Chapter 27

Structural Equation Modeling

Catherine M. Stein, Nathan J. Morris, and Nora L. Nock

Abstract

Structural equation modeling (SEM) is a multivariate statistical framework that is used to model complex relationships between directly and indirectly observed (latent) variables. SEM is a general framework that involves simultaneously solving systems of linear equations and encompasses other techniques such as regression, factor analysis, path analysis, and latent growth curve modeling. Recently, SEM has gained popularity in the analysis of complex genetic traits because it can be used to better analyze the relationships between correlated variables (traits), to model genes as latent variables as a function of multiple observed genetic variants, and assess the association between multiple genetic variants and multiple correlated phenotypes of interest. Though the general SEM framework only allows for the analysis of independent observations, recent work has extended SEM for the analysis of general pedigrees. Here, we review the theory of SEM for both unrelated and family data, the available software for SEM, and provide an example of SEM analysis.

Key words: Multivariate analysis, Latent variables, Modeling, Candidate gene analysis, Complex traits, Path analysis, Structural equation modeling, Association, Population studies, Family studies

1. Introduction

Structural equation modeling (SEM) is a multivariate statistical method that involves the estimation of parameters for a system of simultaneous equations. SEM is a generalized framework that includes regression analysis, pathway analysis, factor analysis, simultaneous econometric equations, and latent growth curve models, to name a few (1). Here, we provide an overview of the methodology behind the SEM framework, how this framework has been extended to analyze related individuals, and currently available software that can be used to conduct SEM analyses. After the overview, we provide a step-by-step procedure for SEM analysis. Finally, we provide some notes on challenges faced when performing these types of analyses.

Robert C. Elston et al. (eds.), *Statistical Human Genetics: Methods and Protocols*, Methods in Molecular Biology, vol. 850,
DOI 10.1007/978-1-61779-555-8_27, © Springer Science+Business Media, LLC 2012

1.1. Overview of Methodology

SEM is used to estimate a system of linear equations to test the fit of a hypothesized "causal" model. Thus, the first step involves visualizing the hypothesized model or creating a "path diagram" based on prior knowledge and/or theories. In path diagrams, rectangles represent observed or directly measured variables and circles/ovals typically represent unobserved or latent constructs which are defined by measured variables. Unidirectional arrows represent causal paths, where one variable influences another directly, and double-headed arrows represent correlations between variables. Some prefer the term "arc" rather than "causal path" (2, 3). Fig. 1 illustrates an example SEM model.

The system of equations can be written as a number of separate equations or with a general matrix notation. SEMs comprise two submodels. First, the measurement model estimates relationships between the observed variables, also referred to as indicators, and latent variables; this is the same framework used in factor analysis. *Please note that here we use the word "indicator variable" in a very different way than in typical statistical models.* In regression and other statistical theories, "indicator variable" implies a binary yes/no sort of variable. Here, as is customary for SEM, "indicator variable" refers to a variable that is directly associated with a latent variable such that differences in the values of the latent variable mirror differences in the value of the indicator (4). Second, the structural model develops the relationships between the latent variables. For clarity of presentation, here we describe the system of equations for this particular example. The measurement model consists of the following equations, using standard notation used by Bollen (1):

$$\begin{aligned} x_1 &= \lambda_1\xi_1 + \delta_1 & y_1 &= \lambda_3\eta_1 + \varepsilon_1 \\ x_2 &= \lambda_2\xi_2 + \delta_2 & y_2 &= \lambda_4\eta_1 + \varepsilon_2 \\ x_3 &= \lambda_3\xi_3 + \delta_3 & y_3 &= \lambda_5\eta_1 + \varepsilon_3, \end{aligned}$$

where the x's and y's are observed indicators for latent variables, the ξ's and η's are latent variables, the λ's are factor loadings, and the ε's and δ's are error, or disturbance, terms. In general matrix notation, the measurement model is written as

$$\begin{aligned} \mathbf{x} &= \Lambda_\mathbf{x}\xi + \delta \\ \mathbf{y} &= \Lambda_\mathbf{y}\eta + \varepsilon. \end{aligned}$$

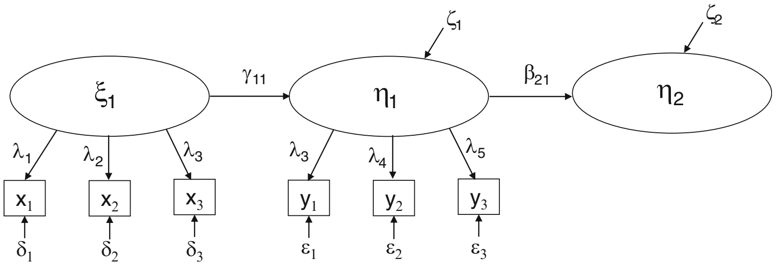

Fig. 1. Example SEM diagram.

Using the path diagram, the arrows point to the *x*'s and *y*'s, so they are modeled as dependent variables. Also, note that the factor loadings for x_1 and y_1 can be set to 1, which can be done for two reasons: so that the model is identifiable and so that the latent variable is on the same statistical scale as the observed variables. Model identification, which is discussed in further detail in Subheading 2.1, can also be achieved in other ways, such as setting the variance for the latent variable to 1. Generally, the indicator with factor loading set to 1 is chosen based on what the analyst deems is the best descriptor of the latent construct, but can be arbitrary. Finally, we can differentiate between *exogenous variables*, which have no directed arcs ending on them, and *endogenous variables*, which have at least 1 arc ending on them.

