A Framework for Structural Equation Models in General Pedigrees

Nathan J. Morris Robert C. Elston Catherine M. Stein

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

Key Words

Latent variable analysis • Path analysis • Extended pedigrees • Complex traits • Genetic linkage analysis • Genetic association

Abstract

Background/Aims: Structural Equation Modeling (SEM) is an analysis approach that accounts for both the causal relationships between variables and the errors associated with the measurement of these variables. In this paper, a framework for implementing structural equation models (SEMs) in family data is proposed. **Methods:** This framework includes both a latent measurement model and a structural model with covariates. It allows for a wide variety of models, including latent growth curve models. Environmental, polygenic and other genetic variance components can be included in the SEM. Kronecker notation makes it easy to separate the SEM process from a familial correlation model. A limited information method of model fitting is discussed. We show how missing data and ascertainment may be handled. We give several examples of how the framework may be used. **Results:** A simulation study shows that our method is computationally feasible, and has good statistical properties. **Conclusion:** Our framework may be used to build and compare causal models using family data without any genetic marker data. It also allows for a nearly endless array of genetic asso-

ciation and/or linkage tests. A preliminary Matlab program is available, and we are currently implementing a more complete and user-friendly R package.

Copyright © 2011 S. Karger AG, Basel

Introduction

Most diseases of interest to modern genetic epidemiologists are complex both in their etiology and measurement. That is, they result from a complicated interplay of various environmental and genetic factors, and they are subject to fuzzy, noisy and often multidimensional disease definitions. As a result, such diseases may display no obvious Mendelian inheritance patterns in spite of the fact that they display obvious familial aggregation.

Such traits call for a modeling approach that accounts for both the causal relationships between variables and the errors associated with the measurement of these variables. One such modeling paradigm is known as Structural Equation Modeling (SEM). SEM is a merging of two separate modeling techniques. The first technique, known as path analysis, originated with Sewall Wright [1]. Wright represented his model graphically by drawing directed arrows between causally related variables. Each arrow represented a linear causal relationship. More recently, Pearl [2] has given a fascinating defense of the causal meaning

of path analysis. Pearl also showed how path models lead to certain conditional independencies between variables. The second origin for SEM goes back to the method of confirmatory factor analysis [3, 4]. Confirmatory factor analysis is not to be confused with exploratory factor analysis. The latter is in general a hypothesis-generating exercise, and it has inherent difficulties with identifiability, leading to the need for factor rotation methods. Confirmatory factor analysis, on the other hand, begins with a set of hypotheses and seeks to compare them.

Path analysis alone is a flexible tool. For example, linear regression is a special case of path analysis. By combining factor analysis and path analysis, SEM supplies the added ability to deal with the attenuation of regression coefficients when the predictor variables are measured with error [5]. SEM, particularly in its most modern formulations, is extremely general [6]. Some of its particular sub-models are confirmatory factor analysis, classical econometric models, path analysis and linear regression. Some modern modeling methods and software can handle continuous, ordinal and categorical latent and observed traits. Some extensions of SEM can combine generalized linear mixed models, multilevel models and latent growth curve analysis. This generality is perhaps best epitomized by the approach of Skrondal and Rabe-Hesketh [7], which they refer to as ‘Generalized Latent Variable Modeling’.

It should be emphasized up front that while algorithmic model searching and comparison may be useful [8, 9], this paper does not advocate such an approach. Instead, we believe that it is perfectly reasonable to start with a small number of hypothesized models and compare them. There are numerous reasonable paradigms for model comparison as has been discussed, for example, in Bollen and Long [10].

There are also several equivalent ways of formulating a structural equation model (SEM). Perhaps the best known approach is known as ‘linear structural relations’ or LISREL [11]. The now classic book by Bollen [3] gives an excellent overview of this approach. At the heart of the LISREL model are two sets of equations: the measurement and the structural equations. The measurement equations relate a set of unobserved or latent variables to a set of observed variables. The structural equations show how the latent factors affect each other. In this work we show how both these equations may be extended for data collected on general pedigree structures. Furthermore, our work allows the statistical modeler to separate the process of modeling familial correlation from the process of developing a SEM.

Both path analysis and factor analysis have a long history in genetic epidemiology. Besides Wright’s already mentioned historic work, numerous methodological and applied papers have utilized SEM for genetic purposes. For example, the work of Neale and Cardon [12] epitomizes a large body of methodological work for twin studies. In the context of allelic association, there is the work of Medland and Neale [13], Posthuma et al. [14] and van den Oord [15]. Some work such as the work of Posthuma et al. [14] is applicable to nuclear families of varying size.

Perhaps the work most similar to what we present here is the excellent work of Todorov et al. [16] and Gianola and Sorensen [17]. Todorov et al. [16] developed an approach to combining linkage analysis with path analysis. However, their model did not include an explicit measurement equation. Moreover, it is difficult to understand how this model can be applied to general pedigrees. Gianola and Sorensen [17] also developed a Bayesian approach similar to ours. However, as with Todorov et al. [16], they did not explicitly include a measurement equation. Moreover, they considered only a polygenic random effect.

Other works similar in spirit to ours are the ones of Boomsma and Dolan [18], Bauman et al. [19] and Eaves et al. [20]. All these papers suggest how factor models may be used with family data. However, these papers suggest only a measurement equation, and not a structural equation. In our work, we show how both the measurement and structural equations may be implemented for general pedigrees.

