Using Factor Scores to Detect G × E Interactive Origin of "Pure" Genetic or Environmental Factors Obtained in Genetic Covariance Structure Analysis

P.C.M. Molenaar, D.I. Boomsma, D. Neeleman, and C.V. Dolan

Department of Psychology, University of Amsterdam (P.C.M.M., C.V.D.), and Department of Psychology Vrije Universiteit (D.I.B., D.N.), Amsterdam, The Netherlands

Moment expressions for individual factor scores can serve as simple tests for the presence of a particular class of interaction factors that are disguised as pure genetic and/or environmental factors. That is, individual genetic and environmental factor scores may be used to construct fourth-order moments of these factors in order to test whether a common genetic or environmental factor in the multivariate genetic factor model is in fact of the interactive origin concerned. Expected fourth-order moments are derived for cases with and without interaction. Application of fourth-order moments of factor scores to detect interactive origin of common factors is illustrated with simulated twin data.

Key words: genotype-environment interaction, fourth-order moments

INTRODUCTION

The statistical analysis of genotype-environment (G × E) interaction can be conducted by means of various approaches, including analysis of variance and regression analysis, often in combination with the use of definite measures to assess the environment [e.g., Jinks and Fulker, 1970; Fulker et al., 1972; Freeman, 1973] or the genotype [Martin et al., 1987]. In human genetics the presence of G × E interaction can in some cases affect estimates of genetic and environmental influences as determined in twin and family studies [Lathrope et al., 1984]. Molenaar and Boomsma [1987] considered the application of nonlinear factor analysis [McDonald, 1967] to the study of

Received for publication September 23, 1988; revision accepted October 4, 1989.

Address reprint requests to P.C.M. Molenaar, Department of Psychology, University of Amsterdam, Weesperplein 8, 1018 XA Amsterdam, The Netherlands.

G × E interaction underlying multivariate continuous measures. The latter approach is based on a special rotation of multiple within-families environmental factors that maximizes third-order moments of factor scores in order to determine whether a second factor that behaves like a within-families environmental factor really is G × E.

In this paper we discuss a general test of G × E interaction underlying multivariate observations that can be applied to covariance structure models with singular genetic, within- and between-families environmental factors, or a subset of these factors. The test we propose can be applied to covariance models where the factors that make up the interaction term are not both present as separate factors in the model. The finding of at most a single factor of each type (a common finding in applied genetic covariance structure analysis) obviates the need to invoke special rotation techniques, but requires a test of fourth-order moments of factor scores, as the presence of G × E interaction does not show up in first-, second-, or third-order statistics or, alternatively, in a general test of the model (e.g., Rao and Morton [1974]: "Gene-environment interaction has little chance of being recognized by goodness-of-fit tests, even in an unrealistically large body of data"). G × E interaction may be suggested, however, by the presence of kurtosis in the raw data.

The test we propose enables detection of G × E interaction *disguised* as pure genetic and/or environmental factors. Specifically, it will be shown that regular genetic covariance structure analysis cannot distinguish between a genetic factor G and a G × C (genotype × between-families environment) interaction factor, or a within-families environmental factor E and a G × E or a C × E interaction factor. The use of factor scores requires the availability of raw data and implies that the test is only reliably applicable to the communal part of the factor model.

MOMENT EXPRESSIONS UNDER INTERACTION

Given the assumptions that gene action is additive and mating is random, the standard multivariate genetic model for vector-valued phenotypes can be written in matrix notation as [e.g., Martin and Eaves, 1977]:

$$P = \Lambda_g G + \Lambda_e E + \Lambda_c C\,(1) + \epsilon \qquad (1)$$

where $P = (P_1, \ldots, P_p)'$ denotes a random p-dimensional column vector of centered phenotypes and $'$ denotes transposition. The random univariate latent variables G, E, and C represent genetic, within-families (nonshared) environmental, and between-families (shared) environmental factors with p-dimensional loadings Λ_g, Λ_e and Λ_c, respectively. ϵ is a random p-dimensional vector composed of influences unique to each phenotype P_i, $i = 1, \ldots, p$. For the moment, we will concentrate on the common factors G, E and C, which are taken to be mutually uncorrelated Gaussian variables with mean zero and unit variance:

$$\sigma_{ii} = 1 \text{ and } \sigma_{ij} = 0,\ i \neq j,\ i,j \subset \{G, E, C\},$$

where σ_{ii} denotes variance and σ_{ij} denotes covariance. On an individual basis, G, E, and C in equation 1 represent individual factor scores, i.e., a person's genetic, nonshared environmental, and shared environmental deviations.

