Multivariate Extensions of a Biometrical Model of Twin Data

DW Fulker

In this paper I would like to discuss some of the problems involved in the multivariate analysis of twin data and describe a maximum likelihood approach that goes some way towards solving them.

Partitioning phenotypic covariance into genetic and environmental components for the purpose of investigating their structure was first suggested by Tukey in 1951. He made the point, novel at the time, that just as mean squares in analysis of variance could be partitioned into components, mean cross products in analysis of covariance could also be partitioned in a completely analogous fashion. Subsequently the structure of these component matrices could be explored to indicate how genetic and environmental influences caused measures to become associated.

The idea of looking at twin data in this way appears to have been first suggested by Kempthorne and Osborne in 1961, ten years later, without reference to Tukey's original discussion of the problem. However, in recognition of his contribution, and because his illustrative example involving four protein levels in single crosses of maize demonstrates the multivariate approach so well, I would like to discuss it briefly by way of introduction.

The analysis is the simplest possible, following a one-way analysis of variance, and is shown in Table I. The top part of the table shows the two 4×4 matrices of mean cross products, one between crosses (B_{ij}), the other within (W_{ij}). Subscripted matrix notation will be used wherever possible in order to emphasize the simple one-to-one or element-for-element correspondence that frequently exists between observed covariances and multivariate models. The estimate of the environmental covariance matrix, $\hat{E}_{ij}$, is given directly by W_{ij}, while the initial estimate of the genetic component ${}_1\hat{G}_{ij}$, is given by one-quarter of the difference between B_{ij} and W_{ij}, there being four replications.

Twin Research: Psychology and Methodology, pages 217–236

TABLE I. ANCOVA of Four Proteins in Maize [Tukey, 1951] *

	Between crosses (8 df)					Expectations				
B_{ij}	860.4	178.9	146.3	86.2		$B_{ij} = E_{ij} + 4G_{ij}$				
		43.3	33.4	17.9		$W_{ij} = E_{ij}$				
			30.8	13.1						
				9.4						
	Within crosses (27 df)					Correlations				
W_{ij}	49.1	10.4	5.6	2.3		RE_{ij}	1.00	0.35	0.52	0.41
		6.0	0.4	0.6				1.00	0.10	0.31
			2.4	0.2					1.00	0.17
				0.6						1.00
$_1\hat{G}_{ij}$	202.8	42.1	35.2	21.0		RG_{ij}	1.00	0.97	0.94	0.99
		9.3	8.3	4.3				1.00	1.01	0.95
			7.1	3.2					1.00	0.81
				2.2	z_i					1.00
$_2\hat{G}_{ij} = z_i z_j$	202.8	43.4	35.7	20.1	14.24	RG_{ij}	1.00	1.00	1.00	1.00
		9.3	7.7	4.3	3.05			1.00	1.00	1.00
			6.3	3.6	2.51				1.00	1.00
				2.0	1.41					1.00

*Note: B_{ij}, between crosses mean cross products; W_{ij}, within crosses mean cross products; $_1\hat{G}_{ij}$, first estimate of genetic covariance matrix; $_2\hat{G}_{ij}$, second estimate of genetic covariance matrix; z_i, first principal component of G_{ij}; RE_{ij}, correlation matrix corresponding to E_{ij}; and RG_{ij}, correlation matrix corresponding to G_{ij}.

The structures of the genetic and environmental matrices are quite different, as can be seen from their corresponding correlation matrices, RG_{ij} and RE_{ij}, shown to the right of the table. The environmental correlations are quite small. Evidently, so far as the environment is concerned, the four proteins are to a large extent independently determined, their levels responding differently to particular conditions of soil and climate. The genetic covariance matrix, on the other hand, shows a very high degree of association, the correlations hardly differing from unity. In contrast to the environment, all four protein levels appear to be determined by a single, in this case, genetic system.

When all the values in a correlation matrix approach 1, its unitary structure can be expressed as a single factor or principal component. Tukey estimated the four loadings, $\hat{z}_i$, of the first principal component from his initial estimate of the genetic component matrix, ${}_1\hat{G}_{ij}$, forming a new estimate, ${}_2\hat{G}_{ij}$, as the corresponding products of these loadings, $\hat{z}_i\hat{z}_j$. This component is shown in the column at the foot of the table and accounts for over 99% of the initial estimate of genetic variance. It appears, by eye, to fit the data very well, providing a more parsimonious explanation of the genetic covariance structure and lending support to the hypothesis of a single genetic system.

This direct structural approach to the analysis of component matrices bypasses the main difficulty of the additive approach that estimates components by taking the difference between matrices of observed mean cross products. In the additive approach, variances may take impossible negative values so that correlations cannot be estimated and even when variances are positive, correlations may fall well outside the range of ± 1. Structural analysis of such matrices, which are the rule rather than the exception, may prove impossible.

However, component matrices based on principal components behave exactly like those based directly on paired observations, their latent roots all being $\geqslant 0$. Matrices conforming to this constraint are sometimes referred to as Gramian and present no problems for further structural analysis, should this be deemed necessary. The problem is, however, how many principal components are needed to account for any particular set of data adequately, and how do we estimate them?