Bauman et al. [19] recognized the utility of the Kronecker product for combining exploratory factor analysis and the multivariate variance component (VC) model. They also pointed out that a factor model may be applied to each VC separately. We follow Bauman et al. in emphasizing the benefits of the Kronecker product in our work as well. We note parenthetically that the animal breeding community has used Kronecker notation frequently [17, 21, 22]. We hope the human genetic community will return to using this helpful notation, introduced into the human genetics literature in 1971 [23].

Methods

The Basic Framework

Table 1 gives an overview of some of the notation we use. Suppose we are interested in modeling t observed traits in K pedigrees. Suppose also that there are n_k individuals in each pedigree. Let y_{ki} be a $t \times 1$ vector of observed traits for the i -th individual in the k -th pedigree, and $y_k = [y_{k1}^T \dots y_{kn_k}^T]^T$. Also, let z_{ki} be a $c \times 1$

Table 1. Some model notation

Notation	Meaning
k/K	Pedigree index/total number of pedigrees
i/n_k	Individual index/total number of individuals in pedigree k
j/ν	Variance component index/total number of variance components
t	Number of observed traits
c	Number of covariates
l	Number of latent factors
y_{ki}	$t \times 1$ vector of observed traits
z_{ki}	$c \times 1$ vector of observed covariates
η_{ki}	$l \times 1$ vector of unobserved or latent factors
ε_{kij}	$t \times 1$ vector of measurement errors
s_{kij}	$l \times 1$ vector of random effects for the structural equation
E_j	$\text{Var}(s_{kij}) = E_j$
Γ_m	$t \times c$ matrix relating z_{ki} to y_{ki}
Λ	$t \times l$ matrix relating η_{ki} to y_{ki}
Z_j	$\text{Var}(s_{kij}) = Z_j$
B	$l \times l$ matrix denoting the causal relationship among the latent factors
Γ_s	$l \times c$ matrix relating z_{ki} to η_{ki}

vector of observed covariates, including a 1 in the first position to account for the mean. Let Γ_m be a $t \times c$ matrix of coefficients relating z_{ki} to y_{ki} . Also, $z_k = [z_{k1} \dots z_{kn_k}]^T$. We assume that there exists an $l \times 1$ vector of unobserved or latent factors η_{ki} , and these latent factors are related to y_{ki} by the factor loading matrix Λ . Note that latent growth curve models may be implemented by specifying certain forms for Λ . Let $\eta_k = [\eta_{k1}^T \dots \eta_{kn_k}^T]^T$. We assume that the covariance structure of the error term may be captured by the sum of ν random effects

$$\left(\sum_{j=1}^{\nu} \varepsilon_{kij} \right).$$

For example, we may have single locus additive genetic, polygenic and independent environmental effects. The covariance of each effect is assumed to be Kronecker factorizable, so that $\text{Var}(\varepsilon_{kij}) = \text{Var}[\varepsilon_{k1j}^T \dots \varepsilon_{kn_k j}^T]^T = M_{kj} \otimes E_j$ and $\text{Var}(\varepsilon_{kij}) = E_j$. Table 2 contains a list of some of the possible VCs that meet this criterion and the matrix M_{kj} related to each VC. We then have the measurement model:

$$y_{ki} = \Gamma_m z_{ki} + \Lambda \eta_{ki} + \sum_{j=1}^{\nu} \varepsilon_{kij} \quad (1)$$

Table 2. Some possible variance components

Component	M_{kj}	Notation
Additive genetic	Matrix of expected proportions of alleles shared IBD at a locus given the marker data	Π_k
Dominant genetic	Matrix of probabilities that a pair share two alleles IBD at a locus given the marker data	κ_k
Additive polygenic	Two times the kinship coefficient matrix	Φ_k
Dominance polygenic	Fraternity coefficient	Δ_k
Shared environmental	$n_k \times n_k$ matrix of unities	J_{n_k}
Independent environment	$n_k \times n_k$ identity matrix	I_{n_k}

Now let B be an $l \times l$ matrix relating η_{ki} to itself and Γ_s be an $l \times c$ matrix of coefficients relating z_{ki} to η_{ki} . We assume that the covariance structure of the residual term may be captured by the sum of ν random effects

$$\left(\sum_{j=1}^{\nu} s_{kij} \right)$$

where $\text{Var}(s_{kij}) = \text{Var}[s_{k1j}^T \dots s_{kn_k j}^T]^T = M_{kj} \otimes Z_j$. Finally, we have the structural equations:

$$\eta_{ki} = B \eta_{ki} + \Gamma_s z_{ki} + \sum_{j=1}^{\nu} s_{kij} \quad (2)$$

Each of the matrices Γ_m , Λ , Z_j , B , Γ_s and E_j are selected to have a specific structure. Many of the entries in these matrices will be forced to 0 while others will be free parameters that need to be estimated. It is often simpler to represent the model graphically. Bollen [3] gives a good introduction to how the matrix and graphical representation of SEM are related. The basic intuition behind (2) is that each variable in η_{ki} is influenced by other variables in η_{ki} , both by some fixed covariates and by a set of unknown genetic and environmental factors. Some of these genetic and environmental factors are shared by individuals, leading to the factorized covariance structure discussed above.

