An introduction to structural equation modeling in genetics

Mykhaylo M. Malakhov

Division of Biostatistics, School of Public Health, University of Minnesota

SCHOOL OF PUBLIC HEALTH
University of Minnesota

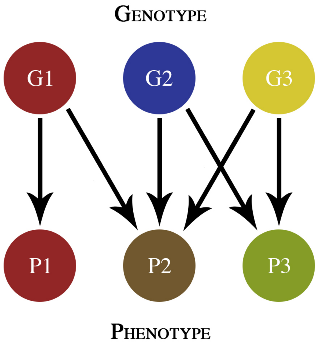
NIH National Institutes of Health
Turning Discovery Into Health

GWAS: from univariate to multivariate

Genome-wide association studies (GWAS) have successfully identified thousands of genetic loci associated with human traits.

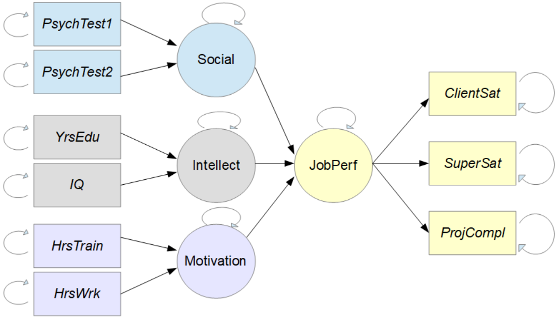
Some problems:

- Widespread genetic pleiotropy
- Research suggests that constellations of phenotypes are affected by shared sources of genetic liability



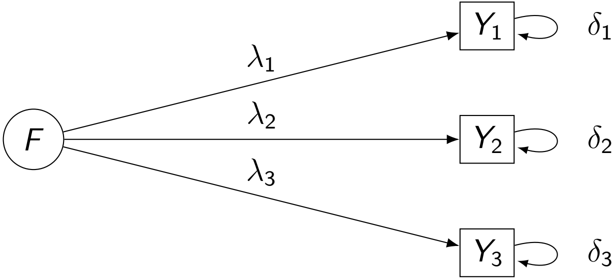
Review of structural equation modeling

What is SEM?



Structural equation modeling (SEM) is a diverse set of methods for posing, fitting, and comparing causal latent variable models.

Example: single-factor SEM



$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

Assumptions:

$cov(F, \delta_j) = 0$

$cov(\delta_i, \delta_j) = 0$ for $i \neq j$

$cov(Y_i, Y_j \mid F) = 0$

Factor analysis: EFA and CFA

Exploratory factor analysis (EFA):

- Goal is to uncover the underlying causal structure
- Any measured variable may be associated with any factor

Confirmatory factor analysis (CFA):

- Goal is to test model fit and make inferences about effects
- The model structure and constraints are pre-specified

In both cases, the model is fit by minimizing the “distance” between the empirical covariance matrix and the model-implied covariance matrix.

Structural equation models of genetic architecture

SEM models of genetic covariance

Genomic SEM

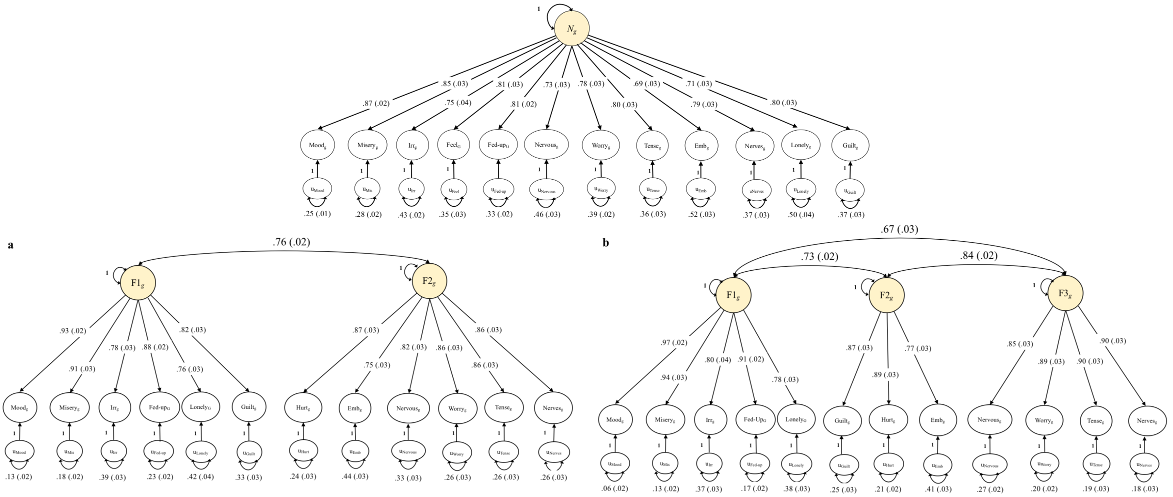
A framework for modeling the genetic architecture of constellations of traits.