With characteristic insight, Tukey indicated that a multivariate extension of the F ratio might be developed to establish the required rank of a component matrix. In the same year Bartlett [1951] published just such a procedure for use in discriminant analysis. This procedure was that later adopted by Bock and Vandenberg [1968] to obtain constrained estimates of the genetic covariance matrix from twin data.

Their method uses the within-pair cross product matrices for dizygotic (DZ) and monozygotic (MZ) twins, DZW_{ij} and MZW_{ij} or **DZW** and **MZW** in conventional matrix notation. The rank of the genetic covariance matrix is estimated

as the number of latent roots greater than unity in the multivariate analysis of variance determinantal equation developed by Bartlett,

$$| \mathbf{DZW} - \Lambda \mathbf{MZW} | = 0,$$

where Λ is the diagonal matrix of roots. Those less than unity, which roughly correspond to negative roots in the component matrix, are set to unity to form Λ^*. This modified matrix of roots is combined with the discriminant functions, X, to estimate the genetic covariance matrix as

$$(\mathbf{X}^{-1})'(\Lambda^* - \mathbf{I})(\mathbf{X}^{-1}).$$

The estimated matrix will be Gramian and can be subjected to further structural analysis by a variety of established procedures.

The method is very straightforward, economical in terms of computer time, and has recently been shown by Bock and Petersen [1975] to provide a constrained maximum likelihood estimate of G_{ij}. It is the only method currently available that provides an explicit solution to the multivariate problem and has been used successfully in a number of applications [Bock and Vandenberg, 1968; Eaves, 1973; Nance et al., 1974].

However, in spite of its considerable advantages, it does suffer from a number of drawbacks. Firstly, it is wasteful of data. In the case of twins, only the within-pairs information is used, that between pairs being excluded from the analysis. The evaluation of more extensive kinships would be even more wasteful, if possible at all. Secondly, the method cannot deal with more than the simplest basic twin model, as we can see if we consider the model below.

$$\begin{aligned}
MZB_{ij} &= SE_{ij} + 2G_{ij} + 2CE_{ij}, \\
MZW_{ij} &= SE_{ij}, \\
DZB_{ij} &= SE_{ij} + 1\tfrac{1}{2}G_{ij} + 2CE_{ij}, \\
DZW_{ij} &= SE_{ij} + \tfrac{1}{2}G_{ij},
\end{aligned}$$

where SE_{ij} is the matrix of specific or within family environmental effects, CE_{ij} the matrix of common or family environmental effects, and G_{ij} the matrix of additive genetic effects.

This model is familiar enough and I do not wish to consider in detail the assumptions underlying it since they have been discussed elsewhere [Jinks and Fulker, 1970; Fulker, 1974]. Suffice it to say the model assumes only additive genetic variance, that genotype-environment covariance and interaction are

both absent, and that MZ and DZ twins share relevant aspects of their environment to the same extent. These assumptions are frequently called into question; much less often are they demonstrated to be false. Only the last assumption is really critical and to my knowledge it has never been clearly demonstrated to be false in any study. Certainly MZ twins share more experiences than DZ twins, but whether these are experiences relevant to the trait in question is often doubtful. In my view, the model is a reasonable approximation for a number of physical and behavioral traits.

The problem of applying the multivariate analysis of variance approach to this model stems from the necessity to set up a proper F ratio which unambiguously establishes the effect in question. The ratio $(\mathbf{DZW})\,(\mathbf{MZW})^{-1}$ establishes the within-pair genetic variance, but no such ratio exists to cope with the common environmental component CE_{ij}. Neither can we, in the absence of CE_{ij}, combine all four observed mean cross product matrices to provide an overall estimate of the remaining components G_{ij} and E_{ij}.

One approach that promises to go some way towards overcoming these limitations is a maximum likelihood approach similar to that used by Taubman [1977] and that advocated by Martin and Eaves [1977]. This approach allows us to use all the available data efficiently and to explore the component structures systematically using χ^2 tests of significance to arrive at reasonable decisions.

We assume that the raw observations follow a bivariate normal distribution, so that the observed mean cross product matrices follow that of the Wishart. If these k matrices, the between and within matrices of twin analysis in the present case, are denoted S_{kij} or S_k for short, with expectations $E(S_k)$, then the following expression provides a log-likelihood ratio statistic, F, following χ^2 in large samples. If we parameterize $E(S_k)$ in terms of the required model and minimize the function with respect to these parameters, we obtain their maximum likelihood estimates. Only the χ^2 value is sensitive to the sample size, the estimates always being maximum likelihood.

$$F = \sum_{k=1}^{m} N_k \left\{ \ln |E(S_k)| - \ln |S_k| + \mathrm{tr}[S_k E(S_k)^{-1}] - P \right\},$$

where N_k = df of the k-th matrix and P = the number of variates. Functions such as these can now be minimized by a number of optimization routines available as packages through most university computing services. These routines frequently employ the first derivatives to minimize the function and the second derivatives to provide standard errors for the parameters, but numerical methods are usually optionally available to avoid the necessity for explicit differentiation. Supplying the derivatives often improves the efficiency of the routines, and Martin and Eaves [1977] offer these for the factor analysis model. The routines used in the present analyses were made available by CERN [1976], the European Centre for Nuclear Research.