There are two types of traits used in SEM: exogenous traits, which are not explained by the model (i.e., they are not affected by other traits), and endogenous traits, which are explained by the model. Readers familiar with the LISREL model will notice that our equations do not distinguish between exogenous and endogenous traits as does the LISREL model. However, this is similar to the approach of Muthén [5], and it represents a generalization of the LISREL model. It is simply important that we allow for a general covariance structure to exist for the exogenous variables. Suppose, for example, that the bottom p traits of η_{ki} are exogenous. Then the bottom p rows of B will be zero, and the lower right $p \times p$ entries in E_j will be free parameters for all $j \in \{1 \dots \nu\}$.

It is also important to note that when we perform SEM we usually believe that most of the covariance between variables can be explained by the causal relationships between those variables. That is, the random genetic and environmental factors are gener-

ally assumed to only affect the entries of η_{ki} one at a time. Thus we force many of the off-diagonal entries in E_j to zero because they are not correlated. When we specify a SEM, we expect that certain conditional independencies will apply, both within the pedigree and on the population level. That is, we often expect to be able to sample unrelated individuals from the population and apply the same SEM with no additional VCs. Thus we would argue that often the same constraints should be placed on E_j for all $j \in \{1 \dots \nu\}$ and that, similarly, the same constraints should be placed on Z_j for all $j \in \{1 \dots \nu\}$. Otherwise, counterintuitive population level results may be encountered. Of course, it may at times be reasonable to break this principle. For example, we may wish to have the major gene VC contribute directly to only one trait and not to others. Similar arguments have appeared in the multilevel modeling literature, as may be seen on p. 582 of Longford and Muthén [24] and on p. 219 of McDonald and Goldstein [25].

By using the Kronecker notation it becomes fairly simple to separate out the development of the SEM part of the model from the modeling of the familial correlations. A statistician may therefore put together a model by mixing and matching various features for both the SEM and typical VC models of familial correlation.

Identifiability and the Implied Mean and Covariance

A SEM is globally identifiable if all of its parameters can be determined uniquely from its associated mean and covariance structure. It is shown in Appendix A that there is a particularly simple form for the mean and covariance structure of the model discussed. Let $H = (I - B)^{-1}$. We have the following structure for the mean:

$$E[y_{ki}] = (\Gamma_m + H \Gamma_p) z_{ki} \text{ and } E[y_k] = [I_{n_k} \otimes (\Gamma_m + H \Gamma_p)] z_k. \quad (3)$$

The variance structure may be written as:

$$\text{Var}[y_k] = \sum_{j=1}^{\nu} M_{kj} \otimes (\Lambda H Z_j H^T \Lambda^T + E_j). \quad (4)$$

Besides being elegant, these formulas provide us with a tool for determining model identifiability. Consider the saturated model with $E[y_{ki}] = C z_{ki}$ and

$$\text{Var}[y_k] = \sum_{j=1}^{\nu} M_{kj} \otimes V_j,$$

where C and V_j contain a free parameter for each cell. Let us define the following conditions:

- (C1) The matrix $[z_{11} \dots z_{1n_1} \ z_{21} \dots z_{2n_2} \dots z_{Kn_K}]^T$ is full rank.
- (C2) There exists at least one pedigree k such that $\{M_{k1} \dots M_{k\nu}\}$ are linearly independent matrices.

If (C1) and (C2) are met, then it can be shown fairly easily that the saturated model is identifiable. One case where (C2) is violated is the inclusion of polygenic, independent and shared environmental components in data comprising only sibships. If (C1) and (C2) are met, then the task of determining if a SEM is identifiable may be reduced to finding whether a unique set of parameters satisfies the following equations:

$$C = (\Gamma_m + H \Gamma_p) \quad (5)$$

$$V_j = \Lambda H Z_j H^T \Lambda^T + E_j \text{ for } j \in \{1 \dots \nu\}. \quad (6)$$

That is, if θ represents all model parameters, then identifiability means we can solve for θ in terms of V_j and C (i.e. $\theta = \theta(C, V_1 \dots V_\nu)$). One simple strategy to prove identifiability is to find a partition of θ such that we can write $\theta = [\theta_1(V_1) \dots \theta_\nu(V_\nu) \ \theta_{\nu+1}(C, \theta_1, \dots, \theta_\nu)]$. In standard cases, there are a large number of intuitive rules for identifiability, for example the two indicator rules for factor analysis that have been summarized by Bollen [3]. Such rules are equally applicable here because we can apply them to each of the $\nu + 1$ equations in (6) and (5) to establish that such a partition exists.

Fitting the Model

It may be computationally demanding to fit the model we are proposing by maximum likelihood techniques. The size of each independent block of traits would be tn_k . Hence, for large pedigrees and a large number of traits, we are likely to run into computational and numerical difficulties. We propose a limited information approach – the minimum χ^2 approach – which is inspired by the approach of Muthén [26]. In the first step we form a limited information estimate for the saturated model discussed in the previous section, and in the second step we form a minimum χ^2 estimate of the model parameters based on the step 1 estimates.

