TCA-TWAS Data Simulation with heritability and genetics correlation

Gary Qiurui Ma, Brandon Jew

August 2019

1 TCA model

Let Z_h^i be the gene expression level of individual $i \in 1, ...n$ in cell type $h \in 1, ...k$ at gene j, and let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be a matrix of m cis-snps, and let $\mathbf{C^1} \in \mathbb{R}^{n \times p_1}$ be a matrix of p_1 covariates. We assume:

$$\begin{aligned} Z_h^i &= (x_i)^T \beta_h + c_i^1 \gamma_h + \epsilon_z \\ \epsilon_z &\sim \mathcal{N}(0, \sigma_z^2) \end{aligned} \tag{1}$$

where x_i is the i-th row of $\mathbf{X}$ (corresponding to the m snps of the i-th individual), β_h is a m-th length vector of corresponding effect size for the m snps in the h-th cell type, which is also the h-th column of $\mathcal{B}$, c_i^1 is the i-th row of $\mathbf{C^1}$ (corresponding to the p_1 covariates of the i-th individual), γ_h is a p_1-th length vector, and ϵ_z an i.i.d. component of variation

We assume the observed bulk level gene expression are convolved signals from k different cell types. We denote $\mathbf{W} \in \mathbb{R}^{n \times k}$ as a matrix of cell-type proportions of k cell types for each of the n individuals, and $\mathbf{C^2} \in \mathbb{R}^{n \times p_2}$ as a matrix of p_2 global covariates potentially affecting the observed bulk level gene expression. TCA model for G_i, the observed bulk level gene expression of the i-th individual in j-th gene, is as follows:

$$\begin{aligned} G_i &= c_i^2 \delta + \sum_{h=1}^{k} w_{hi} z_{hi} + \epsilon_g \\ \epsilon_g &\sim \mathcal{N}(0, \sigma_g^2) \end{aligned} \tag{2}$$

where c_i^2 is the i-th columnn of $\mathbf{C^2}$ (corresponding to the p_2 snps of the i-th individual), δ is a p_2-th length vector of corresponding effect size for the p_2 global covariates in the j-th gene, and ϵ_g is a component of i.i.d. variation that models measurement noise.

2 Data Simulations

2.1 SNPs data generation

For each gene j, the d-th snps cis-snps data for person i is sampled with binomial distribution $x_{id} \sim \text{Binomial}(2, \text{MAF}_d), d \in 1, ...m$. The x_d is centered and scaled to zero mean and unit variance. The p-th cell-type specific covariate for person i is sampled from random normal distribution $c^1_{pi} \sim \mathcal{N}(0, 1)$. The p-th global covariate for bulk level gene expression for persion i is generated similarly: $c^2_{pi} \sim \mathcal{N}(0, 1)$

For d-th cis-snps, sparsity is enforced on the effect size of this snps on k cell-type specific gene expression as follow:

$$\begin{aligned} \beta_{dh} &= Y_1^h \times Y_2 \qquad (3) \\ Y_1 &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\beta) \\ Y_2 &\sim \text{Bernoulli}(1 - pslab) \end{aligned}$$

Where β_{dh} is the d-th entry in β_h (corresponding to the effect size of d-th cis-snps on gene expression for h-th cell type); Y_1^h is the h-th entry in a multivariate normal random variable Y_1, whose covariance matrix is defined by $\mathbf{\Sigma}_\beta \in \mathbf{R}^{k \times k}$ (the effect size of one snps on k cell types are dependent); Y_2 is drawn from a Bernoulli distribution with $pslab$ chance of getting zero.

Covariance of β_d could be expressed as a parametric function of $pslab$ and $\mathbf{\Sigma}_\beta$:

$$\begin{aligned} \text{Var}\,(\beta_{dh}) &= E[(Y_1^h)^2(Y_2)^2] - E[Y_1^h]^2 E[Y_2]^2 \\ &= E[(Y_1^h)^2]E[(Y_2)^2] \\ &= \mathbf{\Sigma}_\beta^{hh} \times (1 - pslab) \qquad (4) \\ \text{Cov}\,(\beta_{dh_1}, \beta_{dh_2}) &= E[Y_d^{h_1} Y_d^{h_2} Y_{2h_1} Y_{2h_2}] - E[Y_d^{h_1}]E[Y_d^{h_2}]E[Y_2]^2 \\ &= \mathbf{\Sigma}_\beta^{h_1 h_2} \times (1 - pslab)^2 \quad (h_1 \neq h_2) \qquad (5) \end{aligned}$$

For p-th covariate, the effect size on gene expression of cell-type k is sampled from a normal distribution $\gamma_{ph} \sim \mathcal{N}(0, \sigma_\gamma^2)$. Similar to γ, δ is generated in the same way.

