Genetic Epidemiology 7:83–91 (1990)

Estimation of Individual Genetic and Environmental Factor Scores

D.I. Boomsma, P.C.M. Molenaar, and J.F. Orlebeke

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

Implicit in the application of the common-factor model as a method for decomposing trait covariance into a genetic and environmental part is the use of factor scores. In multivariate analyses, it is possible to estimate these factor scores for the communal part of the model. Estimation of scores on latent factors in terms of individual observations within the context of a twin/family study amounts to estimation of individual genetic and environmental scores. Such estimates may be of both theoretical and practical interest and may be provided with confidence intervals around the individual estimates. The method is first illustrated with simulated twin data and next is applied to blood pressure data obtained in a Dutch sample of 59 male adolescent twin pairs. Subjects with high blood pressure can be distinguished into groups with high genetic or high environmental scores.

Key words: confirmatory factor analysis, twins, blood pressure

INTRODUCTION

The factor analytic approach to multivariate quantitative genetic problems as suggested by Martin and Eaves [1977] has proven to be very fruitful. This approach not only enables the analysis of genetic and environmental causes of covariance between measures in multivariate data sets, but is also easily generalized to structural modeling approaches such as time-series analysis [Boomsma and Molenaar, 1987]. Moreover, for multivariate problems this approach enables identification of parameters that cannot be estimated in univariate designs. These include estimation of genotype-by-environment interaction without information on measured genotype or environment [Molenaar and Boomsma, 1987], a test of whether the structure of the means of latent genetic and environmental factors can be explained by the same model that explains

Received for publication September 22, 1988; revision accepted October 4, 1989.

Address reprint requests to D.I. Boomsma, Department of Psychology, Vrije Universiteit, De Boelelaan 1115, 1081 HV Amsterdam, The Netherlands.

the covariance structure [Dolan et al., 1989], and estimation of individual factor scores, which is the subject of this paper.

Estimation in factor analysis can be seen as a two-stage procedure [Lawley and Maxwell, 1971]. First, the parameters in the model, e.g., factor loadings and unique variances, are estimated and then these are used to provide estimates of individual factor scores. Multivariate quantitative genetics has limited itself to the first stage. Data from genetically related individuals have been used to estimate loadings of variables on genetic and environmental factors, so that variances and covariances may be explained in terms of these factors. There is no reason, however, why the second step should not be taken and estimates of these factor loadings be used to compute individual factor scores.

As the latent factors used in a factor analysis of twin or family data are genetic and environmental factors, estimation of factor scores amounts to estimation of individual genetic and environmental scores. Although the use of these scores is always implicit in the application of factor analysis, they cannot be determined precisely, but have to be estimated, since the number of common and unique factors always exceeds the number of observed variables.

In this paper we first show how these scores may be obtained using simulated twin data. Since for each twin pair at least three factor scores are to be estimated (i.e., one genetic and two environmental scores in the case of identical twins) the proposed method needs to be used in multivariate data sets. Although it is also possible to obtain these scores in a univariate analysis, this yields highly intercorrelated estimates of independent G and E factor scores. We use simulated data on monozygotic (MZ) and dizygotic (DZ) twins to estimate factor scores using two different methods. Next, we compare these methods by computing correlations between true and estimated factor scores. In the final section an application to blood pressure data is presented.

FACTOR MODEL

The factor analysis model employed in quantitative genetics analyses can be represented as:

$$P_{ij} = h_iG_j + e_iE_j + c_iC_j + U_{ij},$$
$$i = 1, \ldots, p \text{ (variables)}$$
$$j = 1, \ldots, n \text{ (subjects)}$$