In the first step, we begin by estimating the univariate VC models for each of the traits via maximum likelihood and assuming multivariate normality. This gives us an estimate for C and the diagonal elements for all V_j . To estimate the off-diagonal terms of V_j , we maximize the bivariate likelihood of each pair-wise set of traits with the previously estimated parameters fixed. Putting this more formally, let $C_{\cdot i}$ be the i -th row of C and define $\delta_i = (C_{\cdot i} V_{1ii} \dots V_{\nu ii})$, $\delta_{ij} = (V_{1ij} \dots V_{\nu ij})$ and $\delta = (\delta_1 \dots \delta_t \ \delta_{t+1} \dots \delta_{(t-1)t})$. Then let $ll_i(\delta_i)$ be the univariate log-likelihoods and let $ll_{ij}(\delta_i, \delta_j, \delta_{ij})$ be the bivariate log-likelihoods. The stage 1 estimate finds a solution to the equation:

$$U_K = \begin{bmatrix} \frac{dll_1}{d\delta_1} & \dots & \frac{dll_1}{d\delta_t} & \frac{dll_{12}}{d\delta_{12}} & \dots & \frac{dll_{(t-1)t}}{d\delta_{(t-1)t}} \end{bmatrix}^T = 0 \quad (7)$$

Now let δ_0 be the true parameter value, $\hat{\delta}$ be the estimated parameter value, and define

$$J_\delta = \frac{1}{K} \frac{dU_K}{d\delta} \bigg|_\delta.$$

Suppose that

$$\frac{1}{K} \text{Var}(U_K(\delta_0)) \xrightarrow{P} F.$$

Then, using an asymptotic expansion, it may be shown that:

$$\sqrt{K}(\hat{\delta} - \delta_0) \xrightarrow{D} MVN(0, J_{\delta_0}^{-1} F J_{\delta_0}^{-1}). \quad (8)$$

To find an estimate of the covariance, we may replace J_{δ_0} with $J_{\hat{\delta}}$. Also, let u_k be the first derivative of the log likelihood for pedigree k . Then we replace F with the robust estimate:

$$\hat{F} = \frac{1}{K} \sum_{k=1}^K u_k u_k^T. \quad (9)$$

Let θ be a vector containing all the free parameters in the SEM, and $f(\theta)$ be the values of δ corresponding to θ . To form the step 2 estimate, we seek the arguments which minimize the function:

$$Q(\theta) = [\hat{\delta} - f(\theta)]^T W [\hat{\delta} - f(\theta)] \quad (10)$$

for some choice of W . The most obvious choice of W is the inverse of the asymptotic covariance matrix for $(J_{\hat{\delta}}^{-1}\hat{F}J_{\hat{\delta}})$. However, previous studies of similar estimators have found that this estimator requires large sample sizes [27]. Following Muthén et al. [28], we therefore suggest the use of

$$W = \text{diag}(J_{\hat{\delta}}^{-1}\hat{F}J_{\hat{\delta}}^{-1})^{-1}.$$

Let

$$A = \left(\frac{df^T}{d\theta} \right)^T W \left(\frac{df^T}{d\theta} \right)$$

and

$$B = \left(\frac{df^T}{d\theta} \right)^T W J_{\hat{\delta}}^{-1} \hat{F} J_{\hat{\delta}}^{-1} W \left(\frac{df^T}{d\theta} \right).$$

The asymptotic covariance matrix for $\hat{\theta}$ may be estimated as $A^{-1}BA^{-1}$. This estimate of the parameters should maintain asymptotically correct coverage rates, even in the presence of non-normality, given some regularity conditions such as finite fourth moments.

Missing Data

One feature present in nearly all real data is missingness. An easy approach that we have implemented is to throw out individuals with incomplete observations for a given likelihood at the first step when estimating the saturated model parameters. In this way the individual may contribute information to some of the univariate and bivariate likelihoods but not all.

An excellent discussion of robustness and missingness in SEM may be found in Yuan and Bentler [29]. One possible assumption is that the data are missing completely at random (MCAR), meaning that the probability that an observation is missing is independent of the trait values. A less stringent assumption is known as missing at random (MAR). This means that the probability that an observation is missing is independent of the unobserved trait values. To summarize the apparent implication of Yuan and Bentler [29] to our approach, if we assume MAR and multivariate normality, our estimates will maintain correct coverage rates. On the other hand, our approach will also maintain correct coverage rates if we assume MCAR and some regularity conditions such as finite fourth moments.

Ascertainment

A second feature found commonly in family-based studies is ascertainment. That is, pedigrees are selected for inclusion in a study based upon some criteria for the observed traits. Generally, pedigrees are collected by first identifying an individual who is affected or has an extreme phenotype. Such an individual, who causes the pedigree to be sampled, is known as a proband. Often-times the study is designed so that only a restricted group of individuals in a population can become probands. An individual who could potentially become a proband regardless of phenotype is said to be in the proband sampling frame. If the probability that an individual in the proband sampling frame becomes a proband is very small we have the ideal condition known as single ascertainment. The problem of modeling ascertainment accurately is quite difficult and has generated some controversy [30, 31].

In the literature on VC modeling in family data, two common approaches to ascertainment have been investigated [32, 33]. Both methods assume single ascertainment. The first method uses the likelihood conditional on the trait of interest for the proband exceeding a threshold. The second method uses the likelihood conditional on the actual trait value of the proband. The first method is slightly more efficient if ascertainment is actually performed by selecting a proband whose trait value exceeds a specific value. However, this method is often not applicable.

