Behavior Genetics, Vol. 17, No. 1, 1987

Application of Nonlinear Factor Analysis to Genotype–Environment Interaction

Peter C. M. Molenaar[1] and Dorret I. Boomsma[2]

Received 6 Nov. 1985—Final 10 May 1986

*The intention of this paper is to show how the methods of nonlinear factor analysis as developed by McDonald (*Br. J. Math. Stat. Psychol. *20:205–215, 1967) can be used to study genotype–environment interaction. The method is applied to the interaction of genotype and within-family environmental influences. Simulated twin data are used to illustrate how this type of interaction may be detected and estimated. It is shown that estimates of genetic influences are not affected by* G × E *interaction.*

KEY WORDS: genotype–environment interaction; nonlinear factor analysis; twin data.

INTRODUCTION

Various statistical models have been applied to the analysis of genotype–environment ($G \times E$) interaction, especially in plant and animal genetics (Fulker *et al.*, 1972; Freeman, 1973). These approaches include analysis of variance and regression analysis, often in combination with the use of external measures to assess the environment. In human genetics the presence of $G \times E$ interaction can affect estimates of genetic and environmental influences from twin and family studies (Rao and Morton, 1974; Rao *et al.*, 1976; Plomin *et al.*, 1977; Eaves, 1984; Lathrope and Lalouel, 1984).

In this paper we want to show how the theory of nonlinear factor analysis given by McDonald can be applied to study the interaction between genotype and within-family environmental influences. At the level

[1] University of Amsterdam, Department of Psychology, Weesperplein 8, 1018 XA Amsterdam, The Netherlands.

[2] Free University, Department of Experimental Psychology, De Boelelaan 1115, 1081 HV Amsterdam, The Netherlands.

of second-order statistics (i.e., variances and covariances), influences due to this type of interaction cannot be distinguished from within-family environmental influences (Martin *et al.*, 1986). A solution to this problem for multiple continuous variables is offered by the recognition that in factor analysis the product of two orthogonal factors will act just like an additional orthogonal factor (Bartlett, 1953). In order to distinguish between such an interaction factor and a genuine third factor, Bartlett suggested examining the correlation between the factor scores of the presumed interaction factor and a constructed product of the factor scores of its component factors. Such an examination, however, is complicated by the problem of rotational indetermination in factor analysis. McDonald (1967) developed a general approach to solve this problem, which may be generalized to the study of $G \times E$ interaction. One of the advantages of this approach is that there is no need for an index of either measured environment or measured genotype. On the other hand, however, the method can be applied only when at least three variables are measured on the same subject.

THE INTERACTION MODEL

Under the assumptions that gene action is additive, mating is random, and all environmental influences are within families, the genetic model for a single continuous variable may be written as

$$P = hG + eE + iI, \qquad (1)$$

where P is the observed phenotype, G and E are genetic and within-family environmental influences, and $I = G \times E$ represents the interaction between genotype and within-family environmental influences; $h = \sigma_G/\sigma_P$, the square root of heritability, where σ_G and σ_P are the square roots of the unstandardized genetic and phenotypic population variances; $e = \sigma_E/\sigma_P$, the square root of environmentability, where σ_E is the square root of the unstandardized environmental variance; and $i = \sigma_I/\sigma_P$, where σ_I is the square root of the unstandardized variance that can be attributed to the interaction between genotype and within-family environmental influences. It is assumed that P, G, E, and I in Eq. (1) are standardized, i.e., $E(P) = E(G) = E(E) = E(I) = 0$ and $\mathrm{var}(P) = \mathrm{var}(G) = \mathrm{var}(E) = \mathrm{var}(I) = 1$. If, in addition, G and E are mutually statistically independent, the phenotypic variance is equal to

$$V_P = h^2 V_G + e^2 V_E + i^2 V_I = 1 = h^2 + e^2 + i^2. \qquad (2)$$

A multivariate extension of Eq. (1) is obtained by

$$P_k = h_k G_k + e_k E_k + i_k I_k, \qquad k = 1, \ldots, p.$$

If we consider only one common genetic, one common within-family environmental, and one common interaction factor, we get the following factor model:

$$P_k = \Delta_k G + H_k E + J_k I + \epsilon_k, \qquad k = 1, \ldots, p. \tag{3}$$