The measured phenotype (P) (henceforth the j suffix indicating subjects will be omitted) consists of multiple variables that are a function of a subject's underlying genotype (G), nonshared (within-families) environment (E), and common (between-families) environment (C). In addition, there is a unique part (U) to each variable that itself may consist of a genetic and a nongenetic part. At the moment this is immaterial to our analysis, as we will estimate factor scores for the communal part of the model only. h, e, and c are the p-variate factor loadings of measured variables on the latent factors. To estimate these loadings we do not need to know the individual factor scores, as the expectation for the (phenotypic) $p \times p$ covariance matrix only consists of Λ, where $\Lambda = (h,e,c)$ is a $p \times m$ matrix of factor loadings and $m = 3$ is the number of

latent factors, Ψ, which is the m × m correlation matrix of factor scores and Θ, which is a p × p diagonal matrix of unique variances:

$$\Sigma_{pp} = \Lambda \Psi \Lambda' + \Theta$$

In a multivariate analysis of twin data according to this factor model, Σ is a 2p × 2p covariance matrix of observations on twin 1 and twin 2, and Λ is a 2p × 2m matrix of loadings of these observations on latent genotypes and nonshared and common environments (i.e., m = 2 × 3) of twin 1 and twin 2, where the loadings are constrained to be equal for twin 1 and 2. Θ is a 2p × 2p matrix of unique variances that are also equal for both members of a twin pair [e.g., Heath et al., 1989]. Λ and Θ are estimated from the data, and Ψ (2m × 2m) is usually given a priori (e.g., the correlation between G of twin 1 and G of twin 2 is 1 for MZ and .5 for DZ twins; the correlation of C of twin 1 and twin 2 is 1). Next, estimates of factor loadings and unique variances can be used to construct individual scores on the genetic and environmental factors. That is, for each individual a vector f with factor scores of the following form can be considered:

$$f = A'P, f = [\hat{G}, \hat{E}, \hat{C}]$$

where the superscript ′ denotes transposition and A is a matrix with weights that is constant across subjects, depending only on the factor loadings and the unique variances. For each individual the multivariate phenotype (P) is measured and f is to be estimated. Lawley and Maxwell [1971] discuss two well-known estimators for f: the Thurstone regression method [Thurstone, 1935] and the Bartlett estimator [Bartlett, 1937]. Both methods involve a least-squares principle and in each case the factor scores are linear functions of the original variables. Notice, however, that these estimators are derived to optimize different criteria and therefore have distinct characteristics.

The Thurstone regression method is equivalent to finding the linear regression of factor scores on phenotypes, and the weight matrix A is obtained by minimizing the sum of squares of the difference between estimated and true factor scores. This results in:

$$A = \Theta^{-1} \Lambda (I + \Lambda' \Theta^{-1} \Lambda \Psi)^{-1} \Psi$$

The Bartlett estimator is obtained by minimizing the sum of squares of each individual's estimated unique factor scores across all variables. The resulting formula is:

$$A = \Theta^{-1} \Lambda (\Lambda' \Theta^{-1} \Lambda)^{-1}$$

The Bartlett estimator of the common factor scores is a maximum-likelihood estimator when Λ and Θ are known and the observed data have a multivariate normal distribution. As can be seen, the Bartlett estimator does not depend on whether or not the factors are correlated. This implies that to estimate a person's genetic and environmental scores no data from genetically related individuals are required (although these data are of course required to obtain estimates of Λ and Θ). As shown below, this is achieved at the cost of less accurate estimates of factor scores. The regression method can also be used in this case, however, by simply ignoring the existing correlations between latent factors.

SIMULATION

As an illustration, 5-variate data were simulated for 100 MZ and DZ twins, according to the factor model described above. That is, data were simulated with underlying genetic, nonshared, and common environmental factors. Factor scores were simulated to have zero means and unit variance and come from a multivariate normal distribution. Loadings on the genetic and environmental factors were 5 6 7 8 9 for G; 7 7 3 7 7 for E; and 5 9 5 9 5 for C. All unique variance was environmental and not shared between twins. The unique variance was 10 for all 5 variables.