We suggest a way of implementing the second approach. Suppose that trait 1 is the trait upon which ascertainment is performed. We designate the values of trait 1 for the probands as $y_{j=p,i}$ while all the trait values for the i -th trait are y_i . We begin by estimating δ_1 by maximizing the conditional univariate log-likelihood $ll(\delta_1; y_1) - ll(\delta_1; y_{j=p,1})$. We then estimate δ_i and δ_{1i} by maximizing the conditional bivariate likelihoods $ll(\delta_i, \delta_{1i}; y_i, y_1) - ll(\delta_i, \delta_{1i}; y_{j=p,i}, y_{j=p,1})$ for each $i \in \{2 \dots t\}$ with the previously estimated parameters fixed. Next we estimate δ_{ij} by maximizing the conditional log-likelihood $ll(\delta_{ij}, y_{ij}; y_j, y_1^p) - ll(y_{j=p,i})$ for $i \in \{2 \dots t\}$ and $j \in \{i+1 \dots t\}$ with the previously estimated parameters fixed. To estimate the SEM we then minimize the quadratic form defined by (10).

A χ^2 Index of Fit, Model Testing and Model Comparison

The χ^2 index of fit statistic represents a test of the null hypothesis that the model used is correct versus the alternative hypothesis that the model is saturated. One general measurement of the distance between the fitted model and the saturated model is

$$T = \inf_{\theta} Q(\theta) = [\hat{\delta} - f(\theta)]^T W [\hat{\delta} - f(\theta)].$$

Unfortunately, the asymptotic distribution of T is complex. However, by scaling the statistic appropriately it is possible to make T approximately a χ^2 distribution. Satorra [34], on pp. 258–259, describes the formulas for two possible ways of scaling the statistic: the mean scaled approach and the mean/variance scaled approach. Let p_s be the number of parameters in the saturated model, and p_m be the number of parameters in the estimated model. In the mean scaled approach, a scaling factor (c_1) is calculated such that $E[c_1 T] = E[\chi^2_{(p_s - p_m)}]$. It is then assumed that $T_{mean} = c_1 T \sim \chi^2_{(p_s - p_m)}$. In the mean/variance scaled approach, an adjusted number of degrees of freedom (d_2) is first calculated and then a scaling factor is calculated such that $E(c_2 T) = \chi^2_{d_2}$. It is then assumed that $T_{mean-var} = c_2 T \sim \chi^2_{d_2}$. Similar adjustments are available to create χ^2 difference tests for comparing nested models.

It is possible to implement some of the well-known indices of fit. For example, we could define a comparative index of fit (CFI) [35, 36]. Suppose we are comparing models i and k . Let d_i and d_k be the associated degrees of freedom and let D_i and D_k be one of the associated χ^2 indexes defined above. Let $\hat{\lambda}_i = \max(D_i - d_i, D_k - d_k, 0)$ and $\hat{\lambda}_k = \max(D_k - d_k, 0)$. Then, we may define

$$CFI = \frac{\hat{\lambda}_i - \hat{\lambda}_k}{\hat{\lambda}_i}.$$

The typical rules of thumb, such as a cut off of $CFI > 0.95$ [37], may not apply for the sorts of estimating procedures we are using. The use of indices of fit is controversial. Our own inclination is to maintain the hypothesis-testing framework as much as possible in SEM. We would emphasize that a significant χ^2 statistic should be carefully investigated. It is not enough to simply report an acceptable fit index when there is strong evidence against the proposed model.

Fig. 1. Linkage analysis model for platelet count. L1 is a latent trait which underlies the three measurements. a_s represents the major gene VC.

Fig. 2. Genetic association analysis model for platelet count. L1 is a latent trait which underlies the three measurements.

Examples of Linkage and Association

We now briefly suggest some example models that could fit into our framework. We are not necessarily arguing for the correctness of any of the models below as much as demonstrating what can be modeled. Suppose, for example, that we have three measurements for platelet count taken in the teen years approximately a year apart as in Evans et al. [38]. We hypothesize that there is a single latent trait underlying the three measurements. Under this assumption we may develop a model to perform linkage analysis or association analysis as depicted in figure 1 and 2. For linkage analysis we could test $H_0: \text{Var}(a_s) = 0$ versus $H_0: \text{Var}(a_s) > 0$. For association analysis we would test $H_0: \gamma = 0$ versus $H_0: \gamma \neq 0$. Both these tests have reduced degrees of freedom compared to a full multivariate analysis.

Results

Simulation Study

To validate our method of fitting the data we performed a simple simulation study. We simulated 300 nuclear families with a random number of siblings generated as $2 + \text{Poisson}(\lambda = 2)$. The model simulated involves 3 observed traits at 3 different time points for a total of 9 traits. The model is depicted in figure 3 and the matrix form for the model may be found in Appendix B. As may be seen, body mass index (BMI) was simulated to be observed without error, while a latent trait LBP was assumed to underlie systolic (SBP) and diastolic (DBP) blood pressure. BMI was simulated to have a random intercept (I) and slope (S). The random slope was simulated as an additive function of a SNP which is segregating in the pedigree. This is an example of a latent growth curve model.

Fig. 3. Simulated model. SBP_i is the i -th measurement of systolic blood pressure. DBP_i is the i -th measurement of diastolic blood pressure. BMI_i is the i -th measurement of body mass index. S and I are a latent slope and intercept. SNP is a single nucleotide polymorphism. The prefix L is used for latent traits.

