sequence by removing correlations. The method described was originally derived by vector techniques, the problem having been visualised in geometrical terms. It was subsequently seen that the method so obtained was, in effect, identical to the basic form of a statistical transformation described by Rao (1952) which is a step involved in his multivariate analysis and is also involved, with some modification, in the calculation of Mahalanobis' (1928) D^2 statistic. Both the geometrical and the statistical approach are described and the correspondence between the two indicated.

2. The methods

Geometric approach

The correlation between the members of two sets of variables, A and B, may be shown to be equivalent to the cosine of the angle between two vectors **A** and **B** representing the sum of these two variable sets—provided that the set members are transformed to standard (S.-deviation) units to allow this representation.

Proof: To prove $r = \cos\theta$, where θ is the angle between vectors **A** and **B**.

Definition: $\mathbf{A}.\mathbf{B} = |\mathbf{A}||\mathbf{B}|\cos\theta$ (in vector notation).

* Presently professor, Suwon University, Korea.

A.B is the scalar or dot product of vectors **A** and **B** and is calculated as the inner product or the sum of cross products of all members of the set, *i.e.* statistically, $\Sigma(ab)$. $|\mathbf{A}|$ and $|\mathbf{B}|$ are the magnitudes of vectors **A** and **B** respectively which are calculated as the square root of the sum of squares of the members of the particular set in consideration, *i.e.* statistically:

$$|\mathbf{A}|\,|\mathbf{B}| = \sqrt{\Sigma(a^2) \times \Sigma(b^2)}$$

and

$$\therefore \cos\theta = \frac{\mathbf{A}.\mathbf{B}}{|\mathbf{A}|\,|\mathbf{B}|} = \frac{\Sigma(ab)}{\sqrt{\Sigma(a^2) \times \Sigma(b^2)}} = r$$

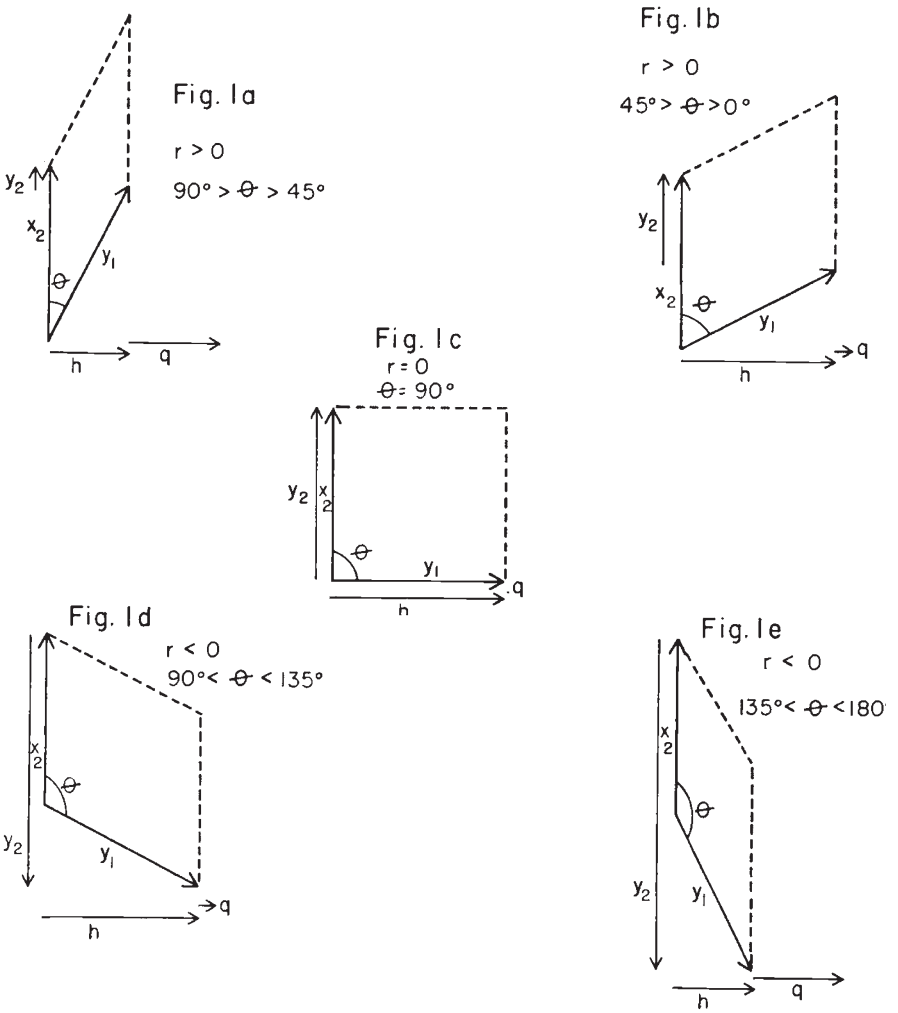

FIG. 1.—Geometric representation of the effects of correlation between trait vectors x_2 and y_1. $\mathbf{y}_2 = \mathbf{x}_2 - \mathbf{y}_1 \cos\theta$ and $q = \mathbf{y}_1 - \mathbf{y}_1 \cos\theta$.