The structural model consists of the following equations:

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\xi_2 + \zeta_2,$$

where the γ and β terms are factor loadings for the latent variables and ζ's are error terms. Here, we can evaluate causal relationships between unobserved variables. In general, the structural model may be rewritten in matrix form as the following:

$$\eta = \alpha + \mathbf{B}\eta + \Gamma\xi + \zeta,$$

where $\boldsymbol{\eta}$ is a $m \times 1$ vector of latent endogenous variables, $\boldsymbol{\xi}$ is an $n \times 1$ vector of latent exogenous variables, $\boldsymbol{\alpha}$ is an $m \times 1$ vector of intercept terms, $\mathbf{B}$ is an $m \times m$ matrix of coefficients that give the influence of $\boldsymbol{\eta}$ on each other, $\boldsymbol{\Gamma}$ is an $m \times n$ matrix of the coefficients of the effect of $\boldsymbol{\xi}$ on $\boldsymbol{\eta}$, and $\boldsymbol{\zeta}$ is the $m \times 1$ vector of disturbances that contain the explained parts of the $\boldsymbol{\eta}$'s. Though it may appear counterintuitive to regress $\boldsymbol{\eta}$ on itself, each variable in $\boldsymbol{\eta}_\mathrm{i}$ is influenced by other variables in $\boldsymbol{\eta}_\mathrm{i}$, so this represents relationships between latent variables and not necessarily feedback loops. We assume that $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ are mutually uncorrelated.

Traditional regression approaches are robust to measurement errors in the outcome but not in the predictors. Also, univariate regression approaches cannot model the correlation between error terms for two different outcomes. SEM allows us to model measurement error for both the predictor and the outcome, and it allows a high degree of flexibility in modeling the correlation between the various error terms. For example, if two of the indicators were lab measurements assayed in one lab, while another two were measurements conducted in another lab, the analyst could model the correlation between the first pair of measurements separately from the second pair. Also, the SEM allows for the decomposition of effects if the direct and indirect effect of variables on the outcome is of interest. For example, the direct effect of η_1 on η_2 is estimated by β_{21}, and the indirect effect of ξ_1 on η_2 is estimated by γ_{11}. Alternatively, one could model the direct effect of ξ_1 on η_2 with the model depicted in Fig. 2, with corresponding coefficient $\gamma_{12.}$ More detail on mediation models can be found elsewhere (5, 6).

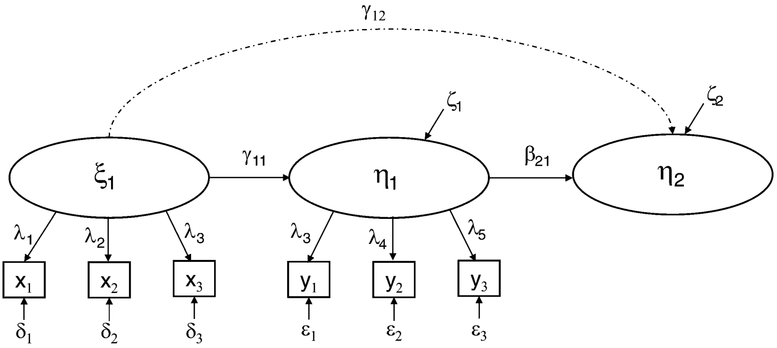

Fig. 2. Example SEM diagram, illustrating the addition of a direct effect in the model.

These models are estimated using the variance–covariance matrix of the data. Usually, maximum likelihood estimation fitting functions are used to fit the system of equations to the data, but this method requires that the data be normally distributed and the observations be independent. Variations that relax the assumption of multivariate normality have been developed, including the robust weighted least squares estimator (WLSMV), which allows for binary and categorical dependent variables (7). To assess the overall model fit, there are a number of fit statistics, including the root mean squared error (RMSEA) and comparative fit index (CFI) (1), and for categorical data, the weighted root mean square residual (WRMR) is appropriate (8). Hu and Bentler (9) categorize these fit statistics as "comparative" or "absolute." One could also compare nested models, as is done with traditional regression models and segregation analysis models, using a likelihood ratio test (LRT) and non-nested models using Akaike's AIC; by contrast, the aforementioned fit statistics (RMSEA, CFI, WRMR, etc.) do not require the models being compared to be nested.

1.2. SEM for Genetics

As pointed out by Pearl (3), SEM was first developed by geneticists. Early models, developed by Sewall Wright (10), were called path analysis. Later models were parameterized in very specific ways. Twin pair data could be used for the estimation of the proportion of variance due to additive genetic, dominance genetic, and shared environmental effects (11), the so-called ACE models. Nuclear family data could also be used for the estimation of additive genetic and shared environmental variance, using the so-called Tau and Beta models (12).

SEM is easily extended for the analysis of genetic and environmental influences on traits. For example, genes may be modeled as unobserved latent constructs with single nucleotide polymorphisms (SNPs) as indicators of these gene constructs (13, 14). If the investigator has a specific polymorphism of interest, it may be modeled as an observed variable. Our work has shown that densely spaced SNPs are best for modeling latent gene constructs, and linkage disequilibrium (LD) between these SNPs may be modeled by correlating the error terms within the measurement model (13).

A word of caution is needed here regarding the selection of genes or SNPs for analysis within a SEM framework. We emphasize that SEM is a hypothesis-driven approach. Thus, it is not agnostic like genome-scan approaches. In genome-wide searches for genes, the analyst conducts linkage or association analysis without a biologic model in mind. SEM is not amenable to this agnostic approach; the implication here is that the model developer must have a set of genes or biological pathways in mind. One approach is to select very specific candidate genes, and only these genes are included in the model. If genome-scan data are available, another approach is to take a two-stage approach by first conducting association analysis between the SNPs and the traits considered within the SEM (14). It should be emphasized up front that while algorithmic model searching and comparison may be useful, we do 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.