Table 3. Bias and coverage rates for simulated model

Matrix	Parameter	Avg. Est.*	True Val.	Bias	Coverage rate*, %	
					95% CI	99% CI
Λ	λ	0.50	0.50	0.0005	94.9	99.0
E_{env}	σ_{e1}	24.96	25.00	-0.0419	94.4	98.7
	σ_{e2}	24.92	25.00	-0.0761	94.3	98.9
B	β	1.00	1.00	0.0035	95.0	99.0
Γ_m	μ_6	19.95	20.00	-0.0529	95.2	99.3
	μ_7	19.95	20.00	-0.0468	94.7	99.0
	μ_8	19.96	20.00	-0.0394	94.9	99.1
Γ_s	μ_1	79.94	80.00	-0.0585	95.1	99.2
	μ_2	79.94	80.00	-0.0626	94.9	98.8
	μ_3	79.93	80.00	-0.0690	95.5	99.3
	γ_{AGE1}	0.00	0.00	-0.0002	94.9	99.0
	γ_{SNP}	1.00	1.00	-0.0026	94.9	98.7
	μ_4	0.00	0.00	0.0050	95.2	98.9
	γ_{AGE2}	0.03	0.03	0.0003	94.4	98.7
	μ_5	19.98	20.00	-0.0151	94.3	98.6
$Z_{polygenic}$	σ_{p1}	0.00	0.00	-0.0011	93.8	98.7
	σ_{p2}	-0.25	0.00	-0.2476	93.1	98.7
	σ_{p3}	0.99	1.00	-0.0071	94.0	98.5
	σ_{p4}	3.93	4.00	-0.0652	93.4	98.0
$Z_{environment}$	σ_{e1}	3.99	4.00	-0.0111	94.7	98.9
	σ_{e2}	24.89	25.00	-0.1093	95.1	98.8
	σ_{e3}	0.99	1.00	-0.0054	94.5	98.8
	σ_{e4}	4.00	4.00	0.0038	94.9	98.7

* Across all 5,000 replicates.

Table 4. Type I error of the χ^2 statistic for the simulated model

	$\alpha = 0.05$	$\alpha = 0.01$
Mean adjusted	0.184	0.1104
Mean and variance adjusted	0.064	0.013

We produced 5,000 data sets from this model. To evaluate coverage rates we found the percentage of simulations with confidence intervals that covered the true parameter. The model was simulated and fit by a Matlab script.

Table 3 displays the results of our simulation. As may be seen, the parameter estimates are close to unbiased, and the coverage rates are generally quite close to the nominal coverage rates. We also calculated the mean ad-

justed and mean and variance adjusted tests for goodness of fit under the true model. As may be seen in table 4, the mean corrected χ^2 statistic was quite liberal. On the other hand, the mean and variance corrected χ^2 statistic performed reasonably well. We would point out, however, that our simulation was somewhat limited in its complexity, and we did not attempt to compare our results with other well-known software packages such as Mx [39].

Discussion

Depending on the problem at hand, there are numerous ways to use the modeling framework we have described. First, we may have no marker data available, but may be interested in building a structural model based on the phenotypes of family members. The framework above will allow us to do this. Second, we may be interested in gene mapping using linkage and/or association. We suggest that it may be best in this case to first develop a model without the use of marker data. We may then proceed to linkage or association using a Wald type test. We note that in the context of multivariate linkage analysis [20, 40], the testing problem is not necessarily simple and may involve complex asymptotic distributions [41]. It may also be possible to develop robust score tests to evaluate linkage or association one marker per locus at a time across the entire genome efficiently. There may be significant power gains by using a multivariate model that correctly reflects the conditional independencies in the data. Third, it may be possible to create multiple SNP models that utilize our knowledge of the underlying biology (cf. [42]). Fourth, we may be interested in estimating the effect size of known genes. It is also possible, using our approach, to dissect the direct and indirect effects. Fifth, we may be interested in using the marker data to help us determine the causal connections within the data. This is often referred to as Mendelian randomization. We note, however, that instrumental variable models in SEM may not be robust to weak instruments [43]. Other methods exist which may work better in such situations [44]. Several works that attempt to infer causal relations using marker data include the work of Li et al. [45] and Chaibub Neto et al. [46].

In the future, we plan to extend the framework discussed above to include ordinal and other categorical traits and multigroup analysis. Multigroup analysis would allow the modeling of certain forms of gene $\times$ gene and gene $\times$ environment interaction. We also plan to do some simulation experiments and show some examples of how the framework may be used in real data.

We are in the process of implementing our framework in an R package that will be made available as a free download. The Matlab code used in our simulation is available upon request.

Acknowledgments

This work was supported by National Cancer Institute (NCI) Grant R25T CA094186 (NJM), National Heart Lung and Blood Institute (NHLBI) grant T32 HL0756 (NJM), National Center for Research Resources (NCRR) Grant P41 RR03655 (RCE, CMS), NCI Grant P30 CAD43703 (RCE), NCRR Multidisciplinary Clinical Research Career Development Programs Grant KL2 RR024990 (CMS) and NHLBI grant R01 HL096811 (CMS).