Should we also wish to fit models that are not automatically constrained, the routines include procedures for specifiying rectangular constraints to keep parameters within upper and lower bounds, and more complex constraints may be forced by devising a penalty function in which the χ^2 value is augmented by a continuous positive function of the violated constraint. This function vanishes when the constraint is satisfied, leaving the required χ^2 goodness-of-fit statistic. In passing it is worth noting that the use of the above procedures make the constrained maximum likelihood estimation of components in univariate analyses so straightforward that there would appear to be no longer any justification for using more primitive unconstrained methods such as weighted least squares.

The main problems encountered in this approach are the problem of uniqueness general to multivariate analysis and the special problem of singularity that may result from forcing constraints once we stray from straightforward principal component models. The uniqueness problem has been thoroughly discussed by Jöreskog [1969] in the context of maximum likelihood factor analysis. Generally it may be overcome in orthogonal analyses by fixing certain component or factor loadings to zero so that the solution is unique. A simple transformation may be used to obtain the conventional loadings should these be desired, although the use of zero loadings may well help in exploring the covariance structure. In correlated factor analyses a second-order structure written onto the correlations between the factors can also be made to produce a unique solution. Exactly the same solutions may be adopted in the component analysis approach we are discussing.

The singularity problem caused by forcing simple constraints will usually be dealt with automatically by the minimization routine ensuring convergence, appropriate parameter estimates, and the χ^2 goodness-of-fit statistic. However, the matrix of second derivatives may become singular too, and standard errors unobtainable unless we subsequently fix certain parameter values. Taken together, these problems require that we feel our way in building up suitable models, and I would like to try to give something of the flavor of this approach with two simple examples.

The first example involves twin data collected by Zuckerman at the Institute of Psychiatry in London, using our volunteer twin register. His area of research is arousal and the need to seek stimulation, a characteristic he measures with a sensation-seeking questionnaire.

There are four subscales in the questionnaire, each measuring a different aspect of sensation seeking. One, Disinhibition (Dis) is concerned with seeking release through activities such as party going, social drinking, and sexual activities. Another, Thrill and Adventure Seeking (TAS) is concerned with a liking for dangerous and exciting sports. Experience Seeking (ES) involves novel sensations and unconventional experiences, mainly in the social context, while Boredom Susceptibility (BS) is concerned with intolerance of routine activities and dull,

predictable people. Zuckerman [1974], reviewing a large body of research with the questionnaire, makes a case for individual differences in the average score on these scales, reflecting the need for different optimal levels of stimulation. He presents evidence implicating constitutional factors, such as levels of platelet monoamine oxidase (MAO), which correlate negatively with sensation seeking, gonadal hormones which correlate positively, the orienting reflex in response to novel stimuli, and the evoked cortical response in reaction to intense ones. As one might expect for such measures, there are marked sex and age differences. While young men frequently express a liking for skydiving and wild parties, most elderly ladies do not.

We [Fulker et al, 1976] decided to investigate the possibility of a constitutional basis for sensation seeking through a twin study of same-sex and opposite-sex twins. The questionnaire was mailed to 422 pairs of male and female twins, in the age range of 18 to 52 years. Scores were age-corrected by analysis of covariance and subscale variances standardized to unity across the whole sample. The comparability of subscale variances permitted univariate analysis of variance for taking a preliminary look at the the structure of the data, a useful procedure when one is trying to feel one's way with respect to an appropriate genetic and environmental model as well as an appropriate structural one.

Consequently we carried out a repeated measures analysis of variance both between and within pairs with individual differences in total scores providing the between subject mean squares, and differences in subscale profiles the mean squares within subjects. To each we fitted a simple univariate additive genetic model with no common environment, one which has been found to fit a number of personality variables [Jinks and Fulker, 1970; Eaves and Eysenck, 1975; Eaves, 1977]. As we can see from Table II the fit was very good for the total sensation-seeking score, and the narrow heritability of 58% was quite high for a personality variable. However, the repeated measures analysis to the right of the table indicates that this simple model was inadequate for the trait profiles, the residual χ^2 being highly significant. The main cause of the problem is not difficult to see, it being the low between pairs mean square for opposite sex DZ twins (DZ_{os}) which is virtually the same size as the mean square within, indicating zero resemblance for these pairs in spite of a moderate degree of resemblance for same-sex DZ pairs. This pattern suggests either different genes are controlling the profiles in men and women, or a form of sex x genotype interaction, the two being formally equivalent. A simplified form of sex interaction model is shown in Table II, apparently accounting for the data very well. Evidently there is a strong common genetic component in mean level of sensation seeking, but the genetic determination of the pattern that goes to make up a particular level is under different genetic control in the two sexes.

