An introduction to structural equation modeling in genetics

Mykhaylo M. Malakhov

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455, USA

1 Introduction

Genome-wide association studies (GWAS) are arguably the most popular study type in the field of statistical genetics. With the increasing availability of inexpensive yet decently accurate microarray sequencing technologies, it has become feasible to genotype cohorts with tens of thousands of individuals. A GWAS then involves finding the univariate associations between each genotyped single-nucleotide polymorphism (SNP) and a phenotype of interest for individuals in the cohort. Such studies have successfully identified thousands of genetic loci associated with human traits, ranging from anthropometric traits such as standing height to complex diseases such as Alzheimer's. Yet despite their popularity, GWAS have a number of severe limitations.

Much research has suggested the existence of widespread genetic pleiotropy – the situation where a genetic variant acts along multiple functional pathways to influence multiple, seemingly unrelated traits (Figure 1). Indeed, the general scientific consensus is that constellations of phenotypes are affected by shared sources of genetic liability. However, GWAS and other commonly-used univariate statistical methods ignore that shared information and treat each SNP-trait association as independent of the rest, resulting in a loss of statistical power and an incomplete picture of the genetic architecture of complex traits. Recently several methods have been proposed that explicitly account for pleiotropic effects and allow us to perform multivariate GWAS. In this expository report we consider one such approach that is based on structural equation models.

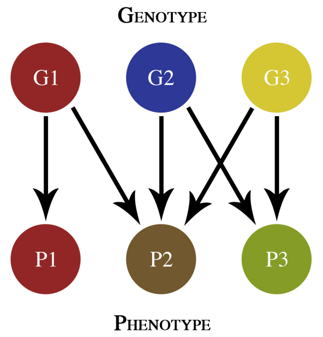


Figure 1: An illustration of additive genetic pleiotropy.

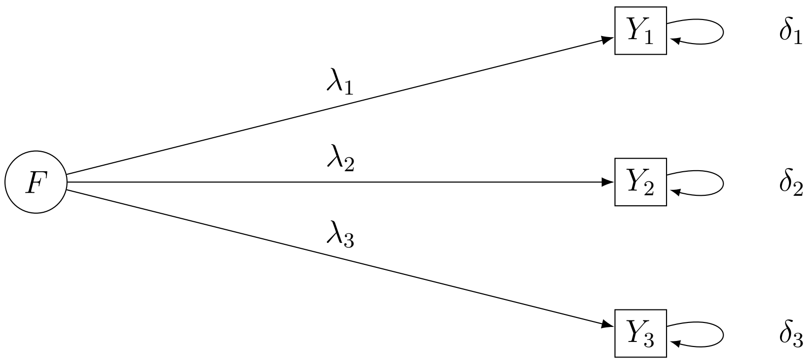


Figure 2: Path diagram for a single-factor structural model.

2 Review of SEM

In order to explain how structural equation modeling (SEM) can be used for genetic association studies, we first review basic concepts and approaches related to SEM more broadly. SEM can be defined as a diverse set of methods for posing, fitting, and comparing causal latent variable models. A *causal* model is one that assumes some causal structure among a set of variables; such models are often represented as path diagrams where arrows between variables indicate the direction of causality between them. A *latent variable* is a variable that is assumed to exist but is not observed explicitly, and a latent variable model is simply a model that includes latent variables. Accordingly, a typical structural equation model includes both measured (i.e. observed) and latent (i.e. unobserved) variables and imposes some set of causal relationships between them.

Figure 2 is a path diagram that represents a simple structural equation model. This model contains one latent variable F and three observed variables Y_1, Y_2, and Y_3. Note that in SEM latent variables are denoted by circles, while observed variables are placed in boxes. When a latent variable is hypothesized to explain several "higher-level" observed variables, such as in this model, it is known as a *factor.* The directed edges between factor F and the three measured variables indicate that F has causal effects on Y_1, Y_2, and Y_3 with effect sizes λ_1, λ_2, and λ_3 respectively. In this setting the λ_i's are commonly referred to as *factor loadings.* For each of the three observed variables, the directed loop represents the (unobserved) residual variance not explained by the common factor. This residual random error has mean δ_i for variable Y_i, where $i = 1, 2, 3$. Crucially, notice that the observed variables in this particular model are correlated only through their relationships with F. This structure allows us to explicitly model the covariance of Y_1, Y_2, and Y_3 according to the hypothesized causal architecture.

Mathematically, the model in Figure 2 can be stated as

$$\begin{aligned} Y_1 &= \lambda_1 F + \delta_1 \\ Y_2 &= \lambda_2 F + \delta_2 \\ Y_3 &= \lambda_3 F + \delta_3 \end{aligned}$$

Standard SEM	**Genomic SEM**
Survey results	GWAS summary statistics
Items	Phenotypes
Factors	Genetic liabilities

Table 1: Comparison of a typical SEM use case with genomic SEM.

where δ_1, δ_2, and δ_3 are normally distributed random variables while F is not stochastic. (Here we slightly abuse notation by now using the δ_i's to refer to the random errors themselves instead of their expectations, but this is common practice and should not cause confusion.) We assume that $Cov(F, \delta_i) = 0$ for all i, $Cov(\delta_i, \delta_j) = 0$ whenever $i \neq j$, and $Cov(Y_i, Y_j \mid F) = 0$ for all i, j. Note that all of these equations and assumptions can be inferred from the path diagram.