Appendix A: Mean and Covariance Structure

We seek here to work out the implied mean and covariance structure. We repeatedly use the well-known identity $(A \otimes B)(C \otimes D) = AC \otimes BD$ if the dimensions of A and B are the same as C and D , respectively. Let $H = (I - B)^{-1}$. Then:

$$\eta_k = (I_{n_k} \otimes B)\eta_k + (I_{n_k} \otimes \Gamma_p)z_k + (I_{n_k} \otimes I_l)\sum_{j=1}^v s_{kj}.$$

Solving for η_k we get:

$$\eta_k = (I_{n_k} \otimes H\Gamma_p)z_k + (I_{n_k} \otimes H)\sum_{j=1}^v s_{kj}.$$

Therefore:

$$E[\eta_k] = (I_{n_k} \otimes H\Gamma_p)z_k$$

$$\begin{aligned} \text{Var}[\eta_k] &= (I_{n_k} \otimes H)\left(\sum_{j=1}^c M_{kj} \otimes Z_j\right)(I_{n_k} \otimes H)^T \\ &= \sum_{j=1}^c M_{kj} \otimes HZ_j H^T \end{aligned}$$

Also:

$$y_k = (I_{n_k} \otimes \Gamma_m)z_k + (I_{n_k} \otimes \Lambda)\eta_k + I_{n_k} \otimes I_l \sum_{j=1}^v \varepsilon_{kj}$$

Therefore:

$$\begin{aligned} E[y_k] &= (I_{n_k} \otimes \Gamma_m)z_k + (I_{n_k} \otimes \Lambda)E[\eta_k] \\ &= [I_{n_k} \otimes (\Gamma_m + H\Gamma_p)]z_k \end{aligned}$$

and:

$$\begin{aligned} \text{Var}[y_k] &= (I_{n_k} \otimes \Lambda)\text{Var}[\eta_k](I_{n_k} \otimes \Lambda)^T + \sum_{j=1}^c M_{kj} \otimes E_j \\ &= \sum_{j=1}^c M_{kj} \otimes (\Lambda HZ_j H^T \Lambda^T + E_j) \end{aligned}$$

Appendix B: Matrix Form for Simulated Model

$$\begin{bmatrix} BMI_1 \\ BMI_2 \\ BMI_3 \\ SBP_1 \\ SBP_2 \\ SBP_3 \\ DBP_1 \\ DBP_2 \\ DBP_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} LBMI_1 \\ LBMI_2 \\ LBMI_3 \\ LBP_1 \\ LBP_2 \\ LBP_3 \\ S \\ I \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

$$\begin{bmatrix} LBMI_1 \\ LBMI_2 \\ LBMI_3 \\ LBP_1 \\ LBP_2 \\ LBP_3 \\ S \\ I \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \mu_1 & 0 & 0 \\ \mu_2 & 0 & 0 \\ \mu_3 & 0 & 0 \\ \mu_4 & \gamma_{AGE1} & \gamma_{SNP} \\ \mu_5 & \gamma_{AGE2} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ AGE \\ SNP \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ \beta & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} LBMI_1 \\ LBMI_2 \\ LBMI_3 \\ LBP_1 \\ LBP_2 \\ LBP_3 \\ S \\ I \end{bmatrix} + \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}$$

$$\begin{aligned} \text{diag}[\text{Var}(p_1 \dots p_8)^T] &= (\sigma_{p_1} \sigma_{p_1} \sigma_{p_1} \sigma_{p_2} \sigma_{p_2} \sigma_{p_2} \sigma_{p_3} \sigma_{p_4}) \\ \text{Cov}(p_i, p_j) &= 0 \text{ for } i \neq j \\ \text{diag}[\text{Var}(e_1 \dots e_8)^T] &= (\sigma_{e_1} \sigma_{e_1} \sigma_{e_1} \sigma_{e_2} \sigma_{e_2} \sigma_{e_2} \sigma_{e_3} \sigma_{e_4}) \\ \text{Cov}(e_i, e_j) &= 0 \text{ for } i \neq j \\ \text{diag}[\text{Var}(\varepsilon_1 \dots \varepsilon_6)^T] &= (\sigma_{\varepsilon_1} \sigma_{\varepsilon_1} \sigma_{\varepsilon_1} \sigma_{\varepsilon_2} \sigma_{\varepsilon_2} \sigma_{\varepsilon_2}) \\ \text{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \text{ for } i \neq j \\ \text{Var}(e_i) &= \sigma_{e_i} \\ \text{Cov}(e_i, e_j) &= 0 \text{ for } i \neq j \\ \text{Var}(\varepsilon_i) &= \sigma \varepsilon_i \\ \text{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \text{ for } i \neq j \end{aligned}$$