We first consider the case in which G in equation 1 is replaced by the product $G^* = G \times C$. The interaction factor G^* will still have unit variance, $\sigma_{G^*G^*} = \sigma_{GG}\,\sigma_{CC} = 1$, and will be uncorrelated with C (and E): $\mathrm{cov}[G^*, C] = \mathrm{cov}[G \times C, C] = \mu_G\sigma_{CC} = 0$, where μ_G denotes the (zero) mean of G. Moreover, the interaction factor G^* will behave exactly like a genetic factor in that it will have unit correlation in MZ twin pairs and an average correlation of 0.5 in DZ twin pairs. Consequently, G^* is indistinguishable from G in a genetic analysis based on second-order moments.

Next, consider the replacement of E in equation 1 by the product $E^* = E \times G$ (similar remarks apply to the replacement of E by $E^* = E \times C$). Again, the interaction factor E^* will have unit variance and will be uncorrelated with G (and C). Also, the interaction factor E^* will behave exactly like a within-families environmental factor in that it is uncorrelated within and between MZ and DZ twin pairs. Consequently, E^* is indistinguishable from E in an analysis of seccond-order moments associated with the standard genetic model.

Reasoning along similar lines, it can be seen that the remaining alternatives do not lead to an exact correspondence with equation 1. Specifically, replacement of G by $G^* = G \times E$ or C by $C^* = C \times E$ gives rise to a model with two within-families environmental factors, as has been discussed in Molenaar and Boomsma [1987]. Replacement of C by $C^* = C \times G$ gives rise to an additional factor resembling a second genetic factor. In this paper we consider the case in which the genetic model given by equation 1 (or a submodel including either G and E, G and C, or E and C) yields a satisfactory fit to the data. The possibility then exists that G really is an interaction factor $G^* = G \times C$, or that E is an interaction factor $E^* = E \times G$ or $E^* = E \times C$.

Simple tests for the presence of interaction can be based on a consideration of higher-order moments of the factors. As G, E, and C are taken to be zero mean Gaussian variables, it is clear that, irrespective of the eventual presence of interaction factors, third-order moments will always be zero. Hence, we will have to take recourse to a consideration of fourth-order moments. For instance, if G is a genuine genetic factor, then $E[G^4] = 3\sigma_{GG}*\sigma_{GG} = 3$. Furthermore, $E[G^2C^2] = \sigma_{GG}\sigma_{CC} = 1$. In contrast, if G is an interaction factor $G^* = G \times C$, then $E[G^{*4}] = 3\sigma_{GG}*\sigma_{GG}*3\sigma_{CC}*\sigma_{CC} = 9$, whereas, $E[G^{*2}C^2] = \sigma_{GG}3\sigma_{CC}*\sigma_{CC} = 3$. Similar fourth-moment expressions can be obtained in order to distinguish between a genuine within-families environmental factor E and an interaction factor $E^* = E \times G$ or $E^* = E \times C$ (see Table I).

The moment expressions in Table I have been derived for pure cases only. That is, G is either a pure genetic factor or a pure G × C interaction factor. We could, on the other hand, consider hybrid cases in which G in equation 1 is replaced by $G^* = aG + bG \times C$, where $b = \sqrt{1 - a^2}$. It then follows that $E[G^{*4}] = 9 - 6a^4$, while $E[G^{*2}C^2] = 3 - 2a^2$. Similar expressions are obtained for hybrid E^* factors. Clearly, the expected values of these fourth-order moments of hybrid factors can vary smoothly between the associated values for the pure cases. Consequently, the proper null hypothesis for test-

TABLE I. Fourth-Order Moments in the Standard Genetic Model With and Without Interaction

	X = G Y = C		X = E Y = G		X = E Y = C	
	G = G	G = G × C	E = E	E = E × G	E = E	E = E × C
$E[X^4]$	3	9	3	9	3	9
$E[X^2Y^2]$	1	3	1	3	1	3

ing against these alternatives is given by the expected values obtained for the genetic model without interaction.