The above theory is applicable for independent observations. Muthén has proposed using a robust maximum likelihood estimator that provides test statistics and standard errors robust to nonindependence of observations (www.statmodel.com). However, in genetic studies, we often have data collected from families, and it is preferable to model the family structures explicitly. One approach to deal with family relationships in SEM was proposed by Todorov et al. (15), whose framework allowed for causal links between measured phenotypes and could include linkage information. However, their approach lacked an explicit measurement model for the traits and was difficult to extend to general pedigrees.

We have developed a generalized framework for modeling familial correlations for SEM (16). By using Kronecker notation, this framework allows for incorporation of both a measurement and structural model, as well as polygenic, environmental, and genetic variance components within the SEM. This allows for linkage and family-based association analyses to be conducted within a complex modeling framework, and can be used to build and compare causal models in family data with or without genetic marker data. Because of the generalized framework of this model, it is easily extended to sophisticated models such as latent growth curve models.

1.3. Software

For general SEM analysis, there are a number of packages available (summarized in Table 1). Of the packages that were not explicitly designed to conduct genetic analyses, none are freely available. Mplus does have the capability to conduct genetic analyses, but they are not as general as those methods described above. For example, SNP genotypes may be incorporated as observed variables, and there are special ways to conduct linkage analysis, but

Table 1
Overview of available SEM software packages

Package	Weblink or base package	Notes
Amos	Add-on to SPSS	
Proc CALIS	Procedure in SAS	
EQS	Multivariate software http://www.mvsoft.com/	
GLLAMM	Add-on to STATA: http://www.gllamm.org/	
HYBALL	http://web.psych.ualberta.ca/~rozeboom/	Free
LISREL	http://www.ssicentral.com/index.html	
Mplus	http://www.statmodel.com/	
Mx	http://www.vcu.edu/mx/	Free Best for twin data Now has GUI
OpenMx	http://openmx.psyc.virginia.edu/	Free R package—based on Mx
NEUSREL	Uses MATLAB http://www.neusrel.com/	
SYSTAT	http://www.systat.com/products.aspx	
Sem	Package for R: http://socserv.socsci.mcmaster.ca/jfox/Misc/sem/index.html	Free
SEGPATH	Weblink broken!	Free
SEPATH	In Statistica: http://www.statsoft.com/products/statistica-advanced-linear-non-linear-models/itemid/5/	
TETRAD	http://www.phil.cmu.edu/projects/tetrad/	Free beta

List partially abstracted from Ed Rigdon's Webpage: http://www2.gsu.edu/~mkteer/

besides that, Mplus cannot currently be used for more sophisticated genetic analyses. Another major consideration in the choice of software is ease of use vs. capability. For example, Amos allows the user to literally draw the SEM diagram that will be fitted, compared to Mplus, which requires the user to write code. However, Mplus may have additional capabilities in terms of specific algorithms and user support. In addition, we recommend that Amos be used with extreme caution, since it is too easy to draw a path diagram without thinking through the parameterization, theoretical implications, etc. A comparison of the most commonly used SEM software packages is provided by Buhi et al. (17).

Currently there are two packages available that implement SEM for genetic analysis. SEGPATH was originally developed for path analysis for sibling pairs and has been extended to conduct segregation analysis, linkage analysis, and interaction effects and analyze multiple phenotypes simultaneously (18) using the method of Todorov et al. (15). However, there are a few limitations of this software.

The likelihood formulation assumes multivariate normality, which makes the analysis of binary or categorical traits impossible without making important assumptions. Also, though Province et al. (18) state that their method has been extended to extended pedigrees, it is not trivial to estimate polygenic effects for such pedigrees without making other assumptions. Finally, at the time of this writing, the weblinks for SEGPATH were broken, so it is unknown whether this package is still available or actively maintained. Second, Mx software (http://www.vcu.edu/mx/) was originally developed for the analysis of twin data. Recently, a graphic user interface (GUI) and R package version have been made available, in which the user can draw SEM diagrams, similar to how Amos is used for general SEM. These recent developments are quite important since the coding language for Mx is not intuitive and rather difficult to implement without very specific examples. The Mx GUI can be used for SEM analysis of general data (unrelated individuals) and twin data, but sibpair data and other general pedigrees cannot be analyzed without extensive programming in the underlying Mx script language. Mx is also limited to traits that follow multivariate normality, though OpenMx can handle binary traits.

In addition, we are currently developing software for our own methodology (16). Recall that our framework includes both a measurement and structural model, allows for general pedigree structures, and is generalizable for both genetic and general SEM analyses. At the time of this writing, MATLAB code for our framework is freely available by request from the authors. We are also developing an R package for our method and will eventually release a GUI version also.

2. Methods

As we described above, software for general SEM is not freely available, and software for SEM with genetics has its limitations. Here, we provide an example using Mplus. We also note that genetics SEM packages will change in availability and functionality in the next couple of years.

Below we provide a worked example using data from the 1000 Genomes Project (Pilot Project 3) generated for the Genetic Analysis Workshop (GAW) held on October 17, 2010. We provide the example worked into two parts. Part 1 (Model 1) shows how to build the latent gene construct for one gene, and evaluate the gene's potential association on Q1, including potential effects of covariates. Part 2 (Model 2) demonstrates how to simultaneously model two genes, and evaluate their potential associations on Q1 (Fig. 1). In the following sections, we provide the Mplus v5.1 code with annotations for the various steps embedded within the code and highlight important findings in the subsequent discussion.