Since a structural equation model consists of a set of linear regressions which must be fit simultaneously, we cannot simply use ordinary least squares approaches to fit the model. Instead, structural models are fit by minimizing a loss function that represents the "distance" between the empirical covariance matrix calculated from the data and the model-implied covariance matrix. In section 4.1 we will describe two approaches for doing so: weighted least squares and maximum likelihood estimation.

Historically SEM has enjoyed the greatest popularity in the social sciences, where it is often used to analyze questionnaire data. SEM is particularly conducive to such analyses since questionnaires consist of multiple related questions (known as *items*) that all seek to measure the same underlying causal construct. For example, a psychological study might analyze participants' responses to three questions about their mood, self-esteem, and level of interest. If the researchers hypothesize that those three items are all different manifestations of major depressive disorder, then the model discussed above would be a good choice for analyzing the study data in light of their assumed hypothesis.

Finally, we introduce two important categories of SEM. *Exploratory factor analysis* (EFA) has the goal of uncovering the underlying causal structure of a given system. When performing an EFA multiple models are considered, each with a different number of factors that are assumed to be associated with all of the observed variables. The model that appears to explain the data "best" is selected. In *confirmatory factor analysis* (CFA), the goal is to test a model's fit and make inferences about factor effects. Thus, the model structure and constraints are pre-specified beforehand for CFA. In other words, EFA involves comparing the plausibility of several hypothesized causal structures while CFA seeks to analyze the data in light of a given hypothesized causal structure.

3 SEM of genetic architecture

Recently Grotzinger et al. (2019) introduced an SEM framework for modeling the shared genetic architecture of constellations of traits [1]. The key idea behind their novel approach is to use structural equations to explicitly model the genetic covariance of multiple traits by treating broad genetic liabilities as latent variables. Table 1 summarizes how the core components of "standard" SEM correspond to their analogs in the genomic SEM framework.

Hence the background discussed in section 2 fully applies to genomic SEM, except that items now represent GWAS traits and factors account for the (hypothesized) genetic causes of those traits. Since genetic factors are modeled as latent variables, they do not correspond to any specific levels of molecular organization (e.g. SNPs or genes) but rather represent broad sources of genetic variation. This grants considerable flexibility to genomic SEM. One primary use case is to conduct a CFA for multivariate genetic associations among traits with (possibly multiple) shared sources of genetic variation. Genomic SEM can also be used to conduct GWAS for constellations of genetically correlated phenotypes; to do so, one considers a single-factor model and includes SNPs one-by-one as measured variables with an effect on the factor alone. Those multivariate GWAS associations can then inform any subsequent analysis, such as the construction of polygenic risk scores for a constellation of phenotypes. Moreover, the authors of the study propose the test statistic Q_{SNP}, calculated from a set of fitted genomic SEM models, that can identify genetic variants which cause divergence between traits.

In addition to its flexibility, another significant advantage of genomic SEM over other multivariate GWAS methods is its ease of use. Genomic SEM only requires GWAS summary statistics, which are freely available for thousands of traits, and a reference panel such as 1000 Genomes. The results of genomic SEM remain unbiased even if the summary statistics come from GWAS with unknown levels of overlap. Furthermore, Grotzinger et al. developed the R package `GenomicSEM` (`https://github.com/GenomicSEM/GenomicSEM`) that includes tools for calculating an empirical genetic covariance matrix, fitting a user-specified structural equation model, calculating the Q_{SNP} statistic, and performing multivariate GWAS.

3.1 Case studies: CFA and multivariate GWAS

To illustrate how genomic SEM is used for factor analysis and multivariate GWAS, we discuss some results from [1]. In their paper, Grotzinger et al. present a factor analysis of neuroticism based on GWAS summary statistics from the UK Biobank (UKBB). First, they obtained summary statistics for 12 neuroticism items (mood, misery, irritability, sensitivity/hurt feelings, fed-up feelings, nervousness, worry, tension, embarrassment, nerves, loneliness, and guilt) from Round 1 of the UK Biobank GWAS performed by the Neale Lab [2]. After applying some standard quality control filters, the authors used their proposed method to compute a genetic correlation matrix for the neuroticism items and then performed an EFA on the correlation matrix using the `fa` package in R. This yielded three potential structural models of neuroticism: a single-factor model, a two-factor model, and a three-factor model. Next, they performed a CFA on each of the model structures by fitting it in `GenomicSEM`. Causal pathways with factor loadings < 0.4 were dropped, yielding the fitted models shown in Figure 3. (Note that these figures use colors to distinguish between latent and observed variables, and place residual errors in their own circles.)