Or, in matrix notation:

$$P = \Delta G + HE + JI + \epsilon, \tag{4}$$

where G, E, and I denote standardized common genetic, within-family environmental, and interaction factors, respectively, and ϵ ($p \times 1$) represents similar influences specific to each observed phenotype P. That is, each specific influence ϵ may itself consist of specific genetic, environmental, and interaction components (Martin and Eaves, 1977). Whatever the precise nature of such specific influences may be, however, is immaterial to our analysis of interaction between common genetic and common within-family environmental factors. For the sake of clarity of presentation, it is therefore sufficient to conceive of specific influences ϵ as representing specific within-family environmental influences. P is a $p \times 1$ column vector of unstandardized observed phenotypes. Δ, H, and J are $p \times 1$ column vectors of factor loadings that are defined as $\Delta = \sigma_{\mathrm{P}} h$, $H = \sigma_{\mathrm{P}} e$, and $J = \sigma_{\mathrm{P}} i$, where σ_{P} is a diagonal matrix of population phenotypic standard deviations.

The variance–covariance matrix of P now can be expressed as

$$\sum_{\mathrm{P}} = \Delta\Delta' + HH' + JJ' + \epsilon^2, \tag{5}$$

where Δ, H, and J are column vectors ($p \times 1$) of factor loadings and ϵ^2 is a $p \times p$ diagonal matrix of specific variances.

In the following we consider Eqs. (4) and (5) with respect to monozygotic (MZ) and dizygotic (DZ) twin data. Notice that the common within-family environmental factor E is uncorrelated within twin pairs. Consequently, the product $I = G \times E$ is also uncorrelated within pairs and the interaction factor I behaves like another common within-family environmental factor E'. Furthermore, remember that G and E are mutually independent zero mean factors, and hence

$$\mathrm{cov}(G, I) = \mathrm{var}(G)\,\mathrm{E}(E) = 0,$$

$$\mathrm{cov}(E, I) = \mathrm{E}(G)\,\mathrm{var}(E) = 0.$$

These results agree with Bartlett's (1953) observation that the product of two independent factors acts just like an additional independent factor.

The exact distribution of I is given by Craig (1936) and Springer (1979, p. 155).

If we consider an alternative factor model in which I is replaced by an additional common within-family environmental factor E', where E' is uncorrelated with G and E,

$$P = \Delta G + HE + JE' + \epsilon, \tag{6}$$

then the latter model will be indistinguishable at the level of second-order moments from the interaction model described by Eq. (4) since the expected dispersion matrix $\sum_P$ according to Eq. (6) is again given by Eq. (5) and hence is the same as for the original interaction model.

We are now in a position to state the main thrust of our approach to the study of $G \times E$ interaction: if a factor model including at least one common genetic (G) and two within-family environmental factors (E and E') is found to yield a reasonable fit to MZ and DZ twin data, then it is possible to identify the presence of $G \times E$ interaction. That is, it is possible to test that $E' = I = G \times E$. Accordingly, given a satisfactorily fitting model described by Eq. (6), we present a way to distinguish between this model and the factor interaction model described by Eq. (4). Such a distinction cannot be made at the level of second-order moments, because we saw that Eqs. (4) and (6) are indistinguishable at this level. Instead, we take a recourse to a consideration of third-order moments and, in particular, focus on $\mathrm{E}(G\,E\,E')$. The latter expression refers to the third-order moment of factor scores in Eq. (6) and it is easily seen that

$\mathrm{E}(G\,E\,E') = 0$ if E and E' are genuine common within-family environmental factors, and

$\mathrm{E}(G\,E\,E') = 1$ if $E' = \mathrm{I}$.

Thus, we always start with the fit of a factor model including at least one common genetic and two within-family environmental factors. Next, the presence of an interaction factor can be identified by computation of the above third-order moment of factor scores. There is one caveat with this approach, however, involving the problem of factor indetermination (cf. Lawley and Maxwell, 1971). As discussed more fully below, the factors E and E' in Eq. (6) are unique up to orthogonal rotations. This state of affairs undermines any straightforward determination of the third-order moment of factor scores, because the estimation of factor scores is dependent upon the particular orientation of the latent dimension in question.