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

Use cases:

- Conduct CFA for multivariate genetic associations among traits
- Perform GWAS for constellations of traits
- Compute polygenic risk scores for constellations of traits
- Identify loci that cause divergence between traits

Application: CFA of neuroticism



Application: CFA of neuroticism

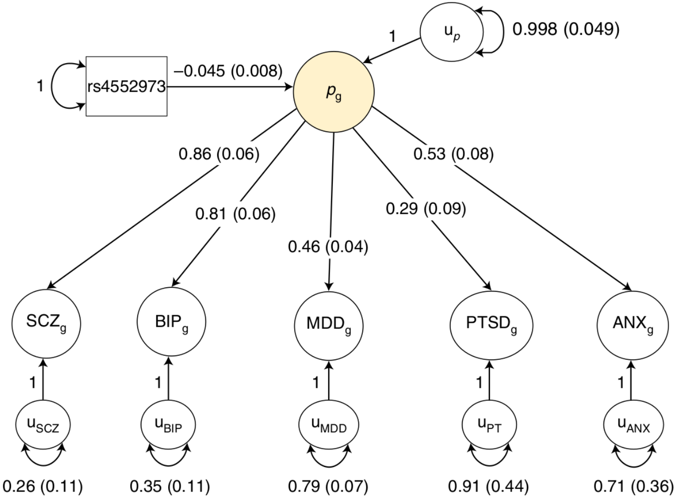
The model χ^2 index:

- Likelihood-ratio statistic comparing the fit of the proposed SEM against the fit of a saturated model
- $\chi^2(54) = 4,884.10$ for single-factor; $\chi^2(53) = 2,758.18$ for two-factor; $\chi^2(51) = 1,879.31$ for three-factor

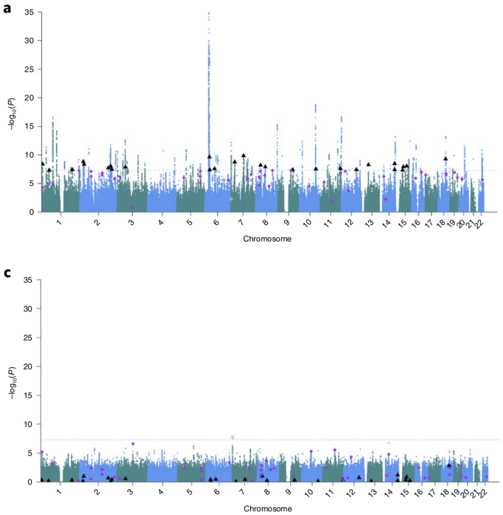
The Q_{SNP} statistic:

- Quantifies how much a SNP's (univariate) effects can be explained by the specified causal pathway model
- Analogous to the Q statistic of heterogeneity in meta-analysis
- 69 significant SNPS in single-factor; 28 significant SNPs in two-factor; 20 significant SNPs in three-factor

Application: multivariate GWAS of p-factor



Application: multivariate GWAS of *p*-factor



- 128 independent loci were genome-wide significant for the *p*-factor
- 27 of these were not identified by any of the contributing univariate GWAS
- Multivariate GWAS had higher power than univariate GWAS (i.e. greater mean χ^2 statistics)

How does genomic structural equation modeling work?