The goodness of fit for a structural equation model is often evaluated using its model χ^2 index, which is a likelihood-ratio statistic comparing the fit of the proposed structural equation model against the fit of a saturated model. Model χ^2 statistics can be computed in genomic SEM as well. (See section 4.2 for further details on goodness of fit measures in

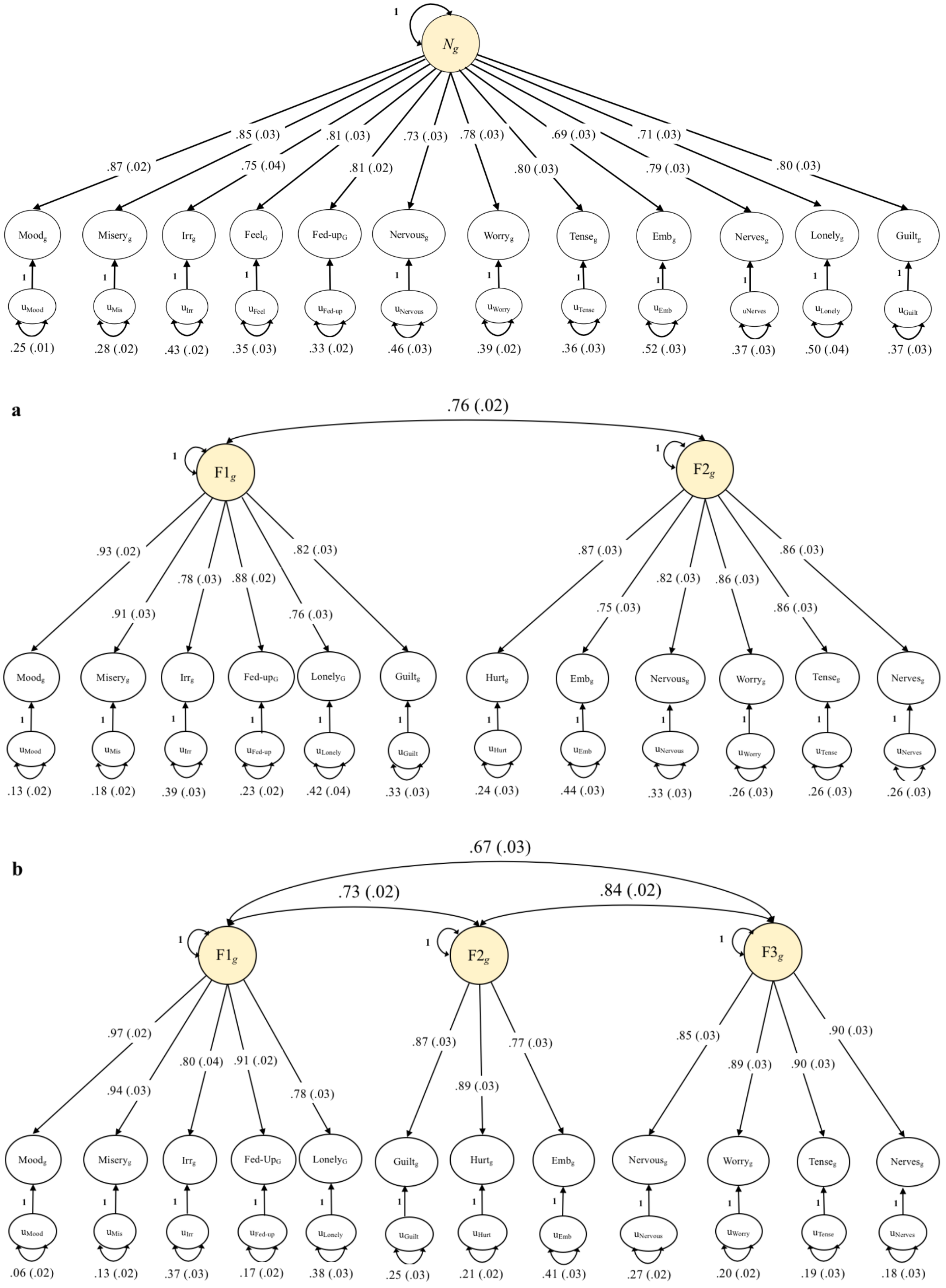
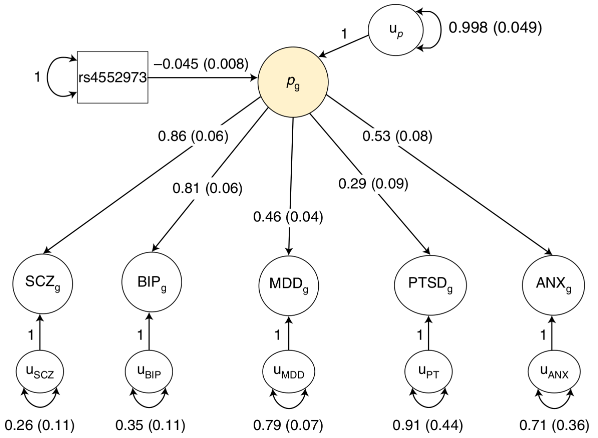
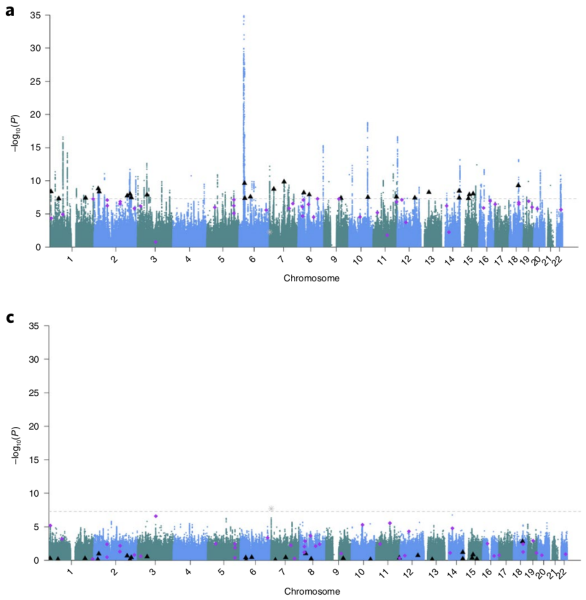