Cell type weight for i-th person, which is a length-k vector, is sampled from Dirichlet distribution: $w_i \sim \text{Dirichlet}(k, \alpha), \alpha \in \mathbf{R}^k$

2.2 Cell Type Specific Heritability

Denoting h-th cell-type gene expression level as a R.V. z_h:

$$z_h = \sum_{d=1}^{m} x_d \beta_h^d + \sum_{p=1}^{p_1} c_p^1 \gamma_h^p + \epsilon_z \qquad (6)$$

where d-th snps being x_d, effect size of d-th snps for h-th cell type being β_h^d, p-th covariates being c_p^1, effect size of p-th covariate for h-th cell type being γ_h^p, noise being ϵ_z. Then heritability for h-th cell type is defined as:

$$\begin{aligned} h_{snps}^2 &= \frac{\mathrm{Var}\,(\sum_{d=1}^m x_d \beta_h^d)}{\mathrm{Var}\,(z_h)} \qquad (7) \\ &= \frac{\mathrm{Var}\,(\sum_{d=1}^m x_d \beta_h^d)}{\mathrm{Var}\,(\sum_{d=1}^m x_d \beta_h^d) + \mathrm{Var}\,(\sum_{p=1}^{p_1} c_p^1 \gamma_h^p) + \mathrm{Var}\,(\epsilon_z)} \end{aligned}$$

The second line follows as snps, covariates and noise are assumed to be independent. Examining the nominator gives us:

$$\begin{aligned} \mathrm{Var}\,(\sum_{d=1}^m x_d \beta_h^d) &= E[(\sum_{d=1}^m x_d \beta_h^d)^2] - E[\sum_{d=1}^m x_d \beta_h^d]^2 \\ &= E[\sum_{d=1}^m x_d^2 (\beta_h^d)^2 + 2 \sum_{d_1 \neq d_2} x_{d_1} x_{d_2} \beta_h^{d_1} \beta_h^{d_2}] \\ &= \sum_{d=1}^m E[x_d^2 (\beta_h^d)^2] + 2 \sum_{d_1 \neq d_2} E[x_{d_1} x_{d_2} \beta_h^{d_1} \beta_h^{d_2}] \\ &= \sum_{d=1} m(E[x_d^2] - 0)(E[(\beta_h^d)^2 - 0]) \\ &= m\,\mathrm{Var}\,(\beta_h) \qquad (8) \end{aligned}$$

The second line stems from the fact that $E[x_d \beta_d^h] = 0$, the fifth line is a result of X being centered and scaled to unit variance. Similarly, denominator in (7) is calculated as:

$$\mathrm{Var}\,(\sum_{p=1}^{p_1} c_p^1 \gamma_h^p) = p_1\,\mathrm{Var}\,(\gamma_h) \qquad (9)$$

Bringing (8) and (9) into (7) brings:

$$\begin{aligned} h_{snps}^2 &= \frac{m\,\mathrm{Var}\,(\beta_h)}{m\,\mathrm{Var}\,(\beta_h) + p_1\,\mathrm{Var}\,(\gamma_h) + \mathrm{Var}\,\epsilon_z} \\ \mathrm{Var}\,(\beta_h) &= \frac{h_{snps}^2 (p_1\,\mathrm{Var}\,(\gamma_h) + \mathrm{Var}\,\epsilon_z)}{m(1 - h_{snps}^2)} \qquad (10) \end{aligned}$$

Bring (10) into (4) essentialy illustrates that we therefore set

$$\mathbf{\Sigma}_\beta^{hh} = \frac{h_{snps}^2 (p_1 \sigma_\gamma^2 + \sigma_z^2)}{(1 - pslab)(1 - h_{snps}^2)m} \qquad (11)$$

In summary, to specify cell-type specific heritability, we specify h_{snps}^2, σ_z, $pslab$, $\sigma_{\gamma_{hj}}$ and calculate diagonal entries in Σ_β.