A 3-factor model was fitted to these data using LISREL-VII [Jöreskog and Sörbom, 1988], and the weight matrix A was computed according to the regression formula and the Bartlett method. The Thurstone regression formula is also used by LISREL-VII for the estimation of factor scores. We did not find any differences between LISREL-VII factor scores and scores that were estimated using our own programs. Table I shows results of fitting the 3-factor model to the simulated data described above.

Once estimates of factor loadings and unique variances have been obtained in a good-fitting model, the weight matrix A can be computed for the estimation of factor scores. Table II shows the matrices that were obtained by the regression and the Bartlett method.

It is clear from the first two matrices that for the estimation of factor scores, both a subject's own observations and observations from his or her cotwin are used. This results in two different weight matrices: one for MZ and one for DZ twins. The Bartlett estimator, in contrast, only uses an individual's own observations, and there is a single formula that applies to both MZ and DZ twins.

Table III shows correlations between true and estimated factor scores for MZ and DZ twins for the regression and the Bartlett method. These correlations show that genetic and environmental scores can be reliably estimated at the individual level. For the regression estimator, correlations are somewhat higher for MZ than for DZ twins and, in both groups, are highest for genetic and common environmental factor scores. Correlations obtained for the Bartlett method are lower than the ones obtained by the regression method. The correlations for MZ twins are not better than for DZ twins, and the lowest correlations are again observed for nonshared environmental scores. The regression method thus gives better estimates of genetic and environmental factor scores when the correlation between true and estimated scores is used as a criterion. This is not surprising, as only the regression method uses information from relatives in the parameter estimation as well as in the construction of the factor scores.

TABLE I. Model Fitting: Estimated Factor Loadings and Unique Variances for Simulated Twin Data

	G	E	C	U^2
V1	4.94	6.96	5.71	9.43
V2	5.87	6.89	9.50	10.13
V3	6.68	2.61	5.56	9.34
V4	7.89	6.63	9.94	9.42
V5	8.93	6.74	5.98	9.00

$\chi^2 = 78.17$ (df = 90; $P = .809$).

TABLE II. Weight Matrices for Simulated Twin Data (Weights Multiplied by 1,000)

Regression estimator

MZ	V1T1	V2T1	V3T1	V4T1	V5T1	V1T2	V2T2	V3T2	V4T2	V5T2
G1	−17	−37	43	−07	54	−17	−37	43	−07	54
E1	70	22	−40	00	38	35	−10	−53	−33	03
C1	−16	43	08	36	−52	−16	43	08	36	−52
G2	−17	−37	43	−07	54	−17	−37	43	−07	54
E2	35	−10	−53	−33	03	70	22	−40	00	38
C2	−16	43	08	36	−52	−16	43	08	36	−52
DZ	V1T1	V2T1	V3T1	V4T1	V5T1	V1T2	V2T2	V3T2	V4T2	V5T2
G1	−35	−44	77	03	70	03	−25	05	−16	32
E1	89	31	−76	−10	20	15	−21	−16	−23	23
C1	−16	42	09	36	−51	−16	42	09	36	−51
G2	03	−25	05	−16	32	−35	−44	77	03	70
E2	15	−21	−16	−23	23	89	31	−76	−10	20
C2	−16	42	09	36	−51	−16	42	09	36	−51

Bartlett estimator

	V1	V2	V3	V4	V5
G	−59	−103	123	−19	137
E	188	10	−171	−69	78
C	−75	114	45	106	−161

When comparing the two methods, another important consideration is the variances of the estimated scores. Table IV shows means and standard deviations of simulated and estimated scores. Bartlett estimates clearly have greater variability than do the regression estimates.