Figure 3: CFA of neuroticism items using genomic SEM, fitted on UKBB summary statistics. Path labels denote standardized effect sizes, with standard errors shown in parentheses.



(a) Structural equation model of the p-factor with SNP-level effects. Standardized effect sizes and standard errors are from the fitted model with the most significant SNP (rs4552973).



(b) Manhattan plots for a multivariate GWAS of the p-factor. The top panel shows P-values for SNP effects, while the bottom panel shows P-values for Q_{SNP}.

genomic SEM.) In this example study the single-factor, two-factor, and three-factor models of neuroticism items respectively yielded model χ^2 indices of $\chi^2(54) = 4,884.10$; $\chi^2(53) = 2,758.18$; and $\chi^2(51) = 1,879.31$. A lower χ^2 statistic indicates better fit, so these results suggest the presence of three broad genetic factors for neuroticism.

The authors also used the CFA of neuroticism to showcase their Q_{SNP} statistic. The Q_{SNP} statistic for a given genetic variant quantifies the extent to which the variant's (univariate) effects on each trait can be explained by the specified structural model. If a variant's Q_{SNP} is statistically significant, then we reject the null hypothesis that our specified structural model explains the variant's effects on all included traits; in other words, we conclude that the variant is responsible for heterogeneity among the traits. As such, Q_{SNP} is analogous to the Q statistic of heterogeneity in meta-analysis. The authors calculated Q_{SNP} for every GWAS variant in each of the three neuroticism models defined above. The single-factor model had 69 significant Q_{SNP} hits, the two-factor model had 28 significant Q_{SNP} hits, and the three-factor model had 20 significant Q_{SNP} hits. These results further support the theory that three separate genetic factors account for neuroticism traits, since the three-factor model can adequately explain the effects of more SNPs than the other two models.

As a second application of genomic SEM, Grotzinger et al. conducted a multivariate GWAS of psychopathology traits. Many cognitive scientists hypothesize that psychopathology traits are manifestations of a single p-factor that has a genetic basis. Thus, rather than explore multiple model structures, the authors only considered a genomic structural equation model with a single factor that affects all of the psychopathology traits. They again obtained summary statistics from the Neale Lab Round 1 GWAS of the UK Biobank, but this time for the psychopathology phenotypes: schizophrenia, bipolar disorder, major depressive disorder, post-traumatic stress disorder, and anxiety. Then they extracted approximately uncorrelated SNPs from the intersection of the respective GWAS for these traits by clumping, and included them one-by-one as observed variables with an effect on the p-factor. That is, if some M SNPs remain after performing clumping on the intersection of GWAS, then M separate genomic SEM models must be fitted. One of these fitted models is shown Figure 4a. Thus, this approach yields estimates of the effect size (and standard error) for the association between each considered variant and the shared genetic factor (in this case, the p-factor of psychopathology). 128 independent loci were genome-wide significant for the p-factor, of which 27 were not identified by any of the contributing (univariate) GWAS. The top panel of Figure 4b is a Manhattan plot showing the $-\log_{10}(P)$ values for all SNPs in the multivariate GWAS.

The authors also calculated the Q_{SNP} statistic for each variant considered in their multivariate GWAS of psychopathology. A Manhattan plot of the results is shown in the bottom panel of Figure 4b. Note that there was only one significant Q_{SNP} hit, which indicates that only one SNP is responsible for heterogeneity among the phenotypes. Hence, all other SNPs are adequately explained by the single-factor model. This finding provides support for the p-factor theory of psychopathology and lends credence to the multivariate GWAS.