2.3 Genetics Correlation Simulation

According to Rheenen, et.al [Rhe+19],for gene j-th expression $\overrightarrow{Z_{h_1}}$, $\overrightarrow{Z_{h_2}}$ for two cell types, we have

$$z_{h_1} = \sum_{d=1}^{m} x_d \beta_{h_1}^{d} + \sum_{p=1}^{p_1} c_p^1 \gamma_{h_1}^{p} + \epsilon_z$$
$$z_{h_2} = \sum_{d=1}^{m} x_d \beta_{h_2}^{d} + \sum_{p=1}^{p_1} c_p^1 \gamma_{h_2}^{p} + \epsilon_z$$

the genetic correlation ρ of the gene expression in two cell-types is defined as

$$\rho^{h_1 h_2} = \frac{\operatorname{Cov}\left(\sum_{d=1}^{m} x_d \beta_{h_1}^{d}, \sum_{d=1}^{m} x_d \beta_{h_2}^{d}\right)}{\sqrt{\operatorname{Var}\left(\sum_{d=1}^{m} x_d \beta_{h_1}^{d}\right) \operatorname{Var}\left(\sum_{d=1}^{m} x_d \beta_{h_2}^{d}\right)}} \tag{12}$$

With (8) we have

$$\begin{aligned}
\rho^{h_1 h_2} &= \frac{\operatorname{Cov}\left(\sum_{d=1}^{m} x_d \beta_{h_1}^{d}, \sum_{d=1}^{m} x_d \beta_{h_2}^{d}\right)}{m\sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}}} \\
&= \frac{m \operatorname{Cov}\left(\beta_{h_1}, \beta_{h_2}\right)}{m\sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}}} \\
&= \rho_{\beta}^{\beta_{h_1}\beta_{h_2}} \\
\operatorname{Cov}\left(\beta_{h_1}, \beta_{h_2}\right) &= \rho^{h_1 h_2} \sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}} \qquad (13)
\end{aligned}$$

where ρ_β is the correlation of effect-size for a gene. (13) essentially requires that correlation of β being the same with the genetic correlation. With β's correlation matrix specified here and the diagonal entries in β's covariance matrix specified by (10), the $\mathbf{\Sigma}_\beta$ in (5) could be calculated as

$$\begin{aligned}
\mathbf{\Sigma}_{\beta}^{h_1 h_2} \times (1 - pslab)^2 &= \rho^{h_1 h_2} \sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}} \\
\mathbf{\Sigma}_{\beta}^{h_1 h_2} &= \frac{\rho^{h_1 h_2}}{(1 - pslab)^2} \sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}} \quad (h_1 \neq h_2) \qquad (14)
\end{aligned}$$

Combinging (14) with (4) presents

$$\Sigma_{\beta}^{h_1 h_2} = \begin{cases} \frac{\rho^{h_1 h_2}}{(1-pslab)^2} \sqrt{\operatorname{Var}\beta_{h_1} \operatorname{Var}\beta_{h_2}} & h_1 \neq h_2 \\ \frac{1}{(1-pslab)} \operatorname{Var}\left(\beta_{h_1}\right) & h_1 = h_2 \end{cases} \tag{15}$$

Where $\operatorname{Var}(\beta_h)$ given by (10). Recall the definition for correlation

$$-1 \leq \rho = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}X \operatorname{Var}Y}} \leq 1$$

enforces the following constraint:

$$-\sqrt{\Sigma_{\beta}^{h_1h_1}\Sigma_{\beta}^{h_2h_2}} \leq \Sigma_{\beta}^{h_1h_2} \leq \sqrt{\Sigma_{\beta}^{h_1h_1}\Sigma_{\beta}^{h_2h_2}}$$
$$pslab - 1 = \rho^{h_1h_1}\rho^{h_2h_2}(pslab-1) \leq \rho^{h_1h_2} \leq \rho^{h_1h_1}\rho^{h_2h_2}(1-pslab) = 1 - pslab$$

The constraint makes sense as when $pslab$ is large, then all entries become zero and correlation disappear. Further notice that this is just a multivariate extension for the cell type specific heritability in section(2.2)

In summary, in order to specify the genetics correlation between effect sizes among different cell types, a correlation matrix $\rho \in R^{K\times K}$ is to be provided. This ρ together with h^2_{snps} (heritability for each cell type) determines the parameter for the distribution of effect size β. Sampling β accordnig to (3),(15) and (10) shall guarantee that both heritablity and genetic correlation is as desired.

