A Comprehensive Guide to Simulating Phenotypes with Causal and Mechanistic Pleiotropic Architectures

Abstract

This document provides a definitive and detailed guide to a generalized mathematical framework for simulating two phenotypes, P_1 and P_2, linked by causality, confounding, and complex pleiotropy. The framework is built upon a fully polygenic architecture where all genetic effects are drawn from normal distributions, operating on standardized genotypes. A core innovation is a sophisticated and biologically plausible model of pleiotropy, parameterized by direct proportions of causal SNPs for P_1 that fall into two disjoint classes: those exerting a **correlated pleiotropic effect** and those exerting an **uncorrelated pleiotropic effect**. The simulation algorithm is robust and data-driven, using empirically calculated genetic variances from provided genotype data to scale all other phenotypic components, ensuring precise adherence to user-specified heritability parameters.

Introduction: The Need for Principled Simulation

A robust simulation framework must generate data that mirrors biological reality. This document provides a blueprint for generating two phenotypes, P_1 and P_2, based on a sophisticated polygenic model that correctly handles data-driven variance components and a nuanced, two-component model of pleiotropy. By providing a clear distinction between user-defined intuitive parameters and the derived model coefficients, it enables the creation of complex datasets where the underlying ground truth is precisely known, serving as an invaluable tool for methodological development in statistical genetics.

The Simulation Model

Conceptual Overview

The model defines the relationships between five core entities: two polygenic scores (G_1, G_2), a common confounder (C), and two phenotypes (P_1, P_2). The key relationships are:

- **Polygenic Basis:** P_1 and P_2 are influenced by polygenic scores G_1 and G_2, respectively. Genotypes (X_i) are assumed to be standardized to have a mean of 0 and variance of 1.
- **Causality:** P_1 may have a direct causal effect α on P_2.
- **Mechanistic Pleiotropy:** A proportion of SNPs influencing P_1 also have a direct effect on P_2. These pleiotropic SNPs are partitioned into two distinct, non-overlapping classes based on their mechanism.

The Structural Equations

The relationships are formalized by the following linear equations:

$$P_1 = G_1 + c_1 C + E_1$$
$$P_2 = \alpha P_1 + G_{\text{pleio}} + G_2 + c_2 C + E_2$$

Detailed Component Definitions

- G_1, G_2**:** Polygenic scores, defined as $G = \sum_{i=1}^{m} X_i \beta_i$. The per-allele effect sizes are random variables drawn from normal distributions: $\beta_{1,i} \sim \mathcal{N}(0, \sigma^2_{\beta_1})$ and $\beta_{2,i} \sim \mathcal{N}(0, \sigma^2_{\beta_2})$.
- G_{pleio}**:** The aggregate pleiotropic score. It is the sum of two components from two disjoint sets of SNPs, which are a subset of the causal SNPs for G_1.
    - **Correlated Pleiotropy ($G_{\text{pleio,corr}}$):** A proportion q_{corr} of G_1's causal SNPs fall into this class ($m_{\text{corr}} = q_{\text{corr}} \cdot m_1$). For each such SNP j, its effect on P_2 is directly proportional to its effect on P_1:

      $$\gamma_j = \tau \cdot \beta_{1,j}$$

      Here, τ is a coupling coefficient that models a shared biological pathway. The resulting score is $G_{\text{pleio,corr}} = \sum_{j=1}^{m_{\text{corr}}} X_j \gamma_j = \tau \sum_{j=1}^{m_{\text{corr}}} X_j \beta_{1,j}$.
    - **Uncorrelated Pleiotropy ($G_{\text{pleio,uncorr}}$):** A proportion q_{uncorr} of G_1's causal SNPs fall into this class ($m_{\text{uncorr}} = q_{\text{uncorr}} \cdot m_1$). For each such SNP k, its effect on P_2 is an independent random draw, unrelated to its effect on P_1:

$$\gamma_k \sim \mathcal{N}(0, \sigma_\gamma^2)$$

This models an independent biological pathway. The resulting score is $G_{\text{pleio,uncorr}} = \sum_{k=1}^{m_{\text{uncorr}}} X_k \gamma_k$.

The total pleiotropic score is $G_{\text{pleio}} = G_{\text{pleio,corr}} + G_{\text{pleio,uncorr}}$.