CONFIDENCE INTERVALS

From standard Kalman filtering techniques [Sage and Melsa, 1971] we derived the standard errors of the regression estimates of factor scores [see also Lawley and Maxwell, 1971, p. 108]:

$$V = \Psi [\Psi^{-1} - \Lambda' \Sigma^{-1} \Lambda] \Psi$$

where Λ and Ψ are matrices of factor loadings and correlations between factor scores and Σ is the estimated covariance matrix. V is the 6×6 covariance matrix of the sampling distribution of estimated factor scores, and the standard errors of the estimated factor scores are given by the square root of its diagonal elements. Table V gives the confidence intervals for the simulated data discussed above. These depend on the factor loadings and unique variances (through Σ) only. As the amount of unique variance increases, confidence intervals around individual estimates become larger and correlations between true and estimated factor scores as shown in Table III may become smaller.

APPLICATION TO BLOOD PRESSURE DATA

The methods outlined above were applied to blood pressure data obtained in 34 MZ (mean age = 16.6; SD = 1.8) and 25 DZ (mean age = 17.0; SD = 1.7) male

TABLE III. Correlations of True and Estimated Factor Scores for Twin 1 and Twin 2 (Decimal Point Omitted)

	E(G1)	E(E1)	E(C1)	E(G2)	E(E2)	E(C2)
MZ (regression method)						
G1	910*	081	211	910*	011	211
E1	020	879*	096	020	−199	096
C1	124	100	936*	124	076	936*
G2	910*	081	211	910*	011	211
E2	088	−218	−008	088	894*	−008
C2	124	100	936*	124	076	936*
DZ (regression method)						
G1	884*	213	100	510*	029	100
E1	295*	773*	369*	132	−185	369*
C1	045	341*	897*	−058	084	897*
G2	510*	−028	031	843*	125	031
E2	065	−213	185	255*	844*	185
C2	045	341*	897*	−058	084	897*
MZ (Bartlett estimates)						
G1	797*	079	072	848*	−189	165
E1	026	721*	−011	−091	006	033
C1	−004	−052	859*	044	−078	838*
G2	797*	079	072	848*	−189	165
E2	093	−011	−141	−024	770*	−058
C2	−004	−052	859*	044	−078	838*
DZ (Bartlett estimates)						
G1	842*	007	016	436*	087	−059
E1	117	676*	244*	150	−144	228
C1	057	254*	741*	−166	−005	803*
G2	450*	−027	−042	815*	−007	−085
E2	078	−195	153	099	777*	−022
C2	057	254*	741*	−166	−005	803*

*$P < 0.001$.

TABLE IV. Means and SD of Simulated and Estimated Factor Scores

	Simulated data		Regression method		Bartlett method	
MZ G1	0.04	1.02	−0.04	0.92	−0.06	1.17
E1	−0.08	0.94	−0.09	0.88	−0.03	1.29
C1	0.02	1.10	0.09	1.00	0.06	1.38
G2	0.04	1.02	−0.04	0.92	−0.04	1.27
E2	0.09	1.06	0.08	0.94	−0.02	1.40
C2	0.02	1.10	0.09	1.00	0.18	1.37
DZ G1	−0.09	1.01	−0.11	0.90	−0.14	1.22
E1	−0.07	0.96	0.02	0.82	0.05	1.22
C1	−0.02	0.90	−0.03	0.79	−0.07	1.05
G2	0.01	0.97	−0.02	0.85	−0.04	1.18
E2	0.06	1.04	0.12	0.83	0.15	1.33
C2	−0.02	0.90	−0.03	0.79	−0.02	1.12

TABLE V. Confidence Intervals Around Estimated Factor Scores

MZ G: ± 0.8132	DZ G: ± 0.9798
E: ± 0.9298	E: ± 1.0992
C: ± 0.8398	C: ± 0.8324

twin pairs. The twins and their parents participated in a larger ongoing project on genetic aspects of cardiovascular risk factors. Addresses of twin pairs living in and outside Amsterdam were obtained from City Council population registries. Zygosity was determined by blood typing, and in four cases also by DNA fingerprinting [Jeffreys et al., 1985].

