Apply Gene-Trait Similarity Regression with Gene level Interaction to Real Data

1 Method

1.1 The Gene-Trait Similarity Model

Denote Y_i as the trait value, X_i the $K \times 1$ covariant vector including the intercept term, and $1 \times l_m$ vector $G_{m,i}$ records the number of distinct alleles on the mth marker for ith individual

For genetic similarity, define S_{ij} to be the genetic similarity between subjects i and j ($i \neq j$). There are many ways to describe the genetic similarity between individuals. Here, two measurements of S_{ij} based on unphased genotype are introduced. One is Average IBS (identical-by-State) method. It defines the genetic similarity between individual i and j as $S_{ij}^{ave} = \frac{1}{M}\sum_{m=1}^{M} s_{m,ij}^{ave}$, where

$$s_{m,ij}^{ave} = \begin{cases} 1, & G_{m,i} \text{ and } G_{m,j} \text{ share the same allele and both of them are homozygous genotypes} \\ \frac{1}{2}, & G_{m,i} \text{ and } G_{m,j} \text{ share the same alleles and both of them are heterozygote genotypes} \\ \frac{1}{2}, & G_{m,i} \text{ and } G_{m,j} \text{ share only 1 allele and one of them is homozygous genotype} \\ \frac{1}{4}, & G_{m,i} \text{ and } G_{m,j} \text{ share only 1 allele and both of them are heterozygote genotype} \\ 0, & \text{o.w} \end{cases}$$

Another way to describe the genetic similarity is so called Typical IBS, here the similarity matrix $S_{ij}^{typ} = \frac{1}{M}\sum_{m=1}^{M} s_{m,ij}^{typ}$ is calculated by

$$s_{m,ij}^{typ} = \begin{cases} 1, & G_{m,i} \text{ and } G_{m,j} \text{ share the same allele(s)} \\ \frac{1}{2}, & G_{m,i} \text{ and } G_{m,j} \text{ share 1 allele} \\ 0, & \text{o.w} \end{cases}$$

We can see the difference between two measurements is that Typical IBS only consider the similarity between two markers while Average IBS also considers whether the marker is homozygous or not. If the information for the pahsed haplotype is known, then we can tell whether the alleles from individual i is identical to the one from individual j. Since the haplotype is unknown, we can only have a guess. Average IBS takes the average value while Typical IBS treats the two individuals always share the same haplotype. It is also worth to point out that while the marker information is bi-allelic such as SNP data, the result from Typical IBS and Average IBS would be quite similar since $s_{m,ij}^{ave}$ can not take the value $\frac{1}{4}$ for the heterozygote genotypes are always the same.

The trait similarity between individual i and j:$Z_{ij}$is computed by

$$Z_{ij} = (Y_i - \mu_i)(Y_j - \mu_j)$$

where we assume $\mu_i = E(Y_i|X_i, G_i) = X_i\gamma$ is the conditional mean of trait with no genotype effect and γ is the effect of the covariant.

Since we need to consider the Gene interaction, 2 Genes, say, Gene A and Gene B are included. Define $S_{AB,ij} = S_{A,ij} \times S_{B,ij}$, the Gene-Trait similarity model considering the interaction is

$$E(Z_{ij}) = \tau_A S_{A,ij} + \tau_B S_{B,ij} + \phi S_{AB,ij} \tag{1}$$

By the definition of Z_{ij}, it should has zero mean.

1.2 The Joint Test

The joint test tests the hypothesis $H_{0,1}$: $\tau_A = \tau_B = \phi = 0$. It is hard to directly derive the test for the testing based on (1). However, we find the score test by taking advantage of the connection between the similarity model and the variance component model (Tzeng et al. (2009)) which is:

$$Y = X\gamma + G_A + G_B + G_{AB} + e \tag{2}$$

where

$$G_A \sim MN(0, \tau_A S_A)$$
$$G_B \sim MN(0, \tau_B S_B)$$
$$G_{AB} \sim MN(0, \phi S_{AB})$$
$$e \sim MN(0, \sigma I)$$

Simple algebra shows that the score statistic under $H_{0,1}$ is

$$T_1 = \frac{1}{2}\mathbf{Y}' P_1 S_1 P_1 \mathbf{Y}$$

where the $P_1 = \sigma^{-2}(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \sigma^{-2}Q$ and $S_1 = S_A + S_B + S_{AB}$. It is easy to see that when $H_{0,1}$ is not true, T_1 which is as strictly increasing function. Therefore larger values of T_1 provides stronger evidence against $H_{0,1}$. This suggests that the testing procedure should be one sided.