References

- Wright S: Correlation and causation. *Journal of Agricultural Research* 1921;20:557–585.
- Pearl J: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- Bollen KA: *Structural Equations with Latent Variables*. New York, John Wiley & Sons, Inc 1989.
- Jöreskog KG: A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 1969;34:183–202.
- Muthén BO: Beyond SEM: General latent variable modeling. *Behaviormetrika* 2002; 29:81–117.
- Tomarken AJ, Waller NG: Structural equation modeling: strengths, limitations, and misconceptions. *Ann Rev Clin Psychol* 2005; 1:31–65.
- Skrondal A, Rabe-Hesketh S: *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, 2004.
- Spirtes P, Glymour CN, Scheines R: *Causation, Prediction, and Search*. MIT Press, 2000.
- Valente BD, Rosa GJM, de los Campos GA, Gianola D, Silva MA: Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics* 2010; 185:633–644.
- Bollen KA, Long JS: *Testing Structural Equation Models*. Sage Publications, 1993.
- Jöreskog KG, Sörbom D: *Lisrel 8: User's reference guide*. Scientific Software International, 1996.
- Neale MC, Cardon LR: *Methodology for Genetic Studies of Twins and Families*. Springer Netherlands, 1992.
- Medland S, Neale M: An integrated phenomic approach to multivariate allelic association. *Eur J Hum Genet* 2009;18:233–239.
- Posthuma D, De Geus EJC, Boomsma DI, Neale MC: Combined linkage and association tests in Mx. *Behav Genet* 2004;34:179–196.
- van den Oord E: Framework for identifying quantitative trait loci in association studies using structural equation modeling. *Genet Epidemiol* 2000;18:341–359.
- Todorov AA, Vogler GP, Gu C, Province MA, Li Z, Heath AC, Rao DC: Testing causal hypotheses in multivariate linkage analysis of quantitative traits: general formulation and application to sibpair data. *Genet Epidemiol* 1998;15:263–278.
- Gianola D, Sorensen D: Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics* 2004;167:1407.
- Boomsma DI, Dolan CV: A comparison of power to detect a qtl in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behav Genet* 1998;28:329–340.
- Bauman LE, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K: Fishing for pleiotropic qtls in a polygenic sea. *Ann Hum Genet* 2005;69:590–611.
- Eaves LJ, Neale MC, Maes H: Multivariate multipoint linkage analysis of quantitative trait loci. *Behav Genet* 1996;26:519–525.
- Henderson CR, Quaas RL: Multiple trait evaluation using relatives' records. *J Anim Sci* 1976;43:1188.
- Mrode RA, Thompson R: *Linear Models for the Prediction of Animal Breeding Values*. Cabi, Cambridge 2005.
- Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. *Hum Hered* 1971;21:523–542.
- Longford N, Muthén B: Factor analysis for clustered observations. *Psychometrika* 1992; 57:581–597.
- McDonald R, Goldstein H: Balanced versus unbalanced designs for linear structural relations in two-level data. *Br J Math Stat Psychol* 1989;42:215–232.
- Muthén B: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 1984;49:115–132.
- Muthén B: Goodness of fit with categorical and other nonnormal variables. *SAGE Focus Editions* 1993;154:205–205.
- Muthén B, Du Toit SHC, Spisic D: Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Technical Report 1997.
- Yuan KH, Bentler PM: Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Soc Methodol* 2000;30:165–200.
- Vieland V, Hodge S: Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *Am J Hum Genet* 1995;56:33–43.
- Ginsburg E, Elston RC, Malkin I, Ginsburg E, Malkin I, Elston RC: *Theoretical Aspects of Pedigree Analysis*. Ramot-tel Aviv University, 2006.
- Rao D, Wette R, Ewens W: Multifactorial analysis of family data ascertained through truncation: a comparative evaluation of two methods of statistical inference. *Am J Hum Genet* 1988;42:506–515.
- de Andrade M, Amos CI: Ascertainment issues in variance components models. *Genet Epidemiol* 2000;19:333–344.
- Satorra A: Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociol Methodol* 1992;22:249–278.
- Bentler PM: Comparative fit indexes in structural models. *Psychol Bull* 1990;107: 238–246.
- Shipley B: *Cause and Correlation in Biology: A user's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, 2000.
- Hu L, Bentler PM: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling* 1999;6:1–55.
- Evans DM, Zhu G, Duffy DL, Montgomery GW, Frazer IH, Martin NG: Multivariate qtl linkage analysis suggests a qtl for platelet count on chromosome 19q. *Eur J Hum Genet* 2004;12:835–842.
- Neale MC, Medical College of Virginia. Department of P, Virginia Institute for Psychiatric and Behavioral G: Mx: Statistical modeling. Department of Psychiatry, Medical College of Virginia Richmond, VA, 1997.
- Hottenga J, Boomsma DI: Qtl detection in multivariate data from sibling pairs; in Neale BM, Ferreira M, Medland SE, Posthuma D (eds): *Statistical Genetics: Gene Mapping through Linkage and Association*. New York, Taylor & Francis Group, 2008.
- Morris NJ, Elston R, Stein CM: Calculating asymptotic significance levels of the constrained likelihood ratio test with application to multivariate genetic linkage analysis. *Stat Appl Genet Mol Biol* 2009;8:Article 39.
- Nock N, Wang X, Thompson C, Song Y, Baechle D, Raska P, Stein C, Gray-McGuire C: Defining genetic determinants of the metabolic syndrome in the framingham heart study using association and structural equation modeling methods, *BioMed Central Ltd*, 2009;3:S50.
- Bound J, Jaeger DA, Baker RM: Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995;90:443–450.
- Morris N, Gray-McGuire C, Stein C: Mendelian randomization in family data, *BioMed Central Ltd*, 2009;3:S45.
- Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA: Structural model analysis of multiple quantitative traits. *PLoS Genet* 2006;2:e114.
- Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS: Inferring causal phenotype networks from segregating populations. *Genetics* 2008.