Studies of G × E interaction in plant and animal genetics have shown that genes that control sensitivity to environment are often different from genes that control average response [Eaves, 1984]. This could be modeled by decomposing the genotype into $G = aG_1 + bG_2$, where $b = \sqrt{1 - a^2}$, while G_1 and G_2 are mutually uncorrelated, zero mean Gaussian variables with unit variance. Accordingly, G × E interaction then can be defined as $E^* = G_2E$. It now turns out that $E[E^{*4}]$ is not affected, but $E[E^{*2}G^2] = 3 - 2a^2$. Of course, this sharp differential prediction could again be tempered by allowing E* to be a hybrid factor.

In a nutshell, these moment expressions can serve as simple tests for the presence of various forms of interaction. Their application requires the availability of the factor scores concerned. Estimation of these factor scores is discussed in a companion paper [Boomsma et al., 1990] and therefore will not be considered here. Instead we will now turn to a simulation study in order to illustrate the validity of the proposed approach.

AN ILLUSTRATIVE APPLICATION

The standard genetic model given by equation 1 can be applied to MZ and DZ twin data. For the expected matrices of mean cross-products between and within MZ and DZ twin pairs we have [e.g., Martin and Eaves, 1977]:

$$\begin{aligned}
\Sigma_{MZB} &= 2\Lambda_g\Lambda'_g + \Lambda_e\Lambda'_e + 2\Lambda_c\Lambda'_c + U^2\\
\Sigma_{MZW} &= \Lambda_e\Lambda'_e + U^2\\
\Sigma_{DZB} &= 1.5\Lambda_g\Lambda'_g + \Lambda_e\Lambda'_e + 2\Lambda_c\Lambda'_c + U^2\\
\Sigma_{DZW} &= 0.5\Lambda_g\Lambda'_g + \Lambda_e\Lambda'_e + U^2
\end{aligned}$$

where U^2 is a p × p diagonal matrix of unique variances. This constrained multigroup model can be fitted to the data by means of LISREL [Jöreskog and Sörbom, 1988; Boomsma and Molenaar, 1986]. Next, the parameter estimates thus obtained are used to estimate the associated factor scores by means of the regression method [Boomsma et al., 1990]. In the final step, estimates of fourth-order moments of factor scores mentioned in Table I are tested for the presence of interaction.

To illustrate the validity of our approach we will first consider applications to simulated data according to the following models: Model I is the standard factor model without interaction; in Model II the genetic factor is an interaction factor: $G^* = G \times C$; in Model III the within-family environmental factor is an interaction factor: $E^* = G \times E$; and in Model IV, $E^* = C \times E$. With each model 5-dimensional vectors of observed phenotypic values have been simulated for 100 MZ and 100 DZ twin pairs, where the factor loadings and unique variances are the same across the four models. True parameter values and maximum-likelihood estimates of the parameters as obtained from the simulated data under each model are presented in Table II (standard errors are given within parentheses).

All parameter estimates are close to their true values and have correspondingly small estimated standard errors. The chi-squared goodness-of-fit statistics of the interaction models II–IV, however, turn out to be higher than expected (remember that in

TABLE II. True and Estimated Factor Loadings (p = 5)

	G	E	C	Unique
True	5	7	5	1
	6	7	3	1
	7	3	5	1
	8	7	3	1
	9	7	5	1
Model I	4.72 (.50)	6.85 (.27)	5.47 (.61)	1.15 (.08)
	5.67 (.47)	6.99 (.27)	3.44 (.63)	0.96 (.06)
	6.77 (.40)	2.92 (.17)	5.34 (.64)	1.11 (.07)
	7.62 (.51)	6.86 (.29)	3.64 (.76)	1.04 (.06)
	8.69 (.57)	6.90 (.30)	5.63 (.86)	1.00 (.06)
	χ^2_{40} = 49.38	(P = .147)		
Model II	4.76 (.50)	6.87 (.27)	5.52 (.61)	1.16 (.08)
	5.81 (.48)	7.04 (.27)	3.50 (.64)	0.95 (.06)
	7.28 (.42)	2.99 (.17)	5.42 (.68)	1.10 (.07)
	7.88 (.52)	6.92 (.29)	3.72 (.78)	1.04 (.06)
	8.91 (.58)	6.96 (.29)	5.72 (.88)	1.00 (.06)
	χ^2_{40} = 61.03	(P = .018)		
Model III	5.74 (.49)	6.83 (.27)	4.61 (.65)	1.16 (.08)
	6.69 (.48)	6.80 (.28)	2.59 (.68)	0.96 (.06)
	7.20 (.41)	2.92 (.17)	5.00 (.67)	1.11 (.07)
	8.64 (.53)	6.71 (.29)	2.80 (.82)	1.03 (.06)
	9.69 (.58)	6.75 (.30)	4.79 (.92)	1.00 (.06)
	χ^2_{40} = 63.04	(P = .012)		
Model IV	4.54 (.50)	7.07 (.28)	5.19 (.60)	1.16 (.08)
	5.48 (.46)	6.92 (.27)	3.16 (.61)	0.96 (.06)
	6.73 (.40)	2.95 (.17)	5.20 (.64)	1.10 (.07)
	7.44 (.51)	6.99 (.29)	3.35 (.75)	1.04 (.06)
	8.52 (.57)	6.92 (.29)	5.33 (.85)	1.00 (.06)
	χ^2_{40} = 67.04	(P = .005)		