Since Q is projection matrix for $\mathbf{X}$, we can show that

$$T_1 = \frac{1}{2\sigma^4}\mathbf{Y}'QS_1Q\mathbf{Y} = \frac{1}{2\sigma^4}(\mathbf{Y} - \mathbf{X}\gamma)'QS_1Q(\mathbf{Y} - \mathbf{X}\gamma) \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

where λ_i is the ordered nonezero eigenvalues of matrix $QS_1Q/(2\sigma^2)$

If σ_e^2 is unknown, we can replace σ_e^2 by $\hat{\sigma_e^2}$ using

$$\hat{\sigma_e^2} = \frac{\mathbf{Y}'Q\mathbf{Y}}{N-1}$$

1.3 The score test for $H_{0,2}$: $\phi = 0$

If we only want to test the hypothesis for the interaction term, we can also construct a similar score test as the joint test. However, we need to estimate τ_A, τ_B and σ first. Here we use the EM algorithm to get the estimation $\hat{\tau_A}$,$\hat{\tau_B}$ and $\hat{\sigma}$(See Appendix). Based on that, we can use score test T_2 for the interaction testing, where

$$T_2 = \frac{1}{2}\mathbf{Y}' P_2 S_2 P_2 \mathbf{Y}$$

where $P_2 = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, $\mathbf{V} = \hat{\tau_A}S_A + \hat{\tau_B}S_B + \hat{\sigma}I$ and $S_2 = S_{AB}$. And we can also derive that the distribution of T_2 is also a weighted χ^2 distribution, that is:

$$T_2 = \frac{1}{2}\mathbf{Y}' P_2 S_2 P_2 \mathbf{Y} \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

where λ_i is the ordered nonezero eigenvalues of matrix $\mathbf{V}^{-\frac{1}{2}} P_2 S_2 P_2 \mathbf{V}^{-\frac{1}{2}}$.

1.4 One Gene Test adjust with the other Gene

Here we construct the score test for the genetic main effect of a gene conditional on the effect of other gene. This test is performed assuming no gene-gene interactions. This is because in scenarios where the interaction effects do exist, the main effects are typically not well-defined, and its significance depends on the scale of the interacting variables.

$$E(Z_{ij}) = \tau_A S_{A,ij} + \tau_B S_{B,ij}$$

Here to test the main effect of Gene A adjusted with Gene B, the corresponding hypothesis is $H_{0,3}$: $\tau_A = 0$ (vice versa for the Gene B main effect test). Similar to previous two tests, the following score test can be applied:

$$T_3 = \frac{1}{2}\mathbf{Y}' P_3 S_3 P_3 \mathbf{Y}$$

where $P_3 = \mathbf{V}_B^{-1} - \mathbf{V}_B^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}_B^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_B^{-1}$, $\mathbf{V}_B = \hat{\tau_B} S^B + \hat{\sigma} I$ and $P_3 = S_A$, and we can also derive that the distribution of $T_{OneGene}$ is also weighted χ^2 distribution, that is:

$$T_3 = \frac{1}{2}\mathbf{Y}' P_3 S_3 P_3 \mathbf{Y} \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

where λ_i is the ordered nonezero eigenvalues of matrix $\mathbf{V}_B^{-\frac{1}{2}} P_3 S_3 P_3 \mathbf{V}_B^{-\frac{1}{2}}$.