With a reasonable univariate genetic and environmental model we felt confident in fitting a multivariate extension to the ten 4 × 4 mean cross products

TABLE II. ANOVA of Sensation Seeking

		Total score analysis				Profile analysis						
				Model 1				Model 1		Model 2		
Twin	Item	df	MS	V(SE)	V(G)	df	MS	V(SE)	V(G)	V(SE)[a]	V(G)	V(G × S)
MZf	B	172	2.96	1	2	516	0.91	1	2	1	2	2
	W	174	0.83	1	0	522	0.37	1	0	1	0	0
MZm	B	57	3.57	1	2	171	1.17	1	2	1	2	2
	W	59	0.82	1	0	177	0.47	1	0	1	0	0
DZf	B	110	2.08	1	1½	330	0.80	1	1½	1	1½	1½
	W	112	1.35	1	½	336	0.47	1	½	1	½	½
DZm	B	24	2.30	1	1½	72	0.99	1	1½	1	1½	1½
	W	26	1.51	1	½	78	0.51	1	½	1	½	½
DZ_{os}	B	49	3.91	1	1½	147	0.64	1	1½	1	1½	½
	W	50	1.55	1	½	150	0.76	1	½	1	½	1½
Estimates				0.83	1.16			0.41	0.29	0.41	0.18	0.43
χ^2				140.59*	23.36*			420.25*	33.64*	402.25*	2.68	15.28*
Residual $\chi^2_8 = 8.59$								$\chi^2_8 = 29.56$*		$\chi^2_7 = 8.52$		

*$P < 0.001$.

[a]V(SE) in model 2 includes SE × sex interactions.

matrices calculated for the five twin groups. The form of the data, omitting the last eight matrices for convenience, together with the model, are shown in Table III.

Maximum likelihood estimates of the three component matrices are shown in Table IV. Initially, an unconstrained estimation procedure was used to see if constraints were necessary. The estimates of SE_{ij} and G_{ij} did, in fact, turn out to be Gramian, but the estimated SG_{ij} did not, one estimated correlation being −2.14. Consequently, a penalty function approach was used to obtain a constrained estimate of SG_{ij}, with the result shown at the foot of the table.

The form of the penalty function was

$$P = Q\sum_{i=1}^{n} \lambda_i^2.$$

The λ_i are the n negative latent roots of the estimated component matrix SG_{ij} at any given time during the minimization, and Q is an arbitrarily large constant modified as the minimization proceeds. The function minimized is the original maximum likelihood function plus P. By starting in a feasible region, where all the λ are positive, making Q progressively larger, and setting a very small limit to λ, most minimization routines, even using gradient methods, will find a satisfactory minimum.

In the present case the constrained estimate of SG_{ij} has three positive roots, the remaining one being nought. Such a matrix has only nine free elements, not ten as in the full rank case. With only nine free parameters one degree of freedom is therefore lost from the residual χ^2 giving a nonsignificant difference χ_1^2 of 1.47. Clearly the constraint not only produces a sensible estimate of SG_{ij} but is also fully consistent with the data. In either case, constrained or unconstrained, the residual χ^2 values indicate a very good fit of the interaction model.

The structure of these constrained component matrices could be explored successfully by any conventional multivariate technique. By inspection their form indicates a general factor for G_{ij}, all the correlations being positive, and a bipolar factor for SG_{ij} in view of both negative and positive correlations. This pattern is consistent with the univariate analysis in which the total score corresponds to the general factor controlled by additive genes, and the profile scores, which involve orthogonal contrasts with plus and minus signs, correspond to the bipolar factor controlled by different genes in the two sexes.

The exploratory approach to factor structures is difficult to combine with significance testing, especially in component analysis. Consequently, the direct structural approach involving a model consisting of factor loadings and specific variances fitted directly to the data was employed, but in a progressive manner to allow for tests of significance of successive aspects of the model. However,

TABLE III. Multivariate Model and ANCOVA: Sensation Seeking*

Twin	Item	Expectation	df	Mean cross products			
				Dis	TAS	ES	BS
MZf	B	$SE_{ij} + 2G_{ij} + 2SG_{ij}$	172	1.35	0.38	0.69	0.70
					1.49	0.64	0.24
						1.56	0.42
							1.29
MZf	W	SE_{ij}	174	0.47	0.12	0.17	0.16
					0.58	0.09	0.11
						0.42	0.04
							0.45
MZm	B	as above	57				–
	W		59				–
DZf	B	$SE_{ij} + 1½G_{ij} + 1½SG_{ij}$	110				–
	W	$SE_{ij} + ½G_{ij} + ½SG_{ij}$	112				–
DZm	B	as above	24				–
	W		26				–
DZ_{os}	B	$SE_{ij} + 1½G_{ij} + ½SG_{ij}$	49				–
	W	$SE_{ij} + ½G_{ij} + 1½SG_{ij}$	50				–

*SE_{ij}, specific environmental covariance matrix; G_{ij}, additive genetic covariance matrix; SG_{ij}, additive genetic × sex interaction covariance matrix. Sensation-seeking scales: Dis, Disinhibition; TAS, Thrill and Adventure Seeking; ES, Experience Seeking; BS, Boredom Susceptibility.