Similar observations have been made by McDonald (1967) in the context of nonlinear factor analysis. In order to cope with the problem of factor indetermination in the identification of interaction between factors, McDonald proposes a general factor rotation procedure that maximizes third-order moments between factor scores. With respect to Eq. (6), notice that the problem of factor indetermination does not occur with the common genetic factor G: the pattern of genetic weights associated with Δ in the expected matrices of mean cross-products within and between MZ and DZ twins ensures that this factor has a determinate orientation in factor space (see below). In contrast, the pattern of weights for H and J is the same, and therefore the problem of factor indetermination does occur with E and E'. In order to specify McDonald's factor rotation procedure for the latter factors, let E^* and E'^* denote orthogonally rotated instances of E and E', respectively. We now look for uniquely rotated factors E^* and E'^* such that E'^* resembles $G \times E^*$ as much as possible. Application of McDonald's rotation procedure to Eq. (6) thus amounts to minimizing the following expression:

$$\min_{\Theta} \mathrm{E}(E'^* - G \times E^*)^2, \tag{7}$$

where

$$E^* = \cos(\Theta)E - \sin(\Theta)E',$$
$$E'^* = \sin(\Theta)E + \cos(\Theta)E',$$

and where Θ is the angle of planar rotation of E and E'. Setting the derivative of Eq. (7) with respect to Θ to zero, it is found that this expression is minimized by taking

$$\tan 2\Theta = \frac{\mathrm{E}(E'^2\, G) - \mathrm{E}(E^2\, G)}{2\mathrm{E}(G\, E\, E')}. \tag{8}$$

In a nutshell, then, one first fits Eq. (6) to MZ and DZ twin data and estimates the factor scores G, E, and E'. Next, Θ is determined from Eq. (8) and E and E' are orthogonally rotated through an angle Θ, yielding E^* and E'^*. Finally, $\mathrm{E}(G\, E^*\, E'^*)$ is computed, and if this statistic is 1, then $E'^* = G\, E^* = I^*$, indicating the presence of interaction between the common genetic and the within-family environmental factors. On the other hand, the above statistic will be zero if E'^* is a genuine environmental factor.

ESTIMATION IN THE INTERACTION MODEL

The factor model given by Eq. (6) can be applied to MZ and DZ twin data. For the expected matrices of mean cross-products between and within MZ and DZ twins, we get

$$\sum_{\text{MZB}} = 2\Delta\Delta' + HH' + JJ' + \epsilon^2,$$

$$\sum_{\text{MZW}} = HH' + JJ' + \epsilon^2,$$

$$\sum_{\text{DZB}} = 1.5\ \Delta\Delta' + HH' + JJ' + \epsilon^2,$$

$$\sum_{\text{DZW}} = 0.5\ \Delta\Delta' + HH' + JJ' + \epsilon^2.$$

If estimates of these four matrices are used as input matrices in LISREL (Jöreskog and Sörbom, 1981) to obtain loadings on the common and specific factors (Boomsma and Molenaar, 1986), the weighting for H and J is exactly the same. Consequently, an additional manipulation is required to arrive at a completely identified LISREL model. This can be done by constraining one of the loadings in H or J at zero. The latter constraint is inconsequential to the obtained goodness of fit, as it involves the choice of a particular orientation of E and E' within the closure of solutions under orthogonal rotation.

The results of the LISREL estimation in the factor model given by Eq. (6) are used to obtain estimates of the factor scores G, E, and E'. Denote the column vector of factor scores of the ith member of a twin pair by

$$f'(i) = [G(i), E(i), E'(i)], \qquad i = 1, 2,$$

and let $P(i)$ be the column vector of observed phenotypes of this ith member. Then the factor scores of each pair of MZ and DZ twins can be estimated by means of the regression method (Lawley and Maxwell, 1971):

$$\begin{bmatrix} \hat{f}(1) \\ \hat{f}(2) \end{bmatrix} = \Phi(I_6 + \Lambda'\Psi\Lambda\Phi)^{-1}\,\Lambda'\Psi^{-1} \begin{bmatrix} P(1) \\ P(2) \end{bmatrix},$$

where I_6 is the 6 × 6 unity matrix,

$$\Lambda = \begin{bmatrix} \Delta & H & J & 0 & 0 & 0 \\ 0 & 0 & 0 & \Delta & H & J \end{bmatrix},$$

$$\Psi = \begin{bmatrix} \epsilon^2 & 0 \\ 0 & \epsilon^2 \end{bmatrix},$$

$$\Phi = \text{cov}[f, f'],$$

the latter being a 6 × 6 unity matrix save for the 4,1 (1,4) element, which is 1 for MZ pairs and 0.5 for DZ pairs.