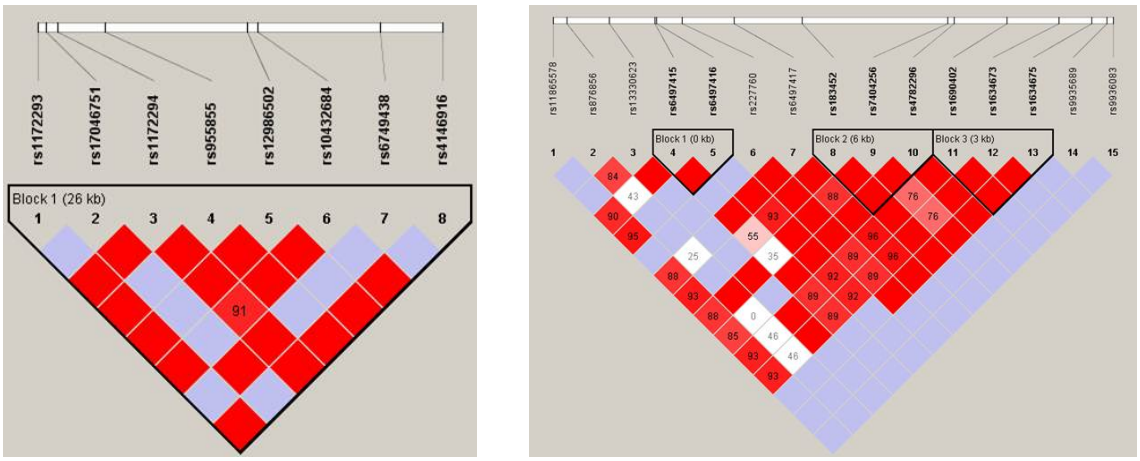
2 Simulation study

2.1 Design for Simulation Study

We studied the performance of our method and other methods using simulated data. Here PCA (Principal Component Analysis) PLS (Partial Least-Square) and simple linear regression are used as bench-marker to jointly test the main effect and the epistasis effect. Only simple Linear regression is applied to compare with our method in epistasis test and one gene test adjust with the other gene. We assume that the true risk model involves causal SNPs located in two genes, Gene A and Gene B. To simulate a realistic LD pattern among the markers, we used real genotype data on Gene RBJ and Gene GPRC5B, two candidate for BMI. Figure 1 shows the LD pattern of the two genes.

Table 1 is a brief summary of the two genes. For gene RBJ, it has 8 SNPs, the left table shows the SNPs feature in RBJ. The right table shows the information for SNP in Gene GPRC5B which has 15 SNPs. For each SNP, two measurements are listed. One is the LD pattern, which shows the average correlation between one SNP and other SNPs in the same gene. The other measurement is minor allele freq (MAF). The reason that we focus on LD and MAF is that we think these two would be the main factors affecting the final result. To convince our result analysis, I transform the original quantitative LD and MAF data into categorical data: SNP is defined to have high LD (H) with other SNPs if its average LD > 0.25, if its LD is between 0.18 and 0.25, it is defined as a normal LD. A SNP labeled with C mean it is a common SNP whose MAF > 0.3, R means its MAF ≤ 0.1, and if its MAF is between 0.1 and 0.3, it is labeled as Median (M)

The trait of each individual in the sample is simulated based on the model listed below according to different purpose. SNP_1^A, SNP_2^A,SNP_1^B,SNP_2^B are indicator variables for causal SNPs and e is



(a) Gene RBJ (b) Gene GPRC5B

Figure 1: LD patten of the two genes

Table 1: LD and MAF information for Gene RBJ and GPRC5B

(a) Gene RBJ

SNP	LD	MAF
1	0.588 (H)	0.115 (M)
2	0.159 (L)	0.062 (R)
3	0.126 (L)	0.482 (C)
4	0.588 (H)	0.115 (M)
5	0.588 (H)	0.115 (M)
6	0.565 (H)	0.119 (M)
7	0.159 (L)	0.062 (R)
8	0.588 (H)	0.115 (M)

(b) Gene GPRC5B

SNP	LD	MAF
1	0.186 (N)	0.146 (M)
2	0.05 (L)	0.066 (R)
3	0.065 (L)	0.044 (R)
4	0.194 (N)	0.195 (M)
5	0.197 (N)	0.143 (M)
6	0.136 (L)	0.159 (M)
7	0.259 (H)	0.46 (C)
8	0.262 (H)	0.336 (C)
9	0.23 (N)	0.482 (C)
10	0.206 (N)	0.371 (C)
11	0.285 (H)	0.394 (C)
12	0.285 (H)	0.394 (C)
13	0.138 (L)	0.155 (M)
14	0 (L)	0.004 (R)
15	0 (L)	0.005 (R)