TABLE IV. Solution to ANCOVA Sensation Seeking

	Components								
	Dis	TAS	ES	BS		Correlations			Residual χ^2
SE_{ij}	0.46	0.13	0.16	0.16	1.00	0.26	0.36	0.33	
		0.56	0.10	0.07		1.00	0.20	0.13	
			0.44	0.05			1.00	0.11	
				0.50				1.00	
G_{ij}	0.88	0.18	0.41	0.57	1.00	0.26	0.42	0.71	
		0.55	0.46	0.05		1.00	0.60	0.08	
			1.08	0.67			1.00	0.75	
				0.73				1.00	
SG_{ij}	0.17	−0.03	0.10	0.05	1.00	−0.12	0.82	0.23	
		0.36	0.06	0.11		1.00	0.50	0.34	$\chi^2_{70} = 73.62$ ns
			0.04	−0.23			1.00	−2.14	
				0.29				1.00	
SG_{ij}	0.15	−0.05	0.06	0.03	1.00	−0.22	0.44	0.13	
constrained		0.35	0.03	0.09		1.00	0.22	0.39	$\chi^2_{71} = 75.09$ ns
			0.15	−0.16			1.00	−0.70	
				0.34				1.00	

no structure beyond the ten SE_{ij} was fitted to the environmental component since these influences are confounded with sex χ environmental interaction in these data.

The results are shown in Table V together with an approximate analysis of χ^2. The addition of each of the two factors and the specific variances produces a large reduction in χ^2, establishing the statistical significance of all these structural components. The full model in line 4 of the table provides a nonsignificant residual $\chi^2_{78} = 85.45$ indicating a satisfactory fit. The differences between this model and the full unconstrained covariance model shown in the final line of the table is not significant ($\chi^2_8 = 11.73$), indicating that the reduced rank model, involving 12 genetic parameters, explains the data as well as the full rank model involving 20. The pattern of loadings for the additive genetic component z_1, given at the foot of the table, confirms the presence of a strong general factor common to men and women, and the bipolar pattern of loadings for the interactive component p_i confirms the sex difference in the genetic determination of trait profiles. Genes that make for high TAS and ES and low Dis and BS in men appear to produce opposite effects in women.

TABLE V. Multivariate Model of Sensation Seeking†

Model	Resid χ^2	df	Diff χ^2	df
SE_{ij}	346.48*	90		
SE_{ij} ; $G_{ij} = z_i z_j$	242.26*	86	104.22*	4
SE_{ij} ; $G_{ij} = z_i z_j$; $S \times G_{ij} = p_i p_j$	163.01*	82	79.25*	4
SE_{ij} ; $G_{ij} = z_i z_j + (s_i^2$ when $i = j)$; $S \times G_{ij} = p_i p_j$	85.45	78	77.56*	4
SE_{ij} ; G_{ij} ; $S \times G_{ij}$	73.62	70	11.73	8

	Estimate of Genetic Parameters		
	Additive		Additive × sex
	z_i	s_i	p_i
Dis	0.63	0.75	0.27
TAS	0.45	0.78	−0.27
ES	0.93	0.36	−0.35
BS	0.71	0.38	0.59
Genetic variance	49%	36%	15%

*$p < 0.001$.

†z_i, Loadings of general additive genetic factor; s_i, specific variances of additive genetic factor; p_i, loadings on additive × sex interaction factor. Resid, Residual; Diff, Difference, here and in Table VI.

Since the former traits involve socially acceptable forms of activity, and the latter include activities which are less socially acceptable, it is perhaps not surprising, even today, that genes controlling these measures should express themselves differently in men and women.

One further analysis was carried out to see if the structure of the environmental component was similar to the genetic component, even given the confounding of sex interaction effects with those of the specific environment. The DZ_{os} twins were dropped from the analysis and a simple SE_{ij}, G_{ij} model fitted to the twins of the same sex, these parameters being understood to represent a main effect confounded with sex interaction, in accordance with the expectations in Table III. In order to test the equality of the two covariance structures, G_{ij} was reparameterized as a weighted composite of SE_{ij},

$$G_{ij} = w_i w_j SE_{ij}.$$

If this model should fit, identical correlation structures for G_{ij} and SE_{ij} are implied.

The results are shown in Table VI. Clearly the hypothesis of equality of correlational structures is supported, especially if we allow for specific variation,

TABLE VI. Testing Genetic and Environmental Correlations Having Same Structures: Sensation Seeking†

Model for like-sexed twins	Resid χ^2	df	Diff χ^2	df
SE_{ij}; $G_{ij} = w_i w_j SE_{ij}$	82.82	66		
SE_{ij}; $G_{ij} = w_i w_j \{ SE_{ij} - (s_i^2 \text{ when } i = j) \}$	64.63	62	18.19*	4
SE_{ij}; G_{ij}	61.61	60	3.02	2

	First two orthogonal components					
	1st	2nd	Resid	s_i^2	Error variance reported	$h^2 = w_i^2/(1 + w_i^2)$
Dis	0.84	0.37		0.00	0.19	53%
TAS	0.56	−0.63		0.19	0.34	55%
ES	0.85	−0.33		0.30	0.23	80%
BS	0.74	0.43		0.36	0.38	79%
Genetic variance	58%	22%	20%			

*$p < 0.002$.