Blood pressure data were obtained when subjects visited the laboratory, where testing took place in a sound attenuated cabin. Subjects were seated in a comfortable chair and were asked to relax as much as possible. Blood pressure was recorded three times during two 8.5 min resting periods with at least 1 hr before the first measurement and with at least 1 hr in-between resting periods. Systolic (SBP) and diastolic blood pressure were measured by the Dinamap 845XT using osillometric technique.

Table VI gives means and standard deviations for the MZ and DZ groups for the six SBP readings. In both groups there was a positive correlation of around 0.3 between SBP and body weight, so weight was included in the model. Model fitting was done separately for the two resting periods in which SBP was measured. Factor scores were also estimated separately for both conditions so that the correlation between factor scores for the two replications could be computed.

For both conditions, a model where the covariance between weight and the three SBP measures was explained by a genetic factor (G) and where the SBP measures had loadings on the nonshared (E) and common environmental (C) factors gave a reasonable fit to the data. All unique variance for SBP was environmental. For weight there was a common environmental component (Cw). Subsequent analyses showed this component to be completely explained by the relationship of body weight with age ($r = 0.6$). In the first condition, genetic heritability for SBP is around 10%, while it is somewhat higher in the second condition (20%), and in both conditions common environment explains a substantial part of the variance. Next in Table VI are the correlations between factor scores from the first and second condition. The correlation is high for the genetic scores and relatively low for the nonshared environmental scores. Finally, Table VI gives the confidence intervals around the factor scores.

Looking at some individual subjects, extreme genetic factor scores were observed in a pair of DZ twins (-2.0 and -2.1) who had average SBP scores of 15 and 16 mm Hg below the group mean. The highest common environmental deviation (2.03) was seen in a pair of DZ twins with average scores of 28 mm Hg and 15 mm Hg above the group mean. The highest nonshared environmental factor score (2.2) was observed for a member of a MZ twin pair, with an average deviation of 25.5 mm Hg, while his cotwin had a blood pressure of 10.7 mm Hg above the mean and showed a nonshared environmental score of -0.27. In yet another subject, his high blood pressure (20 mm Hg above the mean) was due to a high genetic score of 1.91.

TABLE VI. Analysis of Weight and Systolic Blood Pressure (SBP) in 34 MZ and 25 DZ Male Adolescent Twin Pairs

Means and SD for weight (kg) and SBP (mm Hg)

	MZ	DZ
Weight	60.6 (11.11)	62.6 (9.57)
SBP1	122.4 (10.63)	121.9 (12.31)
SBP2	118.4 (9.96)	119.6 (11.49)
SBP3	116.8 (9.12)	119.3 (10.16)
SBP4	123.7 (10.46)	122.7 (11.47)
SBP5	119.4 (8.86)	119.5 (9.14)
SBP6	118.4 (10.74)	119.9 (10.84)

Model fitting: estimated factor loadings and unique variances (standard errors in parentheses)

Rest 1	G	E	C	CWeight	U^2
Weight	7.84 (1.3)	—	—	5.94 (2.0)	6.99 (1.7)
SBP1	3.63 (1.5)	4.96 (1.0)	7.03 (1.3)	—	39.94 (7.3)
SBP2	3.04 (1.4)	7.45 (1.1)	6.28 (1.2)	—	7.52 (9.7)
SBP3	2.70 (1.2)	4.83 (1.0)	5.02 (1.2)	—	34.22 (5.9)

$\chi^2 = 62.22$ (df = 57, $P = .296$)

Rest 2	G	E	C	CWeight	U^2
Weight	7.26 (1.2)	—	—	6.84 (1.5)	6.94 (1.7)
SBP4	4.98 (1.5)	4.50 (0.9)	6.57 (1.3)	—	29.26 (6.0)
SBP5	4.01 (1.2)	4.63 (0.8)	4.73 (1.2)	—	20.56 (4.5)
SBP6	4.93 (1.5)	6.78 (1.2)	4.10 (1.6)	—	26.74 (9.9)