Table 2: Simulation models

Type I error	
Model 1	$Y = \theta_1 SNP_1^A + \theta_2 SNP_1^B + \theta_{12} SNP_1^A \times SNP_1^B + e$
Power	
Model 2	$Y = \beta \times (SNP_1^A \times SNP_1^B + SNP_2^A \times SNP_2^B) + e$
Model 3	$Y = \beta \times (SNP_1^A + SNP_1^B) + e$
Model 4	$Y = \beta \times (SNP_1^A \times SNP_2^A + SNP_1^B \times SNP_2^B) + e$

a normal distributed variable with mean 0 and variance 1. To generate the simulated data, we first download the genotype of RBJ and GPRC5B from HapMap(http://hapmap.ncbi.nlm.nih.gov/) as a gene bank. For the *ith* individual, we assume no LD between the two genes so we pick RBJ and GPRC5B genotypes from two random sampled persons in the gene bank and combine them together as the *ith* individual's genotype . We then picked one or two SNPs to generate the phenotype according to different model.

The Type I error rate is tested mainly based on Model 1. To simplify the computation, in Type I error rate evaluation, SNP_1^A is the first SNP in Gene A and SNP_1^B is the first SNP in Gene B. For joint test, we set $\theta_1 = \theta_2 = \theta_{12} = 0$. For Epistasis test, we consider 3 situations: (1) neither Gene A or Gene B is associated with the trait($\theta_1 = \theta_2 = \theta_{12} = 0$); (2) Gene A has a main effect to the trait ($\theta_1 \neq 0, \theta_2 = \theta_{12} = 0$); (3) both Gene A and Gene B have main effect to the trait, but there is no epitasis effect ($\theta_1 \neq 0, \theta_2 \neq 0, \theta_{12} = 0$). Similar to the epitasis test, for one gene test, we consider two cases: (1) neither Gene A or Gene B is associated with the trait($\theta_1 = \theta_2 = \theta_{12} = 0$); (2) Gene A has a main effect to the trait ($\theta_1 \neq 0, \theta_2 = \theta_{12} = 0$); For each scenario, we run the simulation for 1,000 times with sample 300 to compute the type I error.

To evaluate the power between different methods, we consider three different genetic models(Model 2-4). However, different from Type I error rate evaluation, SNP_1^A, SNP_2^A, SNP_1^B, SNP_2^B are not fixed SNP in corresponding genes. Instead, we try all possible SNP combination sets to test the performance of each methods under different LD and MAF conditions. Model 2 is a typical interaction model which does not contain main effects and the interaction only occurs between Gene A and Gene B. In the pure main effect model 3, there is no epistasis effect. For Model 4, we consider a pure epistasis case where the interaction only occur within genes.We tune β so that the power of different SNP sets are distinguishable.

2.2 Simulation result: Type I error rate

Table 3 shows the result for type I error rate for all the methods at a significance level of $\alpha = 0.05$. All methods seems work well in all different settings.

Table 3: The type I error rate for different test at significance level 0.05

(a) Joint test

Method	$(\theta_1, \theta_2, \theta_{12})$
	(0,0,0)
Typical	0.049
Average	0.051
PCA	0.052
PLS	0.051
LM	0.048

(b) Epistasis test

Method	$(\theta_1, \theta_2, \theta_{12})$		
	(0,0,0)	(2,0,0)	(2,2,0)
Typical			0.046
Average			0.050
LM	0.052	0.053	0.051

(c) One gene test

Method	(θ_1, θ_2)	
	(0,0)	(2,0)
Typical	0.045	0.047
Average	0.046	0.049
LM	0.046	0.047

2.3 Simulation result: Power

3 Real Data Analysis

We have conducted a genetic study with the Warfarin data. In this data set, 2 genes are involved, the first gene contains 7 SNPs and the second gene is a multi-allelic marker which has 6 different genotype. This study involves 301 individuals and records their "ydose"(what is the name of the trait?). Also available are 6 covariant (e.g. gender, age, race) for each individuals in the study.

our method and other proposed approaches are applied to find the association between the trait and the two genes using different tests. Table 4 shows the p value for different analysis approaches to locate the association between trait and candidate genes with the covariant. From this table we can see that all methods show significant association between trait and the two genes. But our method show a stronger evidence of association than the other approaches. This result shows us an example of how to improve the power of finding the association even when the interaction is not statistical significant. The real data analysis also suggests us that our method is more powerful than other proposed methods which is consistent with the result we concluded in our simulation study.