†$w_i w_j$, the relative weight of G_{ij} to SE_{ij} element for element; s_i^2, variance specific to SE_{ij} when $i = j$.

s_i^2, in the environmental component. The values of s_i^2 are similar to the error variation quoted for these tests and probably represent the same source of variation. Analysis of this common genetic and environmental component reveals a very similar general and bipolar structure to that previously found. This finding may indicate that genetic and environmental influences for these measures have similar underlying mechanisms.

Next I would like to consider a rather different example involving a reanalysis of Taubman's [1977] twin study of schooling, income, and occupational status. This study probably represents the most extensive and sophisticated example of structural genetic component analysis available to date, as well as being of great substantive interest.

One question raised by his analysis with Behrman and Wales [Behrman et al, 1977] was why a model involving only one environmental component was used throughout. Inspection of their table of cross-sib correlations for MZ and DZ twins indicated a common environmental covariance matrix (estimated as twice the DZ correlations minus the MZ correlations, in the conventional manner) with one negative variance, three undefined correlations, and three correlations greater than 1. This component was clearly far from Gramian in form, strongly suggesting the necessity for a reduced rank model of common environment.

However, these effects might simply have been the result of sampling variation, so the MZ and DZ correlations were converted to mean cross products and subjected to the constrained maximum likelihood estimation procedure. Since Taubman's original analysis had the form of a kind of path analysis, it was decided to reparameterize the component covariance matrices as variances and correlations to facilitate further investigation. For example, G_{ij}, was replaced by $RG_{ij}G_i^{½}G_j^{½}$, where the RG_{ij} are the correlations and the G_i factors the variances. The Gramian constraint was ensured in two ways. Firstly, rectangular constraints were applied to all the parameters so that the variances could not become negative and the correlations were bounded by ± 1. Secondly, the latent roots of the covariance matrices were all required to be $\geqslant 0$ by means of a penalty function.

The estimated component correlation matrices, together with the components of variances as percentages, are shown in Table VII. The model fits quite well, especially when we consider the power of the test with sample sizes around 4,000. The choice of a single variable for common environment appears to be forced by the data exactly as it was in Tukey's analysis, the unitary values in the correlation matrix indicating the necessity of a single rank model. There appears to be only one common environmental influence general to schooling and subsequent adult status and income, an influence which it seems plausible to equate with the environmental effects of social origins.

One problem with the simple twin model, as soon as we wish to include common environment, is that we can no longer be sure that the effects of nonadditive gene action, assortative mating, and the correlated genetic and environmental influences

TABLE VII. Constrained Parameter Estimates: Taubman's Twin Study [1977]†

	Correlations				Variances (%)
			Specific Environment (SE)		
Schooling (S)	1.00	0.17	0.24	0.10	23
Occupation 1 (O_1)		1.00	0.17	0.07*	48
Occupation 2 (O_2)			1.00	0.14	64
Income (I)				1.00	45
			Genetic (G)		
	1.00	0.60	0.62	0.55	46
		1.00	0.63	0.52	33
			1.00	0.44	28
				1.00	47
			Common Environment (CE)		
	1.00	1.00	1.00	1.00	31
		1.00	1.00	1.00	19*
			1.00	1.00	8*
				1.00	8*

*$p < 0.01$; for all others $p < 0.001$.
†Residual $\chi^2_{16} = 24.99$, $p < 0.1$.

are absent. If present, they bias the estimate of common environment, decreasing it in the case of nonadditivity and being completely confounded with it in the other two cases [Jinks and Fulker, 1970; Fulker, 1974].

Taubman attempts to explore these problems by freeing the DZ genetic correlations ρ from the value of 0.5, which the simple twin model assumes, allowing it to take its own value in the estimation procedure. He found a value of about 0.35 to be consistent with the data, implying considerable nonadditive genetic variation, although a distinct possibility in this particular study was a restricted sampling of family influences. However, with his approach the estimate of common environment is still confounded with assortative mating which, in turn, will force the genetic correlation to become unrealistically low. In addition a single value of ρ might not be realistic if the degree of assortative mating and nonadditive gene action should differ between measures, as might well be true.

To explore these possibilities, the following model was adopted which allowed for a separate sib genetic correlation for each measure (ρ_i):

$$\begin{aligned}
MZB_{ij} &= SE_{ij} + 2G_{ij} + 2CE_{ij}, \\
MZW_{ij} &= SE_{ij}, \\
DZB_{ij} &= SE_{ij} + (1 + \rho_i^{1/2}\rho_j^{1/2})G_{ij} + 2CE_{ij}, \\
DZW_{ij} &= SE_{ij} + (1 - \rho_i^{1/2}\rho_j^{1/2})G_{ij}.
\end{aligned}$$

Parameters estimated by the constrained procedure are given in Table VIII for three different models. Model 1 is the simple model previously fitted with all values of ρ_i fixed at 0.5 and a single common environment. That is, the conventional twin model fits quite well. Model 2 frees the four ρ values but retains a single common environment. This model gives a slightly better fit (the difference χ^2_4 being 9.64, $P \simeq 0.05$), and some of the values of ρ_i are improbably low. The similarity of the estimates in all other respects, together with the merely modest improvement in fit, clearly indicates the data are relatively insensitive to the values of the sib genetic correlation. Put differently, the joint test that the ρ values differ from 0.5 barely reaches the 0.05 probability level.