The final step consists of determining Θ by means of Eq. (8), where

$$\hat{\mathrm{E}}(G\,E\,E') = 1/N \sum \hat{G}\,\hat{E}\,\hat{E}'$$

and where similar estimators for the other third-order moments apply. N is the total number of subjects. Orthogonal rotation of $\hat{E}$ and $\hat{E}'$ through an angle Θ than yields factor scores $\hat{E}^*$ and $\hat{E}'^*$ minimizing Eq. (7). In the simulation studies discussed below, the rotated vectors of estimated factor loadings $\hat{H}^*$ and $\hat{J}^*$ thus obtained will closely resemble the vectors of true factor loadings H and J that have been used in the construction of the data, whereas the original estimates $\hat{H}$ and $\hat{J}$ will not. In practical applications minimization of Eq. (7) may involve a few iterations (mostly three) of this procedure until additional rotations become negligible. Computation of $\hat{\mathrm{E}}(G\,E^*\,E'^*)$ then will indicate the plausibility of Eq. (4), i.e., the presence of $G \times E$ interaction.

In the following sections we discuss a few illustrative applications of the above method to simulated data. In order to illustrate the validity of the proposed method, two data sets have been constructed: one by means of Eq. (4) (including $G \times E$ interaction) and a similar one by means of Eq. (6) (including a second within-family environmental factor). These examples involve five-variate vectors P of observed phenotypes.

EXAMPLE I: DATA SIMULATION

For 200 MZ and DZ twin pairs five-variate phenotypes were simulated according to the factor interaction model described by Eq. (4):

$$P = \Delta G + HE + JI + \epsilon,$$

where for MZ twins

$$\mathrm{cor}[G(1), G(2)] = 1,$$
$$\mathrm{cor}[E(1), E(2)] = 0,$$
$$\mathrm{cor}[I(1), I(2)] = 0,$$

and for DZ twins

$$\mathrm{cor}[G(1), G(2)] = 0.5,$$
$$\mathrm{cor}[E(1), E(2)] = 0,$$
$$\mathrm{cor}[I(1), I(2)] = 0.$$

Random variables were generated using IMSL subroutine FTGEN (IMSL, Inc., 1979). All unique variances of specific influences ϵ are 1, hence ϵ^2 is the 5 × 5 unity matrix. In addition, the elements of the vectors Δ, H, and J have been assigned the following values:

$$\Delta' = 5\ 6\ 7\ 8\ 9,$$

$$H' = 7\ 7\ 3\ 7\ 7,$$

$$J' = 5\ 9\ 5\ 9\ 5.$$

ESTIMATION

The four 5 × 5 matrices of mean cross-products between and within MZ and DZ twins were input for LISREL to obtain estimates of factor loadings on the unique and common factors. To make the model identified, the first loading on the interaction factor was fixed at zero. Next, factor scores were computed for each subject by means of the regression method described above. The factor scores on the second and third factors (i.e., E and E') were then rotated. Remember that in our case there is no need for rotation of the first factor, since its loadings are determined uniquely by the genetic weights.

RESULTS

Table I shows the LISREL factor loadings on the three common factors and the loadings for each variable on the specific environmental factor. The χ^2 for this model was 29.14 with 41 df (P = 0.91). Also in Table I are the factor loadings after rotation. As can be seen the correspondence between the true factor loadings and the estimated loadings after rotation is very reasonable. Rotation took three iterations and yielded the following estimate of the third-order moment of factor scores: $\hat{E}(G\ E^*\ E'^*) = 0.92$. Hence, the method correctly identifies E' as being an interaction factor. It is clear that the estimates of the proportions of variance that can be attributed to G are not influenced by the presence of an interaction between genotype and within-family environmental influences or by orthogonal rotation of E and E'.

EXAMPLE II

With the same parameter values as in the previous example, another data set for 200 MZ and DZ twin pairs was simulated according to Eq. (6). In this case E' is a genuine within-family environmental factor. For this model LISREL gave a χ^2 of 25.9 (df = 41, P = 0.96). As was the case with Example I, the correspondence between the true factor loadings