Thus if a 90° angle exists between vectors $\mathbf{y}_1$ and $\mathbf{x}_2$ representing the set of variables X_1 and X_2 respectively, then $r = \cos\theta = 0$ and y_1 and x_2 are orthogonal. In fig. 1*c* this relationship is illustrated for this particular case

of θ with $\mathbf{x}_2$ and $\mathbf{y}_1$ reprdesented as unit vectors. Furthermore, if we consider the characters y and x to be multiplicative (*e.g.* two components of yield), the area of the figure obtained by connecting up the two lines parallel to $\mathbf{y}_1$ and $\mathbf{x}_2$ is exactly the product of x_2 and y_1 since the two vectors are orthogonal. If any positive or negative correlation exists between x_2 and x_1 then the two vectors are non orthogonal. The graphical consequences of correlation are illustrated in figs. 1*a*, *b*, *d* and *e*, one weak and one strong positive and negative correlation being considered; again the vectors $\mathbf{x}_2$ and $\mathbf{y}_1$ are drawn as unit vectors. The product of vectors $\mathbf{x}_2$ and $\mathbf{y}_1$, in figs. 1*a*, *b*, *d* and *e* is now represented as the series of parallelograms drawn by connecting up lines parallel to $\mathbf{x}_2$ and $\mathbf{y}_2$ and it is clear that once orthogonality is lost the area of these parallelograms is less than unity (unlike fig. 1*c*). The area of these four parallelograms can be adjusted to equal unity by adding to the height (h) a quantity $q = \mathbf{y}_1 - \mathbf{y}_1 \cos\theta$, as shown in these figures.

The main purpose of these illustrations is to demonstrate the effect of removing correlations—*i.e.* the influence of the angle θ. Let us assume that y_1 is the developmentally earlier character influencing x_2. It follows from these figures that y_2, the adjusted value of $\mathbf{x}_2$, has value $\mathbf{y}_2 = \mathbf{x}_2 - \mathbf{y}_1 \cos\theta$ and that this value increases from 0 with $r = +1$ ($\theta = 0°$), through 1 with $r = 0$ ($\theta = 90°$) to $+2$ with $r = -1$ ($\theta = 180°$). Thus when a positive correlation exists the second character in the sequence can be said to be inflated by the first and should be corrected downwards and when a negative correlation exists the second character is deflated and should be upvalued to obtain its independent value. The transformation is applied to the individual members of the sets and results naturally in an r value of 0 between the two transformed sets. The new vector $\mathbf{y}_2$ is orthogonal to $\mathbf{y}_1$ and the product of these lengths is the area of a rectangle and for instance if x_1 and x_2 are multiplicative yield components this area $y_2.y_1$ may be considered as the product, independent of association between components.

The diagrammatic representation becomes more cumbersome to present with three sequential variables and impossible with four. However, vector techniques enable us to work freely in Euclidean n-space and the method is extendable to n variables. The first variable is left unadjusted, the second adjusted by the factor dependent on its connection with the first, the third by a factor dependent on the association with the first two, and so on. Since the actual adjustments are identical to those derived by Rao we shall now proceed to consider his method.

Statistical approach

With observed variables $x_1 \ldots x_p$, Rao (1952) obtains $y_1 \ldots y_p$ transformed, uncorrelated corresponding variables by the following method:

1. $Y_1 = X_1$
2. $Y_2 = X_2 - a_{21}Y_1$
3. $Y_3 = X_3 - a_{32}Y_2 - a_{31}Y_1$

.....

.....

.....