Table 4: The P-values of different approaches in analysis Warfarin data

Methods	Joint test	Epistasis test	One Gene test	
			Gene 1	Gene 2
Typical IBS	5.6×10^{-20}	0.45	1.67×10^{-20}	1.02×10^{-7}
Average IBS	5.32×10^{-21}	0.63	1.44×10^{-22}	9.83×10^{-9}
PCA	9.48×10^{-5}	-	-	-
PLS	9.96×10^{-5}	-	-	-
LM	2.2×10^{-16}	0.844	2.2×10^{-16}	3.23×10^{-8}

4 Appendix

4.1 Score test for joint test and interaction test

For the model:

$$Y = X\gamma + G_A + G_B + G_{AB} + e$$

Define $\theta = \{\sigma, \tau_A, \tau_B, \phi\}$, the REML log-likelihood function $L(\theta)$ for the whole data set is:

$$L(\theta) = -\frac{1}{2}\left[\log|V| + \log|\mathbf{X}'V^{-1}\mathbf{X}| + \mathbf{Y}'P\mathbf{Y}\right]$$

where $V = Var(Y) = \tau_A S_A + \tau_B S_B + \phi S_{AB} + \sigma I$ is the marginal variance of $\mathbf{Y}$ and $P = V^{-1} - V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}$ is the projection matrix for the model.

Simple algebra shows that the score statistic under $H_0 : \tau_A = \tau_B = \phi = 0$ is

$$U_{\sigma_b^2} = \frac{\partial L(\sigma_b^2, \sigma_e^2)}{\partial \sigma_b^2}\bigg|_{\sigma_b^2=0} = \frac{1}{2}[\mathbf{Y}'P_0SP_0\mathbf{Y} - tr(P_0S)]$$

where the $P_0 = \sigma_e^{-2}(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \sigma_e^{-2}Q$. It is easy seen that $E(U_{\sigma_b^2}) = 0$ under H_0^* and when $\sigma_b^2 > 0$, $E(U_{\sigma_b^2}) = \sigma_b^2 \cdot tr(QSQS)/(2\sigma_e^4)$ which is a strictly increasing function. Therefore larger values of $U_{\sigma_b^2}$ provides stronger evidence against H_0^*. This suggests that the testing procedure should be one sided.

If σ_e^2 is known, the second term in our test is a constant. Therefore using the score statistic is equivalent to using the first term of $U_{\sigma_b^2}$ (denoted by T):

$$T = \frac{1}{2}\mathbf{Y}'P_0SP_0\mathbf{Y} = \frac{1}{2\sigma_e^4}\mathbf{Y}'QSQ\mathbf{Y}$$

Since Q is projection matrix for $\mathbf{X}$, we can show that

$$T = \frac{1}{2\sigma_e^4}\mathbf{Y}'QSQ\mathbf{Y} = \frac{1}{2\sigma_e^4}(\mathbf{Y} - \mathbf{X}\mu)'QSQ(\mathbf{Y} - \mathbf{X}\mu) \sim \sum_{i=1}^{c}\lambda_i\chi_1^2$$

where λ_i is the ordered nonezero eigenvalues of teh matrix $QSQ/(2\sigma_e^2)$

If σ_e^2 is unknown, we can replace σ_e^2 by $\hat{\sigma_e^2}$ using

$$\hat{\sigma_e^2} = \frac{\mathbf{Y}'Q\mathbf{Y}}{N-1}$$

4.2 EM algorithm to estimate τ_A, τ_B and σ

The model under the null hypothesis os $\phi = 0$ is

$$Y = X\gamma + G_A + G_B + e$$

Y: $N \times 1$ matrix, records trait value.