Genomic SEM: stage 1

The empirical genetic covariance matrix (without SNP effects) is

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

- h_i^2 is the heritability of phenotype i
- $\sigma_{gi,gj} = r_{gi,gj}\sqrt{h_i^2 h_j^2}$ is the genetic covariance between phenotypes i and j

Genomic SEM: stage 1

The sampling covariance matrix (without SNP effects) is

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

Genomic SEM: stage 2

The SEM can be specified as the measurement model

$$y = \Lambda\eta + \varepsilon$$

and the structural model

$$\eta = B\eta + \zeta.$$

Then the model-implied covariance matrix is

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

where

- Ψ is the latent variable covariance matrix
- Θ is a matrix of covariances among the residuals.

Genomic SEM: stage 2

Diagonally weighted least squares minimizes the fit function

$$F_{WLS}(\theta) = (s - \sigma(\theta))^\intercal D_s^{-1}(s - \sigma(\theta))$$

where

- s and $\sigma(\theta)$ are half-vectorized versions of S and $\Sigma(\theta)$, respectively
- D_s is V_S with its off-diagonal elements set to 0.

The robust covariance matrix of the SEM parameters is estimated by

$$V_\theta = (\hat{\Delta}^\intercal \Gamma^{-1} \hat{\Delta})^{-1} \hat{\Delta}^\intercal \Gamma^{-1} V_S \Gamma^{-1} \hat{\Delta} (\hat{\Delta}^\intercal \Gamma^{-1} \hat{\Delta})^{-1}$$

where

- $\hat{\Delta}$ is the matrix of model derivatives evaluated at the parameter estimates
- Γ is the naive stage 2 weight matrix
- V_S is the sampling covariance matrix of S

Genomic SEM: fit statistics

(A) Model χ^2 reflects the difference $S - \Sigma(\theta)$. It follows $\chi^2(r)$ 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

(B) CFI reflects the extent to which the proposed model fits better than a model where all phenotypes are heritable but genetically uncorrelated. $CFI = \frac{f(\text{independence model}) - f(\text{proposed model})}{f(\text{proposed model})}$, where $f = \chi^2 - r$

(C) AIC balances fit with parsimony. AIC $= \chi^2 + 2f_p$

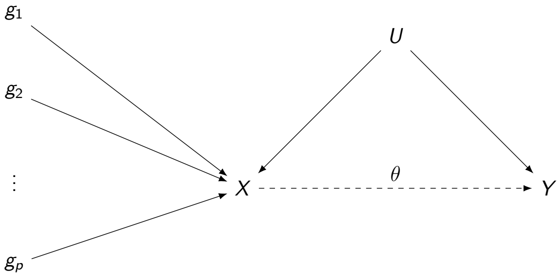
(D) SRMR reflects approximate model fit. It is calculated from $\Sigma(\theta)$ and S

Moving forward: limitations and extensions of genomic SEM

Weaknesses and limitations

1. Genomic SEM is very flexible, but offers little guidance for choosing proper model structures. Poor models can be fit and analyzed too
2. The authors recommend considering all of the (many) model fit indices, but what should we do when the indices imply conflicting conclusions?
3. Genomic SEM does not offer a way to account for ancestral heterogeneity, and it is not applicable for recently admixed populations
   - **Potential solution:** Luo et al. (HMG, 2021) proposed cov-LDSC for estimating h^2 in admixed populations. Can this be extended to genomic SEM?
4. *Despite being based on causal pathway models, genomic SEM does not control for unknown confounders and hence does not yield causal effect estimates*

Transcriptome-wide structural equation modeling (T-SEM)



T-SEM has the same setup, covariance matrices, and estimation methods as (standard) genomic SEM except that TWAS-identified genes replace GWAS-identified SNPs.

References

1. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour 2019;3:513–25.
2. Grotzinger AD, de la Fuente J, Davies G, Nivard MG, and Tucker-Drob EM. Transcriptome-wide and Stratified Genomic Structural Equation Modeling Identify Neurobiological Pathways Underlying General and Specific Cognitive Functions. 2021. medRxiv preprint.
3. Luo Y, Li X, Wang X, et al. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. Human Molecular Genetics 2021;30:1521–34.