$pY_p = X_p - a_{p,\,p-1}Y_{p-1} \ldots a_{p1}Y_1$

Where $a_{21} = \dfrac{\lambda_{21}}{\lambda_{11}} = \dfrac{\text{covariance } X_2Y_1}{\text{variance } X_1} = \dfrac{\text{cov } (X_2X_1)}{\text{v } (X_1)}$

$$a_{31} = \frac{\lambda_{31}}{\lambda_{11}} = \frac{\text{cov } (X_3X_1)}{\text{v } (X_1)}$$

$$a_{32} = \frac{\lambda_{32} - a_{21}\lambda_{31}}{\lambda_{11}}$$

Equation 2 from Rao corresponds to the transformation $\mathbf{y}_2 = \mathbf{x}_2 - r\mathbf{y}_1$ obtained by vector methods except that the latter is in standard units and the former is in unweighted form.

$$a_{21} = \frac{\text{cov } (X_2X_1)}{\text{v } (X_1)} = r \left| \frac{\sigma X_2}{\sigma X_1} \right| = \frac{\text{cov } (X_2X_1)}{\sigma X_2 \sigma X_1} \cdot \frac{\sigma X_2}{\sigma X_1} = b_{X_2X_1}$$

Similarly, the remaining equations of Rao presented may be shown to be equivalent to equations derived by vector methods. Rao's equations and thus these obtained by geometry have considerable respectability and there is no further requirement of justifying the method of removing correlations. When large numbers of variables are being considered Rao (1952) suggests (p. 347) a computational method for obtaining the " a " coefficients involving pivotal condensation of a variance, covariance matrix. Murty and Arunachalam (1967) have published an IBM 1620 program for Mahalanobis' D^2 estimation a section of which may be used (up to statement 212) to give transformed $y_1 \ldots y_p$ data and Mr J. Barnard at this University has adapted the appropriate part of the program for a CDC 3600, which may be obtained on request.

3. Discussion

Data transformation is an acceptable biological procedure, although the particular transformation adopted is usually arrived at by combinations of experience, intuition, and trial and error. The efficiency of the transformation is often judged on a semi aesthetic basis—*e.g.* a curvilinear relationship when expressed on a log scale may appear more pleasing or convenient and may be more amenable to conventional linear statistical analysis. Criteria do however exist to determine the value of transformation and one of the main ones in whether the basis for prediction is increased. The scaling tests of Mather (1949) and Cavalli (1952) may be used for genetic data and they allow a comparison of transformed versus untransformed data—according to which give the better statistical fit, *i.e.* reduce the apparent level of epistasis to a minimum. The predictive value of the transformation suggested above will not be discussed here since it forms the subject of a subsequent paper in this series.

Mather (1949) has stated that any genetic transformation should be applied equally to the basic observations of each cell of the experiment, *i.e.* not merely to means or totals of several observations, and that the same transformations should be applied within all cells of the experiment. Essentially the first of these criteria is followed by the suggested method for the removal of correlations within the set (cell) from which the correlations were calculated. Whether one overall correlation and hence transformation is to

be calculated and applied respectively to all cells of an experiment is debatable and will depend on whether there is heterogeneity of correlation between cells and on the object of the analysis. This point will be dealt with in subsequent papers.

The main advantage of the transformation described here is that it does what it intends to do, *i.e.* removes the effects of correlation and is objective in that once the correlations are known and the causative link stated or postulated no alternatives or trial and error procedures are allowable. The applications of the transformation will depend on the interests of the researcher but two general statements may be made. First, the isolation of sequential characters enables the researcher to examine any part of the sequence *in vitro*. Second, since the difference between transformed and untransformed data can only be due to correlation this numerical difference may be regarded as a convenient measure of all correlations operating on a given trait and analysed as such.

4. Summary

1. Two approaches for removing correlations between correlated traits within populations are presented.

2. The first is based on a geometrical representation of traits as vectors and the correlations as functions of the angles between vectors.

3. The second is a restatement of a statistical method presented by Rao (1952). The identity of the two methods is shown.

4. Some suggestions for utilising the techniques in genetic analyses of traits which develop sequentially are given.

Note and Acknowledgment.—Published as article number 5000 of Michigan State University Agricultural Experiment Station. We wish to acknowledge the adivce and help of Dr M. W. Adams, Dr C. M. Harrison, Dr C. Lee and Mr J. Barnard in the preparation of this manuscript. This work was partically funded by the Malting Barley Improvement Association.

5. References

CAVALLI, L. L. 1952. An analysis of linkage in quantitative inheritance. In E. C. Reeve and C. W. Waddington (Eds.), *Quantitative Inheritance*, Her Majesty's Stationary Office, London, pp. 135-144.

MAHALANOBIS, P. C. 1928. A statistical study of Chinese head measurement. *Man In India*, *8*, 32-64.

MATHER, K. 1949. *Biometrical Genetics*. Methuen, London.

MURTY, B. R., AND ARUNACHALAM, V. 1967. Computer programmes for some problems in biometrical genetics—1. Use of Mahalanobis' D^2 in classificatory problems. *Indian Jour. Gen. & Pl. Breeding*, *27*, 60-69.

RAO, C. R. 1952. *Advanced Statistical Methods in Biometric Research*. John Wiley and Sons, New York. 390 pp.