The Parameter Calculation Algorithm

User-Specified Parameters

- **Genotype Data:** A matrix of standardized genotypes for all individuals at all relevant SNPs.
- **SNP Sets:** Lists of SNPs designated as causal for G_1 (size m_1) and G_2 (size m_2).
- **Causal Effect (α):** The coefficient for the causal path from P_1 to P_2. Set to 0 for no causal effect.
- **Pleiotropy Proportions (q_{corr}, q_{uncorr}):** Direct proportions of G_1's m_1 causal SNPs.
    - q_{corr}: The proportion of G_1's SNPs that exhibit correlated pleiotropy.
    - q_{uncorr}: The proportion of G_1's SNPs that exhibit uncorrelated pleiotropy.
    - *Constraint: It must hold that* $q_{\text{corr}} + q_{\text{uncorr}} \leq 1$.
- **Variance Proportions for P_1 ($h_1^2, v_{c,1}^2$):**
    - h_1^2: Heritability of P_1 (variance explained by G_1).
    - $v_{c,1}^2$: Proportion of variance in P_1 explained by the confounder C.
- **Variance Proportions for P_2 ($h_2^2, v_{c,2}^2, h_{\text{pleio-corr}}^2, h_{\text{pleio-uncorr}}^2$):**
    - h_2^2: Total heritability of P_2.
    - $v_{c,2}^2$: Total proportion of variance in P_2 explained by the confounder C.
    - $h_{\text{pleio-corr}}^2$: Proportion of variance in P_2 from the correlated pleiotropic path.
    - $h_{\text{pleio-uncorr}}^2$: Proportion of variance in P_2 from the uncorrelated pleiotropic path.

- **Causal Path Variance ($v^2_{P_1 \to P_2}$):** (Required only if $\alpha \neq 0$). The proportion of variance in P_2 explained by the entire causal path from P_1.

Step 1: Calculate Parameters for Phenotype P_1

1. **Determine Effect Size Variance for β_1:** To achieve a target heritability h_1^2 with m_1 SNPs, we set the variance of the effect sizes to $\sigma^2_{\beta_1} = h_1^2/m_1$.
2. **Draw Effects and Calculate Empirical Genetic Variance:**
   - Draw m_1 effect sizes: $\beta_{1,i} \sim \mathcal{N}(0, \sigma^2_{\beta_1})$.
   - Calculate the full polygenic score: $G_1 = \sum_{i=1}^{m_1} X_i \, \beta_{1,i}$.
   - Compute its sample variance: $\widehat{\text{Var}}(G_1)$. This value becomes the data-driven anchor for P_1.
3. **Determine Total Phenotypic Variance of P_1:**

$$\text{Var}(P_1) = \frac{\widehat{\text{Var}}(G_1)}{h_1^2}$$

4. **Calculate Confounder and Residual Variances for P_1:**
   - $c_1 = \sqrt{v^2_{c,1} \cdot \text{Var}(P_1)}$.
   - $\text{Var}(E_1) = \text{Var}(P_1) \cdot (1 - h_1^2 - v^2_{c,1})$.

Step 2: Calculate Parameters for Phenotype P_2

This step's logic branches based on whether a causal path exists, as this determines how the scale of P_2 is established.

Case 1: Causal Path Exists ($\alpha \neq 0$)

5. **Determine Total Variance of P_2:** The scale is anchored by the causal effect from P_1.

$$\text{Var}(P_2) = \frac{\alpha^2 \text{Var}(P_1)}{v^2_{P_1 \to P_2}}$$

6. **Calculate Pleiotropic Parameters (τ, σ^2_γ):**
   - First, define and compute the variance of the partial genetic score for correlated pleiotropy, using the β_1 effects drawn in Step 1: $G_{1,p,\text{corr}} = \sum_{j \in \text{corr set}} X_j \, \beta_{1,j}$. Compute its sample variance $\widehat{\text{Var}}(G_{1,p,\text{corr}})$.

- Solve for τ: $\tau = \sqrt{\frac{h^2_{\text{pleio-corr}} \cdot \text{Var}(P_2)}{\widehat{\text{Var}}(G_{1,p,\text{corr}})}}$.
- Solve for σ^2_γ: $\sigma^2_\gamma = \frac{h^2_{\text{pleio-uncorr}} \cdot \text{Var}(P_2)}{m_{\text{uncorr}}}$.

7. **Calculate Effect Size Variance for G_2 ($\sigma^2_{\beta_2}$):** The needed variance for G_2 is what remains of the total genetic variance.