4 Genomic SEM methodology

Now that we have demonstrated how to use and interpret genomic SEM on a conceptual level, we provide the technical details of its methodology [1]. Fitting a genomic structural equation model involves two stages. First, an empirical genetic covariance matrix and its associated sampling covariance matrix are calculated. Second, a loss function representing the "distance" between the empirical covariance matrix and the model-implied covariance matrix is minimized. After the minimization algorithm converges, the sampling covariance matrix is used to robustly estimate standard errors for all estimated coefficients. Competing fitted models can be compared using the model χ^2 statistic and other goodness of fit measures, which we also discuss.

4.1 Model fitting

The empirical genetic covariance matrix (without SNP effects) for k traits is defined as

$$S_{LDSC} = \begin{pmatrix} h_1^2 & & & \\ \sigma_{g1,g2} & h_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{g1,gk} & \sigma_{g2,gk} & \cdots & h_k^2 \end{pmatrix}$$

where h_i^2 is the heritability of phenotype i and $\sigma_{gi,gj} = r_{gi,gj}\sqrt{h_i^2 h_j^2}$ is the genetic covariance between phenotypes i and j. By default, the `GenomicSEM` package calculates SNP heritabilities from summary statistics using linkage disequilibrium score regression (LDSC) [3], although a recent update has introduced the option of heritability calculation with the High-definition Likelihood method [4] instead.

If the contributing GWAS have sample overlaps, then the correlations in S_{LDSC} will be biased. Hence, genomic SEM also relies on a sampling covariance matrix. Without SNP effects, this matrix is defined as

$$V_{S_{LDSC}} = \begin{pmatrix} \text{s.e.}(h_1^2)^2 & & & & & \\ \text{cov}(h_1^2, \sigma_{g1,g2}) & \text{s.e.}(\sigma_{g1,g2})^2 & & & & \\ \vdots & \vdots & \ddots & & & \\ \text{cov}(h_1^2, \sigma_{g1,gk}) & \text{cov}(\sigma_{g1,g2}, \sigma_{g1,gk}) & \text{s.e.}(\sigma_{g1,gk})^2 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \text{cov}(h_1^2, h_j^2) & \text{cov}(\sigma_{g1,g2}, h_j^2) & \text{cov}(\sigma_{g1,gk}, h_j^2) & \text{s.e.}(h_j^2)^2 & & \\ \vdots & \vdots & \vdots & & \ddots & \\ \text{cov}(h_1^2, \sigma_{gj,gk}) & \text{cov}(\sigma_{g1,g2}, \sigma_{gj,gk}) & \text{cov}(\sigma_{g1,gk}, \sigma_{gj,gk}) & \text{cov}(h_j^2, \sigma_{gj,gk}) & \text{s.e.}(\sigma_{gj,gk})^2 & \\ \text{cov}(h_1^2, h_k^2) & \text{cov}(\sigma_{g1,g2}, h_k^2) & \text{cov}(\sigma_{g1,gk}, h_k^2) & \text{cov}(h_j^2, h_k^2) & \text{cov}(\sigma_{gj,gk}, h_k^2) & \text{s.e.}(h_k^2)^2 \end{pmatrix}.$$

Intuitively, $V_{S_{LDSC}}$ is the sampling covariance matrix of the $k(k+1)/2$ non-redundant elements of S_{LDSC}. That is, the diagonal elements are sampling variances and the off-diagonal elements are sampling covariances of the non-redundant elements of S_{LDSC}. These covariances are estimated using an extension of the jackknife resampling procedure from LDSC.

We again emphasize that these covariance matrices are for genomic structural equation models without SNP effects. As discussed above in section 3.1, single-SNP effects are included in the structural equation model when conducting a multivariate GWAS. In this case the empirical covariance matrix S_{LDSC} is expanded to include covariances between the SNP and each of the k phenotypes. Consequently, the sampling covariance matrix V_S becomes much larger and requires SNP variance data estimated from a reference panel. Since these covariance matrices are rather cumbersome yet conceptually similar to the ones without SNP effects, we refer the reader to the original genomic SEM paper [1] for their derivation.