2.1. Develop the Model

SEM is a strongly hypothesis-driven analytical method. One danger with methods like SEM is the temptation to fit all sorts of models that have no grounding in biology or other scientific background. That is why it is essential to develop a model first. Draw the hypothesized relationships. If there are several plausible models, draw them all; in step #4, we will discuss how these models are compared.

There are several issues to keep in mind when developing the model. The measurement model (factor analysis) should be fitted first, followed by the structural model (2). First, as a general rule, when modeling latent constructs, each latent variable requires at least two observed indicator variables, but three is preferable (1); if there are only two indicators, then the latent variable must be correlated with another latent variable. This relates to the issue of model identification, which we will discuss subsequently. When conducting a factor analysis, the factor loadings should form independent clusters (2). Second, the analyst must be mindful of the default procedures of the software being used. For example, many software packages automatically estimate correlations between all latent variables. If the analyst does not want this, he/she must specify the analysis appropriately. Third, it is important to specify the disturbance/error terms and the correlation between them. If disturbance terms are left out, the assumption is made that the variable (x_i or y_i) is perfectly measured. Fourth, many software packages are unable to validly estimate parameters for binary or categorical dependent variables (endogenous variables); more about this in Note 1. Those software packages that can handle categorical outcomes have different computational approaches that should be considered.

Finally, one must consider how to parameterize the latent variables. There are a couple of approaches here. In one approach, the analyst may select one indicator variable for which the factor loading will be set to 1. The result of this will be that the variance of that latent construct will be set to the variance of that specific indicator variable but, at the same time, the importance of that variable to the latent construct cannot be estimated, because there will not be a factor loading. Alternatively, the analyst may fix the variance of the latent variable to 1 and its mean to 0, which then allows factor loadings to be estimated for all indicators. Both approaches are valid, and the decision comes down to interpretation.

It is important to assess whether the model is identified. Identification concerns whether it is possible to uniquely solve for the model parameters in terms of the moments of the observed variables using these equations. A SEM is identifiable if all of its parameters can be determined uniquely from a mean and covariance structure. One quick test to assess model identification is to see if each equation set by the model is a regression, and the covariance of all disturbance variables is zero (2). Another step in this process is to assign a scale to each latent variable that is measured with error. This can be done by either choosing one indicator for each latent variable and setting the factor loading to 1 or setting the variance for the latent variable. Evaluating the identification of a model is easier said than done, and a full discussion of this topic is outside the scope of this review. Bollen (1) provides algebraic arguments to assess model identification, and Pearl (3) presents graphical arguments. Also, see Note 1 for more on model identification.

2.2. Worked Example

Briefly, the 1000 Genomes Project is an international, public–private consortium aimed at building the most detailed map of human genetic variation, with the overarching goal of improving our understanding of the genetic contribution to common human diseases. Initially launched in 2008, three pilot studies have been completed to sequence the full genomes of 1,000 individuals in order to identify rare variants in diverse populations. Pilot Project 3 involved sequencing the coding regions (exons) of 3,205 genes in 697 individuals from seven populations, which revealed 24,487 rare and common genetic variants. To illustrate the latent gene construct SEM approach of Nock et al. (13, 14) using the GAW 17 data (unrelated subjects, Replicate 137), we selected two genes (OR52E4: olfactory receptor, family 52, subfamily E, member 4; OR2T3: olfactory receptor, family 2, subfamily T, member 3), which are biologically related to each other. We focused on Q1 as the phenotype because both OR52E4 and OR2T3 had at least one SNP each that was associated with Q1 in Replicate 137. We have taken a similar approach in our previous work (13, 14); sometimes it is helpful to first do a standard association analysis to identify SNPs associated with the trait(s) of interest, then include those genes within the SEM. In this example, we demonstrate how to model the variation in these genes with latent constructs using multiple SNPs and how to evaluate the potential associations of these genes on the simulated quantitative phenotype, Q1, including the potential effects of covariates [age, sex, population (pop1), and smoking] using Mplus v5.1 (Muthen and Muthen, 1998–2008, www.statmodel.com).

2.2.1. Prepare the Data

Once the model has been developed, the analyst can consider which variables to be included in the dataset, and then, how the datafile will be prepared. Many software packages accept typical flat files with typical delimiters, with each variable in a separate column, and each line of the file representing one individual subject's data. However, some software packages allow a covariance or correlation matrix to be input as data.

For this example, Q1, sex, age, smoking status, Pop1, and SNP genotypes for OR52E4 and OR2T3 were included in a comma-delimited (·csv) file, which was then uploaded into Mplus. In the Mplus code below, the data upload step can be seen in the "DATA:" section. For coding SNP genotype data, we employed an additive genetic model whereby SNPs were coded as 0, 1, or 2 for having 0, 1, or 2 copies of the variant (minor) allele, respectively.

2.2.2. Assess First Model

Next, the analyst can fit the first model. Model fit assessment has two parts: overall fit and component fit (19). Again, each software package differs in how the causal paths, correlations, and factor analyses for latent variables are specified. Regardless of the software package, after the model is fitted, a variety of statistics will be output. These include pathway coefficients and corresponding *p*-values, correlations, R^2 for each indicator for a latent variable, and model fit statistics that are based on maximum likelihood or generalized least squares, such as the chi-square, AIC, BIC, RMSEA, CFI, and other similar statistics. Further discussion of assessment of global fit vs. comparison of nested models can be found in the Subheading 1. In addition, model fit can also be assessed by identification of "Heywood cases," which are negative estimates of variance (1). A word of caution: sometimes, with large sample sizes, the power of significance tests based on the chi-square is so great that even trivial departures lead to rejection of the null hypothesis (19). On the other hand, indexes of model fit are for the most part *ad hoc*. See refs. 20 and 21 for some interesting discussions of these issues. Also see Note 2 for more on analysis of categorical variables.