Model I: no interaction; Model II: G = G × C; Model III: E = G × E; Model IV: E = C × E.

each case the true model has been fitted). Apparently the non-normality of the data simulated according to these models has a deteriorating effect on this statistic.

Table III presents for each model the estimated fourth-order moments mentioned in Table I. Expected deviations from standard values for each model have been underlined. Notice that to distinguish Model III ($E^* = G \times E$) from Model IV ($E^* = E \times C$) one needs to compare $E(G^2E^2)$, which is 2.899 for Model III and 0.764 for Model IV, with $E(E^2C^2)$, which is 0.987 for Model III and 4.012 for Model IV. Table III shows that each model is identified by the expected pattern of fourth-order moments.

As a further illustration of the power of the present approach we consider an application to simulated bivariate phenotypic values for 100 MZ and 100 DZ twins pairs. In this case, four observations are available for each pair, whereas the complete genetic model given by equation 1 involves five factor scores for each DZ twin pair. Hence, the required estimates of these factor scores can no longer be reliably obtained. Instead, we will consider bivariate phenotypic values simulated according to a submodel of equation 1, which only includes genetic (G), within-family environmental (E), and unique (ϵ) factors. In particular, we consider applications to simulated data according

TABLE III. Fourth-Order Moments (p = 5)

	$E(G^4)$	$E(E^4)$	$E(C^4)$	$E(G^2E^2)$	$E(G^2C^2)$	$E(E^2C^2)$
Model I	2.989	3.175	3.514	1.066	1.016	1.065
Model II	8.746	3.102	3.566	1.016	3.501	1.121
Model III	2.975	8.283	3.511	2.899	1.012	0.987
Model IV	2.942	10.677	3.531	0.764	0.996	4.012

Model I: no interaction; Model II: G = G × C; Model III: E = G × E; Model IV: E = C × E.

TABLE IV. True and Estimated Factor Loadings (p = 2)

	G	E	Unique
True	5	6	1
	7	4	1
Model A	4.63 (.40)	6.02 (.25)	0.97 (.07)
	6.77 (.37)	4.11 (.23)	0.97 (.07)[a]
	$\chi^2_7 = 13.01$	$(P = .072)$	
Model B	5.54 (.41)	5.99 (.26)	0.98 (.07)
	7.37 (.39)	4.00 (.23)	0.98 (.07)[a]
	$\chi^2_7 = 20.77$	$(P = .004)$	

Model A: no interaction; Model B: E = G × E.
[a]Error variances constrained to be equal.

TABLE V. Fourth-Order Moments (p = 2)

	$E(G^4)$	$E(E^4)$	$E(G^2E^2)$
Model A	3.075	3.249	1.044
Model B	3.085	8.315	2.727

Model A: no interaction; Model B: E = G × E.

to a standard model (A) and a second model (B) involving E* = G×E interaction. Again, both models have the same factor loadings and unique variances (Table IV). For both models, maximum likelihood estimates, etc., are given in Table IV. The estimated fourth-order moments of factor scores under scrutiny are presented in Table V. It can be seen that even with bivariate data the interaction of G × E is reliably detected.