The second stage involves iteratively computing the model-implied covariance matrix and finding parameter values that minimize its “distance” to the empirical covariance matrix S_{LDSC}. A genomic structural equation model can be specified in terms of two separate models. The measurement model

$$y = \Lambda\eta + \varepsilon \tag{1}$$

represents the relationships between a $k \times 1$ vector of indicators for the observed traits (y) and an $m \times 1$ vector of the latent genetic factors (η). Then Λ is a $k \times m$ matrix of factor loadings, and ε is a $k \times 1$ vector of residuals. Additionally we have the structural model

$$\eta = B\eta + \zeta \tag{2}$$

which encapsulates relationships between the latent variables (i.e. the genetic factors). Here B is an $m \times m$ matrix of regression coefficients that relate the latent variables to each other, and ζ is an $m \times 1$ vector of latent variable residuals. Then the model-implied covariance matrix is

$$\Sigma(\theta) = \Lambda(I - B)^{-1}\Psi((I - B)^{-1})^T\Lambda^T + \Theta \tag{3}$$

where Ψ is the latent variable covariance matrix and Θ is a matrix of covariances among the residuals. Note that for single-factor models, the structural model (Equation 2) is not needed and the model-implied covariance matrix simplifies to

$$\Sigma(\theta) = \Lambda\Psi\Lambda^T + \Theta. \tag{4}$$

Grotzinger et al. derive both weighted-least squares (WLS) and maximum likelihood (ML) estimation methods for genomic SEM, but they only present results obtained using WLS in the paper. In particular, the authors use the diagonally weighted version of WLS because it is more tractable and stable for large matrices. This method minimizes the loss function

$$F_{WLS}(\theta) = (s - \sigma(\theta))^T D_s^{-1} (s - \sigma(\theta)). \tag{5}$$

Here s and $\sigma(\theta)$ are half-vectorized versions of the empirical covariance matrix S and the model-implied covariance matrix $\Sigma(\theta)$, respectively, while D_s is the sampling covariance matrix V_S with its off-diagonal elements set to 0. Maximum likelihood estimation for genomic SEM instead minimizes the loss function

$$F_{ML}(\theta) = \log|\Sigma(\theta)| - \log|S| + tr\{S\Sigma^{-1}(\theta)\} - k. \tag{6}$$

To account for potential sample overlap in the contributing GWAS summary statistics, standard errors for the model parameter estimates are derived using the empirical sampling covariance matrix V_S. Essentially, genomic SEM uses a sandwich estimator to obtain the robust covariance matrix of the model parameter estimates:

$$V_\theta = (\hat{\Delta}^T \Gamma^{-1} \hat{\Delta})^{-1} \hat{\Delta}^T \Gamma^{-1} V_S \Gamma^{-1} \hat{\Delta} (\hat{\Delta}^T \Gamma^{-1} \hat{\Delta})^{-1}. \tag{7}$$

Here $\hat{\Delta}$ is the matrix of model derivatives evaluated at the parameter estimates and Γ is the naive weight matrix for the fitted model.

4.2 Fit and heterogeneity statistics

Comparing goodness of fit for a set of several models is an important step in factor analysis. Even when only considering one model, it is important to verify that the model's structure is appropriate by testing for goodness of fit. SEM relies on a number of different goodness of fit measures, so Grotzinger et al. derive analogous measures for genomic SEM in particular.

Model χ^2, which we considered when discussing a genomic SEM analysis of neuroticism in section 3.1, is an index of the exact fit of a structural equation model. Roughly speaking, it reflects the difference between the empirical covariance matrix S and the model-implied covariance matrix $\Sigma(\theta)$. The difference between χ^2 statistics of two nested models can also be used to formally test whether one model fits the data better than the other. Model χ^2 is typically derived for structural models using a formula that depends on the sample size, but since no meaningful measure of sample size exists for genomic structural equation models, the authors found a novel derivation of model χ^2 that incorporates the sampling covariance matrix of the model residuals. This derivation is somewhat lengthy, but it nicely produces a test statistic that follows a $\chi^2(r)$ distribution with $r = k^* - f_p$, where k^* is the number of non-redundant elements in S and f_p is the number of freely-estimated model parameters.

Several other goodness of fit measures are derived from model χ^2. The comparative fit index (CFI) reflects the extent to which the proposed model fits better than a null model. In particular, the CFI statistic derived for genomic SEM compares the proposed model to one where all phenotypes are heritable but genetically uncorrelated. It can be calculated as

$$CFI = \frac{f(\text{independence model}) - f(\text{proposed model})}{f(\text{proposed model})}$$

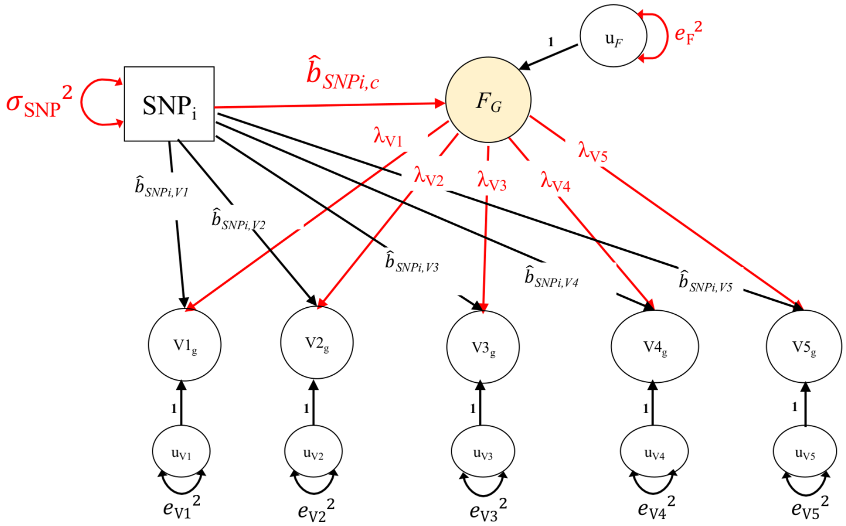


Figure 5: Example computation of the Q_{SNP} statistic. In step 1 the parameters colored in red are estimated. In step 2 the red-colored parameters are fixed to their values from step 1, while the parameters colored in black are freely estimated.