We provide the example worked into two parts. Part 1 (Model 1) shows how to build the latent gene construct for one gene, OR52E4, and evaluate the gene's potential association on Q1, including potential effects of covariates. Part 2 (Model 2) demonstrates how to simultaneously model two genes, OR52E4 and OR2T3, and evaluate their potential associations on Q1 (Fig. 1). The following provides the Mplus v5.1 code for the worked examples, with annotations for the various steps embedded within the code.

Example: Part 1: OR52E4 Gene on Q1 Simulated Trait:

Mplus v5.1 Input File:

```
TITLE: GAW17 Q1 UNRELATEDS REP137 OR52E4

 DATA:
  FILE IS " _ "; !directory of where data file is located
  FORMAT IS FREE;
  LISTWISE = ON; !ensures no missing data used

 VARIABLE:
  NAMES ARE id sex age smoke pop1 Q1
  !OR52E4
  C11S773 C11S774 C11S775 C11S776 C11S777

 USEVARIABLES ARE id sex age smoke pop1 Q1
  C11S773 C11S774 C11S775 C11S776 C11S777;

  MISSING ARE .; !symbol used to denote missing data

  IDVARIABLE IS id;

ANALYSIS:
  TYPE=GENERAL;
  ESTIMATOR=MLR; !robust maximum likelihood estimator used for non-normal data
```

```
MODEL:
   OR52E4 BY C11S774 C11S773 C11S775 C11S776 C11S777; !gene defined by all
SNPs
   OR52E4 ON pop1; !adjustment for population structure
   Q1 ON age smoke sex OR52E4; !model for evaluating potential association of
OR52E4 on Q1
```

Below, we provide the Mplus v5.1 output (partial) from the code above.

Mplus v5.1 Output File (Partial):

```
GAW17 Q1 UNRELATEDS REP137 OR52E4
SUMMARY OF ANALYSIS
Number of groups                            1
Number of observations                    697
Number of dependent variables               6
Number of independent variables             4
Number of continuous latent variables       1
THE MODEL ESTIMATION TERMINATED NORMALLY
TESTS OF MODEL FIT
Chi-Square Test of Model Fit
      Value                               114.007*
      Degrees of Freedom                   29
      P-Value                               0.0000
      Scaling Correction Factor             1.129
        for MLR

*  The chi-square value cannot be used for chi-square difference tests. See chi-square
difference testing in the index of the Mplus User's Guide.

CFI/TLI
      CFI                 0.900
      TLI                 0.866

Loglikelihood
      H0 Value                          -7691.066
      H0 Scaling Correction Factor          2.823
        for MLR
      H1 Value                          -7626.710
      H1 Scaling Correction Factor          1.860
        for MLR
RMSEA (Root Mean Square Error Of Approximation)
      Estimate                              0.065
SRMR (Standardized Root Mean Square Residual)
      Value                                 0.041

MODEL RESULTS
STANDARDIZED MODEL RESULTS
STDYX Standardization
                                          Two-Tailed
              Estimate     S.E.  Est./S.E.   P-Value

 OR52E4   BY
   C11S774      0.809     0.045    18.060     0.000
   C11S773      0.341     0.066     5.152     0.000
   C11S775      0.328     0.053     6.214     0.000
   C11S776      0.820     0.035    23.457     0.000
   C11S777      0.863     0.041    21.028     0.000

 OR52E4   ON
```

```
   POP1             0.036    0.041    0.897    0.370

 Q1      ON
   OR52E4           0.173    0.040    4.321    0.000
[We will refer to the above result later on]
 Q1      ON
   AGE              0.294    0.034    8.727    0.000
   SMOKE            0.185    0.035    5.312    0.000
   SEX             -0.035    0.035   -1.010    0.313

Intercepts
   Q1              -0.659    0.140   -4.693    0.000
   C11S773          0.411    0.031   13.195    0.000
   C11S774          0.202    0.054    3.714    0.000
   C11S775          1.574    0.066   23.748    0.000
   C11S776          0.316    0.057    5.580    0.000
   C11S777          0.245    0.059    4.185    0.000

 Residual Variances
   Q1               0.845    0.027   31.539    0.000
   C11S773          0.884    0.045   19.588    0.000
   C11S774          0.345    0.073    4.753    0.000
   C11S775          0.892    0.035   25.782    0.000
   C11S776          0.327    0.057    5.702    0.000
   C11S777          0.255    0.071    3.606    0.000
   OR52E4           0.999    0.003  338.658    0.000

R-SQUARE
   Observed                          Two-Tailed
   Variable      Estimate    S.E.  Est./S.E.   P-Value
   Q1               0.155    0.027    5.782    0.000
   C11S773          0.116    0.045    2.576    0.010
   C11S774          0.655    0.073    9.030    0.000
   C11S775          0.108    0.035    3.107    0.002
   C11S776          0.673    0.057   11.729    0.000
   C11S777          0.745    0.071   10.514    0.000
```

As expected, given the large sample size, the χ^2 test statistic is statistically significant; however, given a CFI value of ≥ 0.90, an RMSEA value ≤ 0.06, and a SRMR value ≤ 0.08, the overall fit of the model is basically good (9). As such, Model 1 results are interpretable, and we highlight that the standardized path coefficient of OR52E4 is statistically significant, although its magnitude is less than that of age or smoking. Furthermore, we note that this single gene model only explains ~0.15 of the variance in Q1.