Table I. Results of First Simulation Study

LISREL				Rotation				
G	E	I	ϵ	G	E^*	I^*	ϵ	
Estimated factor loadings (Example I)								
5.64	8.98	—	1.00	5.64	7.43	5.05	1.00	
6.95	11.68	3.05	0.99	6.95	7.78	8.98	0.99	
7.47	5.63	2.20	1.05	7.47	3.42	4.98	1.05	
8.93	11.60	3.06	0.98	8.93	7.87	9.06	0.98	
9.67	8.87	0.10	0.97	9.67	7.28	5.07	0.97	
Total variances								Total
31.81	80.64	—	1.00	31.81	55.21	25.50	1.00	113.50
48.30	131.79	9.30	0.98	48.30	60.53	80.64	0.98	190.45
55.80	31.70	4.48	1.10	55.80	11.70	24.80	1.10	93.40
79.74	134.56	9.36	0.96	79.74	61.94	82.08	0.96	224.73
93.51	78.68	0.01	0.94	93.51	53.00	25.70	0.94	173.15
Percentages of variance								
28.0	71.1	—	1.0	28.0	48.6	22.4	1.0	
25.4	69.2	4.8	0.5	25.4	31.8	42.3	0.5	
59.7	33.9	5.2	1.2	59.7	12.5	26.6	1.2	
35.5	59.9	4.2	0.4	35.5	27.6	36.5	0.4	
54.0	45.4	0.0	0.6	54.0	30.6	14.8	0.6	

and the estimated loadings after rotation was very close. The rotation procedure gave an estimated third-order moment $\hat{E}(G\ E^*\ E'^*) = 0.04$. Hence, the method correctly identifies E' as being a genuine within-family environmental factor.

DISCUSSION

Two important restrictions of the model in its present form are that it requires at least three observations on each subject and that each of the orthogonal factors that make up the interaction factor must also be present as a separate factor in the model. As summarized by Eaves (1984), however, studies of $G \times E$ interaction in plant and animal genetics have shown that genes that control sensitivity to the environment are often different from genes that control average response and, also, that different genes control sensitivity to different environments. It is possible, however, to generalize the proposed method to interaction models in which $G \times E$ interaction factors occur without the presence of the constituting genotype and/or environmental factors.

Based on the examples given above the conclusion seems warranted that the proposed method constitutes a viable first step toward a general approach of $G \times E$ interaction. In itself the proposed method leads to a valid analysis of $G \times E$ interactions underlying at least three observed phenotypes.

REFERENCES

Bartlett, M. S. (1953). Factor analysis in psychology as a statistician sees it. In Uppsala Symposium on Psychological Factor Analysis, *Nordisk Psykologi's Monograph Series No. 3*, Ejnar Mundsgaards, Copenhagen.

Boomsma, D. I., and Molenaar, P. C. M. (1986). Using LISREL to analyze genetic and environmental covariance structure. *Behav. Genet.* **16**:237–250.

Craig, C. C. (1936). On the frequency function of xy. *Ann. Math. Stat.* **7**:1–15.

Eaves, L. J. (1984). The resolution of genotype × environment interaction in segregation analysis of nuclear families. *Genet. Epidemiol.* **1**:215–228.

Freeman, G. H. (1973). Statistical methods for the analysis of genotype-environment interactions. *Heredity* **31**:339–354.

Fulker, D. W., Wilcock, J., and Broadhurst, P. L. (1972). Studies in genotype-environment interaction. I. Methodology and preliminary multivariate analysis of a dial cross of eight strains of rat. *Behav. Genet.* **2**:261–287.

IMSL, Inc. (1979). *IMSL Library Reference Manual Edition 7*, IMSL Inc., Houston, Tex.

Jöreskog, K. G., and Sörbom, D. (1981). *LISREL: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*, National Educational Resources, Chicago.

Lathrope, G. M., and Lalouel, J. M. (1984). Path analysis of family resemblance and gene-environment interaction. *Biometrics* **40**:611–625.

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths, London.

Martin, N. G., and Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**:79–95.

Martin, N. G., Eaves, L. J., and Heath, A. C. (1986). Detecting the contributions of measured and unmeasured genotypes and environments and their interactions to liability for a disease (submitted for publication).

McDonald, R. P. (1967). Factor interaction in nonlinear factor analysis. *Br. J. Math. Stat. Psychol.* **20**:205–215.

Plomin, R., DeFries, J. C., and Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* **84**:309–322.

Rao, D. C., and Morton, N. E. (1974). Path analysis of family resemblance in the presence of gene-environment interaction. *Am. J. Hum. Genet.* **26**:767–772.

Rao, D. C., Morton, N. E., and Yee, S. (1976). Resolution of cultural and biological inheritance by path analysis. *Am. J. Hum. Genet.* **28**:228–242.

Springer, M. D. (1979). *The Algebra of Random Variables*, John Wiley & Sons, New York.

Edited by C. R. Cloninger