Trait simulation

Huwenbo Shi

August 8, 2016

1 Types of simulations

1.1 GWAS simulation

1.1.1 Single quantitative trait simulation

- Users specify
  1. Genotype matrix
  2. Effect size vector for SNPs
  3. Desired heritability
  4. Desired missing rate
- Model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $\boldsymbol{y}$: trait values for n individuals
- $\boldsymbol{X}$: the $n \times p$ genotype matrix, provided by the user (allelic? standardized?)
- $\boldsymbol{\beta}$: the p-vector of effect sizes, provided by the user
- $\boldsymbol{\epsilon}$: the n-vector of environmental effect, drawn by the program such that the desired amount of heritability is achieved.

1.1.2 Single case-control trait simulation

- Users specify
  1. Genotype matrix
  2. Effect size vector for SNPs
  3. Desired heritability
  4. Prevalence of the cases
  5. Desired number of cases and controls
- Simulation procedure
  1. First, simulate liability based on $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
  2. Then, select liability threshold based on prevalence
  3. Finall, draw cases and controls according to the specified numbers

1.2 Two correlated traits

- Users specify
  1. Genotype matrix
  2. SNP effect size vectors for 2 traits

3. Desired heritability for 2 traits
4. Desired missing rate for 2 traits (or prevalence and number of cases and controls for case-control traits)
5. Desired phenotypic correlations between the two traits

- Model: $\boldsymbol{y}_1 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{y}_2 = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\delta}$
- Simulation procedure
  1. First, simulate the genetic component $\boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{X}\boldsymbol{\gamma}$
  2. Compute $\mathrm{Var}[\boldsymbol{X}\boldsymbol{\beta}]$, $\mathrm{Var}[\boldsymbol{X}\boldsymbol{\gamma}]$, and $\mathrm{Cov}[\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\gamma}]$
  3. Find σ_ϵ^2, σ_δ^2, and $\sigma_{\epsilon\delta}$ and draw $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, from bivariate normal (parameterized by σ_ϵ^2, σ_δ^2, and $\sigma_{\epsilon\delta}$), such that the desired amount of heritability and phenotypic correlation are achieved.
  4. Dichotomize liability for case-control traits

1.3 Variance component simulation

1.3.1 Two variance components (σ_g^2, σ_e^2)

- Users specify
  1. Heritability of the trait (σ_g^2)
  2. Genotype matrix
  3. Missing rate
  4. Mean effect
- Model $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\epsilon}$
- $\boldsymbol{y}$: n-vector of trait values
- $\boldsymbol{X}$: Standardized $n \times p$ genotype matrix
- $\boldsymbol{u}$: p-vector of SNP effect ($\boldsymbol{u} \sim N(0, \sigma_g^2/p)$)
- $\boldsymbol{\mu}$: mean effect (can be a function of genotypes, i.e. $\boldsymbol{\mu}_i = \tilde{\boldsymbol{\mu}}_i + \boldsymbol{X}_{ik}$?)
- Simulate from $N\left(\boldsymbol{\mu}, \sigma_g^2 \boldsymbol{X}\boldsymbol{X}^\intercal/n + \sigma_e^2 \boldsymbol{I}\right)$

1.3.2 Mutiple variance components (σ_g^2, σ_i^2, σ_e^2)

- Users specify
  1. σ_g^2, and all σ_i^2
  2. Genotype matrix
  3. All other design matrices for each variance component
  4. Missing rate
  5. Mean effect
- Model $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{u} + \sum_i \boldsymbol{W}_i \boldsymbol{v}_i + \boldsymbol{\epsilon}$
- $\boldsymbol{X}$: standardized $n \times p$ genotype matrix
- $\boldsymbol{W}_i$: design matrices for each variance component

1.4 General multiple correlated traits and multiple variance components simulation

- Users specify

1. Mean effects $\boldsymbol{\mu}_i$ for m traits
2. Genotype matrix $\boldsymbol{X}$
3. Design matrices for each variance component $\boldsymbol{W}_i$
4. Heritability: σ^2_{gj} (heritability of trait j)
5. Other variance components: σ^2_{ij}
6. Genetic covariances: σ_{gjk} for the pair of traits $\boldsymbol{y}_j$ and $\boldsymbol{y}_l$
7. Random noise covariances: σ_{ejk} for the pair of traits $\boldsymbol{y}_j$ and $\boldsymbol{y}_l$
8. Other covariances: σ_{ijk} for the pair of traits $\boldsymbol{y}_j$ and $\boldsymbol{y}_l$

- Simulate from the model

$$N\left(\begin{bmatrix} \boldsymbol{\mu}_i \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \boldsymbol{C}_g \otimes \frac{\boldsymbol{X}\boldsymbol{X}^\intercal}{n} + \sum_i \boldsymbol{C}_i \otimes \frac{\boldsymbol{W}_i\boldsymbol{W}_i^\intercal}{n} + \boldsymbol{C}_e \otimes \boldsymbol{I}\right), \text{ where } \boldsymbol{C}_* = \begin{bmatrix} \sigma^2_{*1} & \cdots & \sigma_{*1m} \\ \vdots & \ddots & \vdots \\ \sigma_{*m1} & \cdots & \sigma^2_{*m} \end{bmatrix}$$