2.2.3. Fit Other Models and Compare

Once the analyst has examined the results of the first model, decisions must be made on how to modify the model. This is the crux of SEM modeling. The goal is to find the most plausible, best-fitting model. When assessing changes to make to the initial model, a variety of issues may be considered. Which path coefficients are not statistically significant? Should they remain in the model because they are biologically or epidemiologically important? What other relationships are worth examining—how would these be depicted in the model? Once another model is fitted, another set of statistics will be output as above (path coefficients, R^2, and fit statistics).

Specifically, the following statistical issues should be considered when comparing models. Are the R^2 values good? Do some indicators of latent variables have lower R^2 values and, if so, should those be removed? Sometimes a path coefficient, though not statistically significant by itself, may contribute to the overall fit of the model, such that the inclusion of that variable results in a better AIC, RMSEA, and/or CFI. Then it is up to the analyst which model is "better." It is then important to replicate findings in an independent dataset (19).

Finally, we must comment about the modification of models before arriving at one with the "best fit." Wright (10) advocated careful thought and prior knowledge for the comparison of alternative models. All models should be based on substantive theory and causal conjectures (2). We and others (2, 3) recommend against a "quasi-random walk" through a sequence of models and instead promote theoretical justification of all models.

Example: If we add in another gene, OR2T3, which is biologically related to OR52E4, to our first model, the fit of the model is slightly better and the amount of variance explained in Q1 increases to ~0.19.

Example: **Part 2**: OR2T3 and OR52E4 Genes on Q1 Simulated Trait (Model 2)

Mplus v5.1 Input File:

```
TITLE: GAW17 Q1 UNRELATEDS REP137 OR2T3 AND OR52E4
 DATA:
  FILE IS " _ "; !location of data file
 FORMAT IS FREE;
  LISTWISE = ON; !ensures no missing data used

 VARIABLE:
  NAMES ARE id sex age smoke pop1 Q1
  !OR2T3
  C1S11507 C1S11510 C1S11511 C1S11520 C1S11522 C1S11523
  !OR52E4
  C11S773 C11S774 C11S775 C11S776 C11S777

 USEVARIABLES ARE id sex age smoke pop1 Q1
  C1S11507 C1S11510 C1S11511 C1S11520 C1S11522 C1S11523
  C11S773 C11S774 C11S775 C11S776 C11S777;

  MISSING ARE .;
  IDVARIABLE IS id;

ANALYSIS:
  TYPE=GENERAL;
  ESTIMATOR=MLR;
MODEL:
  OR2T3 BY C1S11510 C1S11507 C1S11511 C1S11520 C1S11522 C1S11523;
  OR52E4 BY C11S774 C11S773 C11S775 C11S776 C11S777;
  OR2T3 ON pop1;
  OR52E4 ON pop1;
  Q1 ON age smoke sex OR2T3 OR52E4; !potl association of both genes are being
evaluated
  OR2T3 WITH OR52E4;  !correlation added because of known biological relatedness
```

Mplus v5.1 Output File (Partial):

```
INPUT READING TERMINATED NORMALLY
GAW17 Q1 UNRELATEDS REP137 OR2T3 AND OR52E4
SUMMARY OF ANALYSIS
Number of groups                              1
Number of observations                      697
Number of dependent variables                12
Number of independent variables               4
Number of continuous latent variables         2

THE MODEL ESTIMATION TERMINATED NORMALLY
TESTS OF MODEL FIT
Chi-Square Test of Model Fit
     Value                            228.214*
     Degrees of Freedom                95
     P-Value                             0.0000
     Scaling Correction Factor           1.554
      for MLR

*  The chi-square value for cannot be used for chi-square difference tests.  See chi-
square difference testing in the index of the Mplus User's Guide.
CFI/TLI
     CFI                               0.911
     TLI                               0.894
Loglikelihood
     H0 Value                      -8178.021
     H0 Scaling Correction Factor      3.508
      for MLR
     H1 Value                      -8000.741
     H1 Scaling Correction Factor      2.163
      for MLR
RMSEA (Root Mean Square Error Of Approximation)
     Estimate                          0.045

SRMR (Standardized Root Mean Square Residual)
     Value                             0.046

STANDARDIZED MODEL RESULTS
STDYX Standardization
                                                Two-Tailed
                    Estimate      S.E.  Est./S.E.  P-Value

 OR2T3    BY
   C1S11510         0.889     0.031     28.819     0.000
   C1S11507         0.570     0.068      8.381     0.000
   C1S11511         0.780     0.035     22.557     0.000
   C1S11520         0.545     0.060      9.015     0.000
   C1S11522         0.201     0.070      2.874     0.004
   C1S11523         0.625     0.054     11.581     0.000

 OR52E4   BY
   C11S774          0.806     0.045     17.883     0.000
   C11S773          0.345     0.066      5.258     0.000
   C11S775          0.328     0.053      6.160     0.000
   C11S776          0.821     0.034     23.823     0.000
   C11S777          0.865     0.040     21.612     0.000
```

```
OR2T3    ON
  POP1            0.146    0.037    3.960    0.000

 OR52E4   ON
  POP1            0.036    0.041    0.877    0.380

 Q1         ON
  OR2T3           0.205    0.045    4.551    0.000
  OR52E4          0.109    0.042    2.568    0.010

 Q1         ON
  AGE             0.299    0.033    8.969    0.000
  SMOKE           0.181    0.034    5.347    0.000
  SEX            -0.036    0.034   -1.068    0.285

 OR2T3    WITH
  OR52E4          0.309    0.064    4.854    0.000

 Intercepts
  Q1             -0.709    0.139   -5.114    0.000
  C1S11507        0.095    0.036    2.610    0.009
  C1S11510        0.142    0.054    2.652    0.008
  C1S11511        0.287    0.052    5.511    0.000
  C1S11520        0.165    0.036    4.571    0.000
  C1S11522        0.129    0.023    5.669    0.000
  C1S11523        0.145    0.039    3.678    0.000
  C11S773         0.412    0.031   13.119    0.000
  C11S774         0.203    0.054    3.748    0.000
  C11S775         1.575    0.066   23.761    0.000
  C11S776         0.317    0.057    5.586    0.000
  C11S777         0.246    0.059    4.191    0.000

 Residual Variances
  Q1              0.808    0.029   27.616    0.000
  C1S11507        0.675    0.077    8.719    0.000
  C1S11510        0.210    0.055    3.825    0.000
  C1S11511        0.392    0.054    7.262    0.000
  C1S11520        0.703    0.066   10.665    0.000
  C1S11522        0.960    0.028   34.184    0.000
  C1S11523        0.609    0.067    9.036    0.000
  C11S773         0.881    0.045   19.469    0.000
  C11S774         0.351    0.073    4.833    0.000
  C11S775         0.893    0.035   25.586    0.000
  C11S776         0.326    0.057    5.769    0.000
  C11S777         0.252    0.069    3.638    0.000

  OR2T3          0.979    0.011    90.297    0.000
  OR52E4         0.999    0.003   345.578    0.000

R-SQUARE

  Observed                          Two-Tailed
  Variable      Estimate     S.E.  Est./S.E.   P-Value
  Q1               0.192    0.029      6.554     0.000
[we will refer to the above result later on]
  C1S11507         0.325    0.077      4.191     0.000
  C1S11510         0.790    0.055     14.409     0.000
  C1S11511         0.608    0.054     11.279     0.000
  C1S11520         0.297    0.066      4.507     0.000
  C1S11522         0.040    0.028      1.437     0.151
  C1S11523         0.391    0.067      5.790     0.000
  C11S773          0.119    0.045      2.629     0.009
  C11S774          0.649    0.073      8.942     0.000
  C11S775          0.107    0.035      3.080     0.002
  C11S776          0.674    0.057     11.911     0.000
  C11S777          0.748    0.069     10.806     0.000
```