In both these models the effects of assortative mating are still confounded with common environment, even though ρ has taken account of nonadditive gene action. However, if we drop common environment from the model but still keep the ρ values free, to give model 3, the effect of assortative mating is accommodated by the ρ values in addition to the effect of nonadditive gene action. These additional effects may be seen in the increased estimates of ρ. Now, though, the model fails quite badly, the residual χ^2_{16} being 40.63, $P < 0.001$. Clearly a model assuming no common environment is quite unrealistic, even allowing for assortative mating, nonadditivity and possible restricted sampling. Since if there is at least some common environment, its component must be at least rank one, and something between model 1 and 2 would seem to be required by the data. As the χ^2 values and the estimates of the variances differ only slightly between these two models, parsimony favors the conventional model 1. Probably, as appears to be true for IQ [Jinks and Fulker, 1970], assortative mating and nonadditivity balance out, making a ρ of 0.5 quite a realistic assumption.

Bearing in mind the limitation that common environmental effects may still include some effects of correlated genes and environment, we can explore the structure of the component variances and correlations in model 1 further (shown in Table VII) by means of the modified path model shown in Figure 1.

In this model only three of the four variables have been selected for analysis since they can be plausibly related longitudinally. These are schooling (S) and the two measures relating to the individual some 30 years later, namely, occupational status (Oc_2) and income (Inc). On the left of the figure are the three influences G, SE, and CE that affect schooling. These influences are also assumed to affect income and status some 30 years later. However, in addition income and status are assumed to be influenced by the residual genetic and environmental effects shown in the right of the figure. No residual common environment is needed in view of the rank one structure of this component. Following the conventions of path analysis [Wright, 1954], the casual influences of the seven latent variables on the three measures S, Inc, and Oc_2 are represented by straight arrows bearing the path coefficients that indicate their relative influence when all the other

TABLE VIII. Taubman's [1977] Data: Different Sib Genetic Correlations

		% Variances						
		SE	G	CE	ρ_i	χ^2	df	P
Model 1 all $\rho_i = 0.5$, $RCE_{ij} = 1.0$	S	23	46	31	0.50	24.99	16	< 0.1
	O_1	48	33	19	0.50			
	O_2	64	28	8	0.50			
	I	45	47	8	0.50			
Model 2 all ρ_i free, $RCE_{ij} = 1.0$	S	18	45	37	0.19	15.35	12	ns
	O_1	36	46	18	0.33			
	O_2	50	40	11	0.15			
	I	31	61	8	0.42			
Model 3 all ρ_i free, no CE	S	13	87	0	0.67	40.63	16	< 0.001
	O_1	30	70	0	0.68			
	O_2	47	53	0	0.56			
	I	29	71	0	0.55			

[a] ρ_i, Genetic sib correlation for i-th measure.

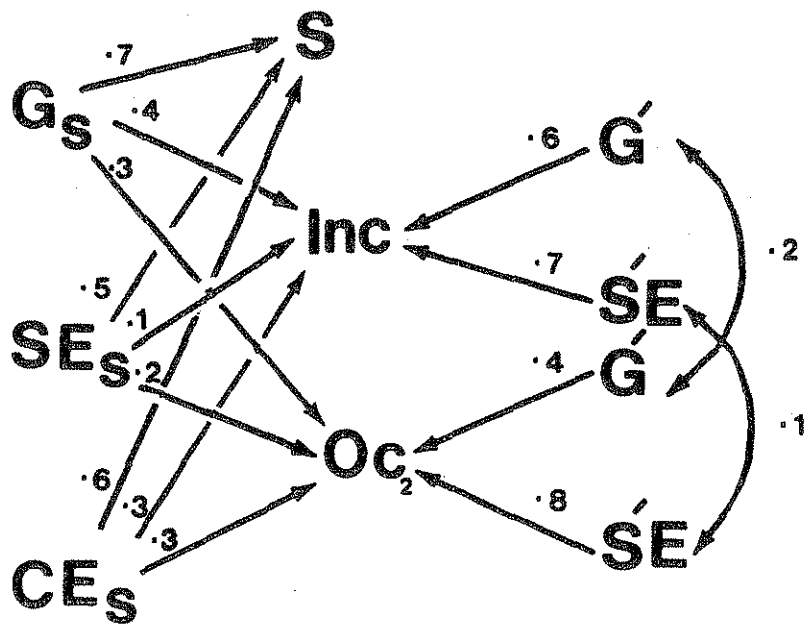

Fig 1. Path analysis of Taubman's [1977] twin study. Latent variables: G_S, genetic influences on schooling; SE_S, specific environmental influences on schooling; CE_S, common family influences on schooling; G^1, residual genetic influences; SE^1, residual specific environmental influences. Observed variables: S, schooling (years); Inc, adult income (log $); Oc_2, adult occupational status.

factors in the system are held constant. The relationships between the residual effects are represented by curved arrows simply indicating the existence of a correlation. Coefficients have been rounded to one decimal place for simplicity.