2.4 Bulk Heritability Estimate

Recall equation(1) and equation(2)

$$\begin{aligned} Z_h^i &= (x_i)^T\beta_h + c_i^1\gamma_h + \epsilon_z \\ \epsilon_z &\sim \mathcal{N}(0, \sigma_z^2) \\ G_i &= c_i^2\delta + \sum_{h=1}^{k} w_{hi}z_{hi} + \epsilon_g \\ \epsilon_g &\sim \mathcal{N}(0, \sigma_g^2) \end{aligned}$$

Suppose the genetic effects $(X\beta)$ have some covariance structure $\Sigma_{X\beta} \in \mathbb{R}^{k\times k}$, the covariance effects $(C^1\gamma)$ have some covariance structure $\Sigma_{C^1\gamma} \in \mathbb{R}^{k\times k}$. Then the covariance of Z across cell types $\Sigma_Z = \Sigma_{X\beta} + \Sigma_{C^1\gamma} + \text{diag}(\sigma_z^2)$

$$\begin{aligned} \text{Var}(G) &= \text{Var}(\sum_{h=1}^{k} w_h z_h + \sum_{p=1}^{p_2} c_p^2\delta_p + \epsilon_g) \\ &= \text{Var}(\sum_{h=1}^{k} w_h z_h) + \text{Var}\,(\sum_{p=1}^{p_2} c_p^2\delta_p) + \sigma_{\epsilon_g}^2 \\ &= \left(\sum_{h=1}^{k}\sum_{l=1}^{k}\text{Cov}(w_h z_h, w_l z_l)\right) + p_2\sigma_\delta^2 + \sigma_{\epsilon_g}^2 \\ &= \left(\sum_{h=1}^{k}\sum_{l=1}^{k} E[w_h z_h w_l z_l]\right) + p_2\sigma_\delta^2 + \sigma_{\epsilon_g}^2 \text{ (assuming each } z \text{ is centered)} \\ &= \left(\sum_{h=1}^{k}\sum_{l=1}^{k} E[w_h w_l]E[z_h z_l]\right) + p_2\sigma_\delta^2 + \sigma_{\epsilon_g}^2 \qquad (16) \end{aligned}$$

Where

$$E[z_h z_l] = \Sigma_{z\{h,l\}} = \Sigma_{X\beta} + \Sigma_{C^1\gamma} + \mathrm{diag}(\sigma_g^2)$$

$$E[w_h w_l] = \begin{cases} \frac{\tilde{\alpha_h}(1-\tilde{\alpha_h})}{\alpha_0+1} + \tilde{\alpha_h}^2 & h = l \\ \tilde{\alpha_h}\tilde{\alpha_l}(1 - \frac{1}{\alpha_0+1}) & h \neq l \end{cases}$$

Define each entry $\{h, l\}$ of $\Sigma_\alpha \in \mathcal{R}^{k\times k}$ to be $E[w_h w_l]$. Then

$$\begin{aligned} \mathrm{var}(G) &= \mathrm{sum}(\Sigma_\alpha \odot \Sigma_Z) + p_2\sigma_\gamma^2 + \sigma_{\epsilon_g}^2 \\ &= \mathrm{sum}(\Sigma_\alpha \odot \Sigma_{X\beta}) + \mathrm{sum}(\Sigma_\alpha \odot \Sigma_{C^1\gamma}) \\ &+ \mathrm{sum}(\Sigma_\alpha \odot \mathrm{diag}(\sigma_g^2)) + p_2\sigma_\gamma^2 + \sigma_{\epsilon_g}^2 \end{aligned} \tag{17}$$

where the 'sum' operator is the sum of each entry of the argument matrix. Heritability of bulk expression in this model can then be defined as

$$h^2_{\mathrm{bulk}} = \frac{\mathrm{sum}(\Sigma_\alpha \odot \Sigma_{X\beta})}{\mathrm{var}(G)} \tag{18}$$

Therefore, by varying σ_γ and σ_{ϵ_g}, bulk level heritability could be modified as well.

References

[Rhe+19] Wouter van Rheenen et al. “Genetic correlations of polygenic disease traits: from theory to practice”. In: *Nature Reviews Genetics* (2019). ISSN: 1471-0064. DOI: `10.1038/s41576-019-0137-z`. URL: `https://doi.org/10.1038/s41576-019-0137-z`.