Although both gene constructs are significantly associated with Q1 (Model 2), the magnitude of the path coefficient for OR2T3 is larger than for OR52E4. It is also interesting to note that the path coefficient of OR52E4 is attenuated when OR2T3 is included (Model 2) compared to that when OR2T3 is not included (Model 1). We can also see via the magnitude and significance of the standardized coefficients that population structure (pop1) is more influential on OR2T3 than on OR52E4.

2.2.4. Presentation of "Final" Model

Once the best model has been selected, presenting it for publication is also not trivial. Often, the path diagram is very complex, including indicators for latent variables with corresponding factor loadings, correlations between all latent variables, and, of course, the causal pathways. The authors will want to present P-values for each path coefficient, and also some assessment of the goodness-of-fit of the final model. Clearly, presentation of every model considered with their fit statistics would be out of the question. The authors should make every attempt to draw the path diagram simply and clearly. One may consider listing factor loadings for latent variables in a separate table, and providing correlations between latent variables in another table, so that the path diagram is not too cluttered.

McDonald and Ho (2) propose several guidelines for presenting SEM results in addition to those stated above. The report should give theoretical grounds for the presence or absence of each causal path ("arc") in the model and also some discussion about the use of causal pathways instead of correlations. If space allows, the full covariance matrix of the observed variables should be provided; if not, means and standard deviations of each variable are sufficient. Also, the global χ^2 statistic should also be provided, in addition to other fit statistics, such as RMSEA and CFI.

In Fig. 3, we present the final model (Model 2) from our worked example. Here, we present SNPs in rectangles, genes as ovals since they are modeled as latent variables and provide the factor loadings±standard errors above the single-headed arrows directed from the SNPs to the gene. Q1 is an observed continuous trait, and thus is presented as a rectangle. Similarly, age, sex, and smoking status are observed covariates, and thus represented by rectangles. The correlation between OR52E4 and OR2T3 is represented by a double-headed arrow between the two latent variables.

3. Notes

1. As might be expected with models that include many pathways and many variables, particularly many latent variables, model convergence might be a problem. One thing to look at is the existence of (phenotypic) outliers in the data. If there are

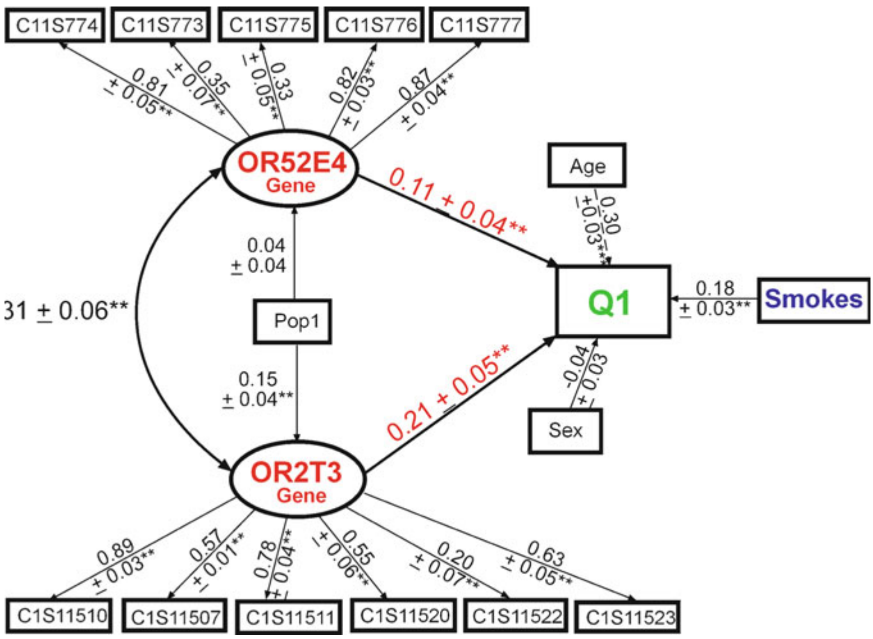

Fig. 3. Modeling the aggregate effects of common and rare variants in multiple potentially interesting genes using latent variable SEM. Model of the associations between two genes (11 SNPs) and potential associations with Q1 (CFI = 0.91; RMSEA = 0.04; SRMR = 0.03). Standardized loadings and standard errors are shown above the *arrows*. $^*p \leq 0.05$; $^{**}p \leq 0.01$. Residuals are not shown for clarity.