X: $N \times p$ matrix, records the covariant.

γ: $p \times 1$, the effect of covariant.

G_A: the gene effect for Gene A, treated as an random effect. $G_A \sim N(0, \tau_A S_A)$, S_A is the similarity matrix which records the genetic similarity between individuals.

G_B: the gene effect for Gene B, also treated as an random effect. G_A and G_B are assumed to be independent.

e: $N \times 1$ matrix, the error term. $e \sim N(0, \sigma I)$.

Define $U = A^TY$ with the restriction that $A^TA = I_{N-p}$ and $AA^T = I - P_X$.

It is easy to find out that : $E(U) = 0$ and $Var(U) = A^TVA$ where $V = Var(Y) = \tau_A S_A + \tau_B S_B + \sigma I$.

$$\begin{aligned} Cov(U, G_A) =& Cov(A^TX\gamma + A^TG_A + A^TG_B + A^Te, G_A) \\ =& Cov(A^TG_A, G_A) \\ =& A^TCov(G_A, G_A) \\ =& \tau_A A^T S_A \end{aligned}$$

In the same way, we have$Cov(U, G_B) = \tau_B A^T S_B$ and $Cov(G_A, G_B) = 0$ since they are independent.

Therefore, the joint distribution of $(U, G_A, G_B)^T$ is

$$\begin{pmatrix} U \\ G_A \\ G_B \end{pmatrix} \sim MN\left(\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} A^TVA & \tau_A A^T S_A & \tau_B A^T S_B \\ \tau_A S_A A & \tau_A S_A & 0 \\ \tau_B S_B A & 0 & \tau_B S_B \end{pmatrix}\right)$$

so we can have the following conditional mean and variance,

1. the conditional mean and variance for U are

$$\begin{aligned} E(U|G_A, G_B) =& A^T(G_A + G_B) \\ Var(U|G_A, G_B) =& \sigma I_{N-p} \end{aligned}$$

2. the conditional mean and variance for G_A, since $Cov(G_A, G_B) = 0$

$$\begin{aligned} E(G_A|G_B, U) =& E(G_A|U) \\ =& \tau_A S_A A(A^TVA)^{-1}A^TY \\ Var(G_A|G_B, U) =& Var(G_A|U) \\ =& \tau_A S_A - \tau_A^2 S_A A(A^TVA)^{-1}A^T S_A \end{aligned}$$

Simple algebra shows that $A(A^TVA)^{-1}A^T = P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$, so that

$$\begin{aligned} E(G_A|G_B, U) =& \tau_A S_A PY = g_A \\ Var(G_A|G_B, U) =& \tau_A S_A - \tau_A^2 S_A P S_A = v_A \end{aligned}$$

3. Similarly, the conditional mean and variance for G_B are

$$\begin{aligned} E(G_B|G_A, U) =& \tau_B S_B PY = g_B \\ Var(G_B|G_A, U) =& \tau_B S_B - \tau_B^2 S_B P S_B = v_B \end{aligned}$$

Define $\theta = \{\sigma, \tau_A, \tau_B\}$, according to the EM algorithm, we need to first compute the $\log L(\theta^{(t)}|U, G_A, G_B)$,

$$\begin{aligned} \log L(\theta^{(t)}|U, G_A, G_B) =& f(U, G_A, G_B|\theta^{(t)}) \\ =& \log f(U|G_A, G_B, \theta^{(t)}) + \log f(G_A|\theta^{(t)}) + \log f(G_B|\theta^{(t)}) \end{aligned}$$