where $f = \chi^2 - r$. It follows that CFI values range from 0 to 1, with higher values indicating better fit. The SEM literature generally considers CFI $\geq$ 0.90 to indicate acceptable fit and CFI $\geq$ 0.95 to indicate good fit. Another measure of goodness of fit is AIC, which is a relative fit index that seeks to balance fit with parsimony. It can also be derived from the model χ^2 by

$$\text{AIC} = \chi^2 + 2f_p.$$

AIC can be used to compare multipled (not necessarily nested) genomic SEM models, with lower AIC considered to be better.

The last goodness of fit index implemented for genomic SEM is the standardized root mean squared residual (SRMR). This index is also popular in SEM, and is defined as the standardized root mean squared difference between the empirical and model-observed covariance matrices. Hence, it can be naturally calculated from the S and $\Sigma(\theta)$ matrices in genomic SEM. This index reflects approximate model fit, with lower values indicating better fit. Generally, SEM literature regards SRMR values to indicate acceptable fit when < 0.10 and good fit when < 0.05.

An important contribution of Grotzinger et al. is the Q_{SNP} statistic, which quantifies the extent to which a specified structural equation model explains the univariate effects of a SNP on each modeled trait. Q_{SNP} is calculated through the two-step procedure illustrated in Figure 5. First, a single-factor genomic SEM model with a single SNP effect on the latent factor is fitted – this is the same model used when conducting a multivariate GWAS. In the second step, direct causal pathways from the SNP to each of the traits are added. The effect sizes for those pathways are then estimated while holding the parameters from step 1 fixed at their fitted values. The model χ^2 index of this step 2 model is called the Q_{SNP} statistic.

5 Limitations and extensions

The greatest strength of genomic SEM lies in its considerable flexibility. Any structural equation model that relates GWAS traits to latent genetic factors and SNP effects can be fitted using genomic SEM (provided the model is identifiable). This opens the door to a wide variety of analyses, including model designs that are more nuanced than the examples discussed in this introduction. And since it only requires GWAS summary statistics and a reference panel, genomic SEM is easily applicable to thousands of traits. Moreover, the `GenomicSEM` R package makes it easy for users to specify any model, fit it, and obtain fit and heterogeneity statistics. Because of this genomic SEM has quickly gained popularity in the fields of behavioral and psychiatric genetics, where it has been used to analyze the shared genetic basis of traits such as addiction and lack of self-regulation [5] or cannabis use and schizophrenia [6].

But the broad flexibility afforded by genomic SEM is also a potential weakness. Few rules exist for choosing proper model structures, and the framework will not stop one from fitting and analyzing a poor model. Grotzinger et al. recommend considering all of the model fit indices (model χ^2, CFI, AIC, and SRMR) to choose good models, but they offer little guidance on the strengths and weaknesses of each index within the genomic SEM framework. For example, it is unclear how model selection decisions should be made when the indices imply conflicting conclusions. Even if there was a universally-accepted way to test for goodness of fit and compare the relative fits of several models, we would still be unable to conclude that a model with “acceptable” or “good” fit truly captures the causal pathways present in reality. Researchers who rely on SEM should always keep in mind that SEM is not a causal inference method: it does not provide a way to control for unobserved confounders. The SEM literature is replete with terms such as “causal pathways” and “causal factors,” but SEM (and genomic SEM in particular) operate on the assumption that the user-specified causal structure is correct. Any interpretation of findings needs to state that the factor loadings and other coefficient estimates are valid given the specified structural equation model; the estimates may or may not hold in general. Of course, these challenges are not unique to genomic SEM. We merely seek to highlight that structural equation models in genetics are prone to the same uncertainties and limitations inherent to SEM more broadly.