outliers in the observed variables, the removal of these datapoints may enable the model to converge. In addition, model over- or under-specification may result in problems with model convergence. As stated before, evaluating the identification of the model is difficult. If faced with this concern, it is best to draw the model, then write out the simultaneous equations evaluated within the model, and then apply the aforementioned rules described by Bollen (1) to assess identification. A final method to increase model convergence is to increase the sample size. Since SEM is really the estimation of simultaneous regression equations, a similar rule of thumb may be applied: at least 20 observations per variable are recommended, but more is better.

2. The analysis of binary or categorical traits is not trivial. Most of the original SEM methodology was developed for quantitative traits and made the assumption of multivariate normality and linear causal effects. The so-called “Asymptotic Distribution Free” approach to model fitting relaxes the assumption of multivariate normality. However, it does not relax the assumption of linearity, and it has been shown that in finite samples, its behavior is quite poor (22). Numerous methods have been developed that explicitly model categorical traits using a threshold model. That is, it is assumed that an underlying quantitative

multivariate normal trait exists which belongs to a specific category if it falls into a specific range. Some software packages such as Mplus, GLLAMM, and OpenMx support such explicit models for various types of categorical traits. For instance, the MLR estimator in Mplus is robust to non-normality and can be used for categorical variables. In general, we recommend that modelers who have categorical traits avoid using software that does not support an explicit model for such categorical traits.

References

1. Bollen K (1989) Structural equations with latent variables John Wiley & Sons, New York
2. McDonald RP, Ho MH (2002) Principles and practice in reporting structural equation analyses. Psychol. Methods **7**: 64–82
3. Pearl J (2000) Causality: Models, Reasoning, and Inference Cambridge University Press, New York, New York
4. Bollen K (2001) Indicator: Methodology, in Internatinoal Encyclopedia of the Social and Behavioral Sciences (Smesher, N. and Baltes, P., Eds.) pp 7282–7287, Elsevier Sciences, Oxford
5. Sobel M (1982) Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology **13**: 290–312
6. Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers. Soc. Psychol. **51**: 1173–1182
7. Muthén BO (1984) A general structural equation model with dichotomous ordered categorical and continuous latent variable indicator. Psychometrika **49**: 115–132
8. Hancock GR, Mueller RO (2006) Structural Equation Modeling: A Second Course Information Age Publishing, Inc., Greenwich CT
9. Hu LT, Bentler PM, (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural equation modeling **6**: 1–55
10. Wright S (1923) The Theory of Path Coefficients: A Reply to Niles's Criticism. Genetics **8**: 239–255
11. Neale M, Cardon LR (1992) Methodology for Genetic Studies of Twins and Families Kluwer Academic Publishers, Dordrecht, The Netherlands
12. Rao DC (1985) Application of path analysis in human genetics, in Multivariate analysis (Krishnaiah PR, Ed.) pp 467–484, Elsevier Science Publishers
13. Nock NL, Larkin EK, Morris NJ, Li Y, Stein CM (2007) Modeling the complex gene x environment interplay in the simulated rheumatoid arthritis GAW15 data using latent variable structural equation modeling. BMC. Proc **1** Suppl 1: S118
14. Nock NL, Wang X, Thompson CL, Song Y, Baechle D, Raska P, Stein CM, Gray-McGuire C (2009) Defining genetic determinants of the Metabolic Syndrome in the Framingham Heart Study using association and structural equation modeling methods. BMC Proc. **3** Suppl 7: S50
15. Todorov AA, Vogler GP, Gu C, Province MA, Li Z, Heath AC, Rao DC (1998) Testing causal hypotheses in multivariate linkage analysis of quantitative traits: general formulation and application to sibpair data. Genet Epidemiol **15**: 263–278
16. Morris NJ, Elston RC, Stein C M (2011) A Framework for Structural Equation Models in General Pedigrees. Hum Hered **70**: 278–286
17. Buhi ER, Goodson P, Neilands TB (2007) Structural equation modeling: a primer for health behavior researchers. Am J Health Behav. **31**: 74–85
18. Province MA, Rice TK, Borecki IB, Gu C, Kraja A, Rao DC (2003) Multivariate and multilocus variance components method, based on structural relationships to assess quantitative trait linkage via SEGPATH. Genet Epidemiol **24**: 128–138
19. Bollen K (1998) Structural Equation Models, in Encyclopedia of Biostatistics (Armitage, P. and Colton, T., Eds.) pp 4363–4372, John Wiley & Sons, Sussex, England
20. Mulaik S (2007) There is a place for approximate fit in structural equation modeling. Personality and Individual Differences **42**: 883–891
21. Barrett P (2007) Structural equation modeling: Adjudging model fit. Personality and Individual Differences **42**: 815–824
22. Curran PJ, Finch JF, West SG (1996) The robustness of tests statistics to nonnormality and specification error in confirmatory factor analysis. Psychol. Methods **1**: 16–29