   Computational Note: Calculating $\widehat{\text{Var}}(\alpha G_1 + G_{\text{pleio}})$
   This term must be calculated directly to correctly account for the covariance between G_1 and G_{pleio}.

   a. **Construct G_{pleio}:**
      - Calculate the correlated part: $G_{\text{pleio,corr}} = \tau \cdot G_{1,p,\text{corr}}$.
      - Draw effects $\gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma)$ for the m_{uncorr} SNPs.
      - Calculate the uncorrelated part: $G_{\text{pleio,uncorr}} = \sum_{k \in \text{uncorr set}} X_k \gamma_k$.
      - Sum them: $G_{\text{pleio}} = G_{\text{pleio,corr}} + G_{\text{pleio,uncorr}}$.

   b. **Construct the composite score:** For each individual, create the score $Y = \alpha G_1 + G_{\text{pleio}}$.

   c. **Compute the variance:** Calculate the sample variance of the resulting vector Y. This is $\widehat{\text{Var}}(\alpha G_1 + G_{\text{pleio}})$.

   Now, solve for the variance needed from G_2:

$$\widehat{\text{Var}}(G_2)_{\text{needed}} = h^2_2 \text{Var}(P_2) - \widehat{\text{Var}}(\alpha G_1 + G_{\text{pleio}})$$

$$\sigma^2_{\beta_2} = \frac{\widehat{\text{Var}}(G_2)_{\text{needed}}}{m_2}$$

Case 2: No Causal Path ($\alpha = 0$)

8. **Determine Effect Size Variance for G_2 ($\sigma^2_{\beta_2}$):** With no causal anchor, G_2 must create the scale. The proportion of variance from G_2 is $h^2_{G_2 \to P_2} = h^2_2 - (h^2_{\text{pleio-corr}} + h^2_{\text{pleio-uncorr}})$.

$$\sigma^2_{\beta_2} = \frac{h^2_{G_2 \to P_2}}{m_2}$$

9. **Determine Total Variance of P_2:** Draw the β_2 effects, calculate $G_2 = \sum X_i \beta_{2,i}$, find its empirical variance $\widehat{\text{Var}}(G_2)$, and then scale up to find the total variance.

$$\text{Var}(P_2) = \frac{\widehat{\text{Var}}(G_2)}{h^2_{G_2 \to P_2}}$$

10. **Calculate Pleiotropic Parameters (τ, σ^2_γ):** Now that $\text{Var}(P_2)$ is known, use the same formulas as in Case 1 to solve for τ and σ^2_γ.

Step 3: Calculate Remaining Non-Genetic Parameters for P_2

This step is common to both cases, as the scale of P_2 is now fully determined.

11. **Confounder Coefficient c_2:**

$$c_2 = \sqrt{v^2_{c,2} \cdot \text{Var}(P_2)} - \alpha c_1$$

12. **Residual Variance $\text{Var}(E_2)$:**

$$\text{Var}(E_2) = \text{Var}(P_2) \cdot (1 - h^2_2 - v^2_{c,2}) - \alpha^2 \text{Var}(E_1)$$

Simulation Procedure

13. **Parameter Calculation:** Execute Steps 1-3 to solve for all model coefficients.

14. **One-Time Effect Size Draw:** Realize all random genetic effects.
    - The effects $\beta_{1,i}$ were already drawn in Step 1.
    - Draw unique effects for G_2: $\beta_{2,i} \sim \mathcal{N}(0, \sigma^2_{\beta_2})$.
    - For the m_{uncorr} SNPs, draw their unique effects: $\gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma)$.

15. **Per-Individual Simulation Loop:** For each individual in the genotype data:
    a. Calculate all genetic scores:
       - $G_1 = \sum X_i \beta_{1,i}$.
       - $G_2 = \sum X_i \beta_{2,i}$.
       - $G_{\text{pleio,corr}} = \tau \sum_{j \in \text{corr set}} X_j \, \beta_{1,j}$.
       - $G_{\text{pleio,uncorr}} = \sum_{k \in \text{uncorr set}} X_k \, \gamma_k$.
    b. Draw non-genetic components: $c \sim \mathcal{N}(0,1)$, $e_1 \sim \mathcal{N}(0, \text{Var}(E_1))$, $e_2 \sim \mathcal{N}(0, \text{Var}(E_2))$.
    c. Construct the final phenotypes:

- $P_1 = G_1 + c_1 c + e_1$.
- $P_2 = \alpha P_1 + (G_{\text{pleio,corr}} + G_{\text{pleio,uncorr}}) + G_2 + c_2 c + e_2$.