DISCUSSION

The present approach to the detection of G × E interaction is based on the estimation of fourth-order moments of factor scores and hence pertains to the communal part of the factor model [cf., Boomsma et al., 1990]. The factor scores concerned are estimated according to the regression method involving the use of estimated loadings and unique variances. Consequently, this approach involves the use of a threefold composition of estimators and therefore can be expected to yield a threefold accumulation of estimation errors. In addition, estimates of fourth-order moments in themselves appear to be unreliable [Ratcliff, 1979]. In spite of all this, the results of our applications to simulated data involving 100 MZ and 100 DZ twin pairs and 5-variate or 2-variate observations turn out to be quite comforting. It appears, then, that the power of the present approach may not be too small to detect the presence of interaction in realistic situations. Moreover, the regression method for the estimation of factor scores can be improved by correcting fo the uncertainty inherent in the estimated factor loadings and unique variances. These issues require extensive elaboration, however, and will be considered in a forthcoming paper.

The obtained maximum likelihood solutions appear to be reliable in spite of the non-normality of phenotypic observations due to factor interaction. In contrast, unweighted least-squares estimates of these parameters obtained with the same data turn out to be entirely inferior. The reason for the good performance of maximum likelihood estimators has been indicated by Gourieroux, Monford, and Trognon [1984] who show that these estimators keep their desirable qualities under quite general conditions, i.e., if the data follow a distribution of arbitrary exponential type [see also Browne and Shapiro, 1988]. Our simulation results indicate, however, that the chi-squared goodness-of-fit statistic is sensitive to departures from normality. To assess the fit of a model, then, one should also consider the magnitudes of normalized residuals, for example.

In our illustrative applications we did not consider the sometimes intricate issue of model selection. Instead, we only fitted the true model to the data in order to show the validity of our approach to the detection of factor interaction. It remains to be seen how much the presence of interaction, implying non-normality of the observations, affects the course of model selection. One intriguing development, based on artificial intelligence techniques [Glymour et al., 1987], may become of use in this respect.

In conclusion, a consideration of fourth-order moments of factor scores in the standard genetic model seems to be a promising way to detect the presence of factor interactions. The restriction to multiplicative factor interactions is defensible because it involves the most important term in a Taylor expansion of any nonlinear function describing the exact form of interaction at stake. The proposed method may be of service to detect factor interaction disguised as pure genetic and/or environmental factors and is easy to apply. That is, whatever nonlinear functional dependency between G and E may exist, it will be generically described to an important degree by a multiplicative interaction described in this paper. At present we do not know whether failure to find evidence for this multiplicative interaction precludes G × E interaction of any type.

REFERENCES

Boomsma DI, Molenaar PCM (1986): Using LISREL to analyze genetic and environmental covariance structure. Behav Genet 16:237–250.

Boomsma DI, Molenaar PCM, Orlebeke JF (1990): Estimation of individual genetic and environmental factor scores. Genet Epidemiol this issue.

Browne MW, Shapiro A (1988): Robustness of normal theory methods in the analysis of linear latent variate models. Br J Math Stat Psychol 41:193–208.

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

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

Fulker DW, Wilcock J, Broadhurst PL (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.

Glymour C, Scheines R, Spirtes P, Kelly K (1987): "Discovering Causal Structure, Artificial Intelligence, Philosophy of Science, and Statistical Modeling." New York: Academic Press.

Gourieroux C, Monford A, Trognon A (1984): Pseudo-maximum likelihood methods: Theory. Econometrica 17:287–304.

Jinks JL, Fulker DW (1970): Comparison of the biometrical genetical, MAVA and classical approaches to the analysis of human behavior. Psychol Bull 73:311–349.

Jöreskog KG, Sörbom D (1988): "LISREL VII: A Guide to the Program and Applications." Chicago: Spss Inc.

Lathrope GM, Lalouel JM, Jacquard A (1984): Path analysis of family resemblance and gene-environment interaction. Biometrics 40:611–625.

Martin NG, Eaves LJ (1977): The genetic analysis of covariance structure. Heredity 38:79–95.

Martin NG, Eaves LJ, Heath AC (1987): Prospects for detecting genotype × environment interactions in twins with breast cancer. Acta Genet Med Gemellol 36:5–20.

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

Molenaar PCM, Boomsma DI (1987): Application of nonlinear factor analysis to genotype-environment interaction. Behav Genet 17:71–80.

Rao DC, Morton NE (1974): Path analysis of family resemblance in the presence of gene-environment interaction. Am J Hum Genet 26:767–772.

Ratcliff R (1979): Group reaction time distributions and an analysis of distribution statistics. Psychol Bull 86:446–461.

Edited by D.C. Rao and G.P. Vogler