When considering potential future improvements to genomic SEM, ancestral heterogeneity immediately comes to mind. A major focus of genetics research in recent years has been on the statistical challenges arising from differences in allele frequencies, heritabilities, and genetic correlation structures among individuals of different ethnic ancestries. To avoid potential confounding from such differences, genetic analyses have traditionally been performed only on ethnically homogeneous populations. The vast majority of GWAS only include individuals of European descent, so as a result there is a noticeable gap in knowledge about the genetic makeup and associations of non-European populations. Thus, multi-ethnic methods for GWAS, polygenic risk score construction, and fine mapping have the potential to accelerate discoveries of particular significance to understudied populations. Genomic SEM, however, is only valid for GWAS performed on individuals of the same ethnic ancestry. A useful extension of the genomic SEM framework would be to allow for empirical and sampling

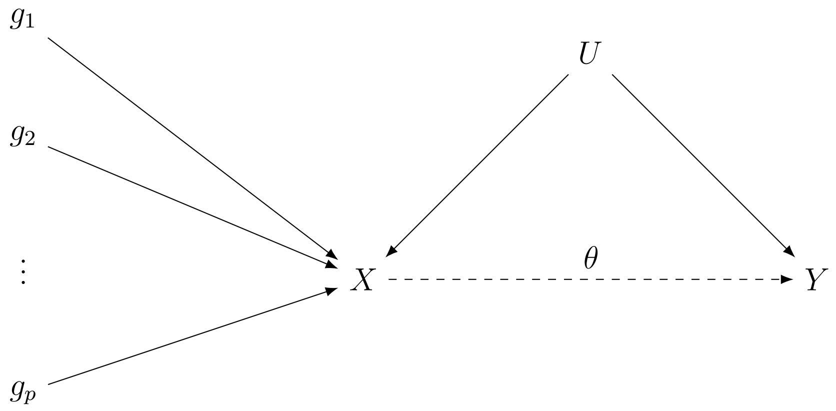


Figure 6: Path diagram of the TWAS model. $g_1, \ldots, g_k$ are eQTLs, X is the expression level of a given gene, U is an unobserved confounder, and Y is the phenotype of interest.

covariance matrix estimation from a multi-ethnic set of GWAS, perhaps through covariate-adjusted linkage disequilibrium score regression (cov-LDSC) [7] or a related method. This would open the possibility of using genomic SEM on a wider set of GWAS, and potentially lead to novel insights for recently admixed populations.

Another extension of genomic SEM that bears mentioning is transcriptome-wide structural equation modeling (T-SEM). This method is described in a recent preprint and has already been incorporated into the `GenomicSEM` package [8]. Transcriptome-wide association studies (TWAS) estimate the association between a gene's genetically-regulated expression levels and a phenotype. The basic TWAS setup is shown in Figure 6. First the gene's expression level (X) is predicted from expression quantitative trait loci (eQTLs), and then the phenotype (Y) is predicted from $\hat{X}$. Under certain assumptions, TWAS are able to control for bias from unobserved confounding variables and hence identify (putatively) causal genes. Just like GWAS, however, a TWAS is limited to a single phenotype. Now, T-SEM is an extension of TWAS that allows one to consider multiple genetically correlated traits. The assumptions, covariance matrices, and estimation methods of T-SEM are analogous to those of (standard) genomic SEM except that effect sizes and correlations of TWAS-identified genes are used instead of effect sizes and correlations of GWAS-identified SNPs. As a result, T-SEM facilitates analyses of structural equation models at a higher level of genomic organization than in (standard) genomic SEM. Assuming that the TWAS assumptions hold, T-SEM might also benefit from the confounding correction offered by TWAS.

6 The verdict

Genomic SEM was proposed in 2019, and since then structural equation models have been increasingly gaining popularity in genetics research. The flexibility of the genomic SEM framework, despite its shortfalls, has launched a new direction for the analysis of genetically correlated traits. By modeling underlying genetic factors as latent variables and explicitly accounting for genetic covariance between measured phenotypes, this approach has led to multiple advances in our understanding of psychopathology and cognitive traits. The use of structural equation models in genetics is still in its infancy and some uncertainties remain, but this frame of thinking is bound to yield many more extensions and applications across the spectrum of genetics research.

References

1. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3,** 513–525. `http://dx.doi.org/10.1038/s41562-019-0566-x` (May 2019).
2. Neale Lab. *Round 1 GWAS of the UK Biobank* (2017). `http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank`.
3. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47,** 291–295. `http://dx.doi.org/10.1038/ng.3211` (Feb. 2015).
4. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics* **52,** 859–864. `http://dx.doi.org/10.1038/s41588-020-0653-y` (June 2020).
5. Linnér, R. K. *et al.* Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nature Neuroscience* **24,** 1367–1376. `http://dx.doi.org/10.1038/s41593-021-00908-3` (Aug. 2021).
6. Johnson, E. C. *et al.* The relationship between cannabis and schizophrenia: a genetically informed perspective. *Addiction* **116,** 3227–3234. `http://dx.doi.org/10.1111/add.15534` (May 2021).
7. Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Human Molecular Genetics* **30,** 1521–1534. `https://academic.oup.com/hmg/article/30/16/1521/6275363` (Aug. 2021).
8. Grotzinger, A. D., de la Fuente, J., Davies, G., Nivard, M. G. & Tucker-Drob, E. M. Transcriptome-wide and Stratified Genomic Structural Equation Modeling Identify Neurobiological Pathways Underlying General and Specific Cognitive Functions. medRxiv preprint (May 2021).