$\chi^2 = 61.82$ (df = 57, $P = .308$)

Correlations between factor scores of Rest1 and Rest2

G: 0.98 E: 0.63 C: 0.79

Confidence intervals

Rest 1	MZ G: ± 0.989	DZ G: ± 1.202
	E: ± 1.157	E: ± 1.193
	C: ± 1.228	C: ± 1.225
Rest 2	MZ G: ± 1.075	DZ G: ± 1.269
	E: ± 1.285	E: ± 1.371
	C: ± 1.268	C: ± 1.258

DISCUSSION

Confirmatory factor analysis of family data can be seen as consisting of two steps: estimation of heritabilities and determining a weight matrix that can be applied to individual data to obtain genetic and environmental scores. The weight matrix obtained in a representative sample can be applied to the general population, and genetic and environmental factor scores can be used to investigate their relationship with other variables that are measured at the phenotypic level. It would thus be possible to estimate, in large groups of unrelated individuals, the genetic covariance between a measured phenotype and a latent genetic factor. Results obtained in a twin study can be generalized to the general population by estimating factor scores for family members, for example, from the regression formula for DZ twins. Even better would be, of course, to include ordinary siblings and parents in an analysis of twins. When no information on genetically related individuals is available, the Bartlett estimator or a regression estimator for individual subjects can be used.

The structural model discussed in this paper can be easily generalized to include unique genetic and environmental factors in addition to the factors that are common to all observed variables [cf., Martin and Eaves, 1977]. However, as there is only a single phenotype associated with these unique factors, the estimated factor scores will become highly correlated, but can still be obtained. In order to increase their reliability, it is necessary to use multiple phenotypic indicators.

Estimation of individual genetic and environmental scores is also possible in a longitudinal analysis using Kalman filtering techniques [Sage and Melsa, 1971]. This makes it possible to estimate changes in factor scores over time at the level of individual subjects and to obtain individual genetic and environmental developmental profiles.

We have shown [Molenaar and Boomsma, 1987; Molenaar et al., 1990] how estimates of certain joint moments of the distribution of true factor scores can be applied to the analysis of genotype by environment interaction, and we have demonstrated in this paper that these scores may be estimated reliably.

Knowledge about the reasons why certain subjects exhibit high phenotypic scores may be both of theoretical and practical interest. Risk assessment may be improved by the knowledge that a high phenotypic score is caused by a high genetic or a high environmental deviation. It is also conceivable that the power of certain types of analyses will increase when a distinction between high genetic and high environmental scores can be made. The use of factor scores depends, however, on a good fitting multivariate model. It is thus necessary to obtain several reliable indicators of the underlying latent construct we want to measure.

REFERENCES

Bartlett MS (1937): The statistical conception of mental factors. Br J Psychol 28:97–104.

Boomsma DI, Molenaar PCM (1987): The genetic analysis of repeated measures. I: Simplex models. Behav Genet 17:111–123.

Dolan CV, Molenaar PCM, Boomsma DI (1989): LISREL analysis of twin data with structured means. Behav Genet 19:51–62.

Heath AC, Neale MC, Hewitt JK, Eaves LJ, Fulker DW (1989): Testing structural equation models for twins using LISREL. Behav Genet 19:9–36.

Jeffreys AJ, Wilson V, Thein SL (1985): Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73.

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

Lawley DN, Maxwell AE (1971): "Factor Analysis as a Statistical Method." London: Butterworths.

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

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

Molenaar PCM, Boomsma DI, Neeleman D, Dolan CV (1990): Using factor scores to detect G × E origin of "pure" genetic or environmental factors obtained in genetic covariance structure analysis. Genet Epidemiol. This issue.

Sage AP, Melsa JL (1971): "Estimation Theory with Applications to Communications and Control." New York: McGraw-Hill Book Company.

Thurstone LL (1935): "The Vectors of the Mind." Chicago: University of Chicago Press.

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