Since S_A and S_B are singular, we have,

$$f(G_A) = \frac{1}{((2\pi)^{rank(S_A)}|\tau_A S_A|_+)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}G_A^T(\tau_A S_A)^- G_A\Big)$$

where $|\tau_A S_A|_+$ is the Pseudo-Determinant, and $(\tau_A S_A)^-)$ is the Generalized inverse.
Define $rank(S_A) = q_A$, $rank(S_B) = q_B$, we have,

$$\begin{aligned} f(G_A) =& \frac{1}{((2\pi)^{q_A}\tau_A^{q_A}|S_A|_+)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2\tau_A}G_A^T(S_A)^- G_A\Big) \\ f(G_B) =& \frac{1}{((2\pi)^{q_B}\tau_B^{q_B}|S_B|_+)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2\tau_B}G_B^T(S_B)^- G_B\Big) \end{aligned}$$

therefore,

$$\log f(G_A) = constant - \frac{q_A}{2}\log\tau_A - \frac{1}{2}\log(|S_A|_+) - \frac{1}{2\tau_A}G_A^T S_A^- G_A$$
$$\log f(G_B) = constant - \frac{q_B}{2}\log\tau_B - \frac{1}{2}\log(|S_B|_+) - \frac{1}{2\tau_B}G_B^T S_B^- G_B$$
$$\log f(U|G_A, G_B) = constant - \frac{N-p}{2}\log\sigma - \frac{1}{2\sigma}\Big[(Y - G_A - G_B)^T(I - P_X)(Y - G_A - G_B)\Big]$$

EM algorithm treats G_A and G_B as missing values. So instead of estimating G_A and G_B and then plugging them into the $\log L$, EM algorithm calculate the expectation of $\log L$ given U and $\theta^{(t-1)}$, then based on $E(\log L|U, \theta^{(t-1)})$, $\theta^{(t)}$ are calculated by taking partial derivative.

$$E\Big(\log L\Big|U, \theta^{(t-1)}\Big) = E\Big(\log f(G_A)\Big|U, \theta^{(t-1)}\Big) + E\Big(\log f(G_B)\Big|U, \theta^{(t-1)}\Big) + E\Big(\log f(U|G_A, G_B)\Big|\theta^{(t-1)}\Big)$$

It is easy to find out that to estimate $\tau_A^{(t)}$, we just need to consider $E\Big(\log f(G_A)\Big|U, \theta^{(t-1)}\Big)$, so let

$$\frac{\partial E\Big(\log f(G_A)\Big|U, \theta^{(t-1)}\Big)}{\partial \tau_A} = 0$$

we have,

$$-\frac{q_A}{2\tau_A} + \frac{1}{2\tau_A^2}E\Big(G_A^T S_A^- G_A\Big|U, \theta^{(t-1)}\Big) = 0$$

$$E\Big(G_A^T S_A^- G_A\Big|U, \theta^{(t-1)}\Big) = \big(g_A^{(t-1)}\big)^T S_A^- g_A^{(t-1)} + tr(S_A^- v_A^{(t-1)})$$

plugging the expression of $g_A^{(t-1)}$ and $v_A^{(t-1)}$, finally we have

$$\tau_A^{(t)} = \tau_A^{(t-1)} + \frac{[\tau_A^{t-1}]^2}{q_A}\Big[Y^T P S_A P Y - tr(S_A P)\Big]$$

In the same way, we can estimate $\tau_B^{(t)}$ by

$$\tau_B^{(t)} = \tau_B^{(t-1)} + \frac{[\tau_B^{t-1}]^2}{q_B}\Big[Y^T P S_B P Y - tr(S_B P)\Big]$$

To estimate $\sigma^{(t)}$, we just need to consider $E\Big(\log f(U|G_A, G_B)\Big|\theta^{(t-1)}\Big)$, so the final expression of $\sigma^{(t)}$ is

$$\sigma^{(t)} = \frac{1}{N-p}\Big\{\Big[(Y^*)^T(I - P_X)Y^*\Big] + tr\Big[(I - P_X)\Big(\tau_A^{(t-1)}S_A - (\tau_A^{(t-1)})^2 S_A P S_A + \tau_B^{(t-1)}S_B - (\tau_B^{(t-1)})^2 S_B P S_B\Big)\Big]\Big\}$$

where $Y^* = Y - \tau_A^{(t-1)}S_A P Y - \tau_B^{(t-1)}S_B P Y$

References

Tzeng, J.-Y., Zhang, D., Chang, S.-M., Thomas, D. C., and Davidian, M. Gene-trati simiarity regression for multimarker-based association analysis. *Biometrics*, 65:822–832, 2009.