This diagram indicates aspects of the genetic and environmental influences on adult status and income, not all of which are obvious from simple inspection of the correlations and variances given in Table VII. Most of these conclusions follow from Taubman's analysis too, but the path diagram has the advantage of providing a convenient summary.

Firstly, both genes and common environment for schooling subsequently influence adult status and income, roughly to the same extent, all four paths being between 0.3 and 0.4. Secondly, specific environmental effects on schooling, that is, chance and accidental factors, exert an almost trivial influence later, their paths being between 0.1 and 0.2. Thirdly, by far the greatest influences on adult income and status are residual genetic and specific environmental factors. Fourthly, these strong residual factors are largely independent of each other with respect to the two adult measures, their correlations being merely 0.1 and 0.2.

This analysis suggests, then, that insofar as schooling influences adult status, home environment is almost as important as genetic endowment, but that large independent genetic and environmental influences unrelated to home environ-

ment play the major role. One could hazard a guess that these later genetic influences are related more to temperament and special skills than to IQ, which we know has a powerful influence on schooling. The environmental factors probably relate to market imperfections and luck.

These two examples have, I hope, indicated something of the scope of the maximum likelihood approach to the multivariate structural analysis of genetic and environmental influences using twins. It could, of course, be extended with no difficulty to include additional kinships. It is possible to handle a variety of models and the method is probably statistically optimal. Its only drawback seems to be the demands it makes on computer time, and the development of more efficient algorithms geared to the needs of particular problems can be expected to remove this limitation.

ACKNOWLEDGMENTS

I would like to thank Owen White for generously sharing with me his extensive knowledge of multivariate analysis and optimization. I would also like to thank Nick Martin and Lindon Eaves for many helpful discussions.

REFERENCES

Bartlett MS (1951): The goodness of fit of a single hypothetical discriminant function in the case of several groups. Ann Eugen 16:199–214.

Behrman J, Taubman P, Wales T (1977): Controlling for and measuring the effects of genetic and family environment in equations for schooling and labour market success. In Taubman P (ed): "Kinometrics: The Determinants of Socioeconomic Success Within and Between Families." Amsterdam: North Holland.

Bock RD, Petersen AC (1975): A multivariate correction for attenuation. Biometrika 62(3): 673–678.

Bock RD, Vandenberg SG (1968): Components of heritable variation in mental test scores. In Vandenberg SG (ed): "Progress in Human Behavior Genetics." Baltimore: Johns Hopkins University Press.

CERN (1974): Minuit; A package of programs to minimise a function of a variable, compute the covariance matrix and find the true errors. Geneva, Switzerland: CERN.

Eaves LJ (1973): The structure of genotypic and environmental covariation for personality measurements: An analysis of the PEN. Br J Soc Clin Psychol 12:275–828.

Eaves LJ (1977): Inferring the causes of human variation. J Roy Statist Soc A 140:123–146.

Eaves LJ, Eysenck HJ (1975): The nature of extraversion: A genetical analysis. J Pers Soc Psychol 32:102–112.

Fulker DW (1974): Applications of biometrical genetics to human behaviour. In van Abeelen JHF (ed): "The Genetics of Behaviour." Amsterdam: North Holland.

Fulker DW, Zuckerman M, Eysenck SB (1976): A genetic and environmental analysis of sensation seeking. Mimeo. Dept of Psychology, Institute of Psychiatry, London.

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

Jöreskog KG (1969): A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34:183–202.

Kempthorne O, Osborne RH (1961): The interpretation of twin data. Am J Hum Genet 13:320–329.

Loehlin JC, Vandenberg SG (1968): Genetic and environmental components in the covariation of cognitive abilities: An additive model. In Vandenberg SG (ed): "Progress in Human Behavior Genetics." Baltimore: Johns Hopkins University Press.

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

Nance WE, Nakata M, Paul TD, Yu P (1974): The use of twin studies in the analysis of phenotypic traits in man. In Janevich DT, Skalko RG, Porter IH (eds): "Congenital Defects. New Directions in Research." New York: Academic Press.

Taubman P (ed) (1977): "Kinometrics: The Determinants of Socioeconomic Success Within and Between Families." Amsterdam: North Holland.

Tukey JW (1951): Components in regression. Biometrics 7:33–69.

Wright S (1954): The interpretation of multivariate systems. In Kempthorne O, Bancroft TA, Gowen JW, Lush JL (eds): "Statistics and Mathematics in Biology." Ames, Iowa: Iowa State College Press.

Zuckerman M (1974): The sensation seeking motive. In Maher B (ed): "Progress in Experimental Personality Research." London: Academic Press.