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 H_i as the $L \times 1$ haplotype vector of the ith individual in a sample of n subjects. Denote a $1 \times l_m$ 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 haplotype similarity between subjects i and j ($i \neq j$). However, Two measurements of S_{ij} based on unphased genotype are introduced. One is Average IBS (identical-by-descent) method. Then we have $S_{ij}^{ave} = \frac{1}{M}\sum_{m=1}^{M} s_{ij}^{ave}$, where

$$s_{ij}^{ave} = \begin{cases} 1, & G_{m,i} = G_{m,j} \text{ and } G_{m,i}, G_{m,j} \text{ are homozygous genotypes} \\ 0.5, & G_{m,i} = G_{m,j} \text{ and } G_{m,i}, G_{m,j} \text{ are heterozygote genotypes} \\ 0.5, & G_{m,i} \text{ and } G_{m,j} \text{ share only 1 allele} \\ 0, & \text{o.w} \end{cases}$$

Another way to describe the genetic similarity using the unphased genotype is so called Typical IBS, here the similarity matrix s_{ij}^{typ} is calculated by

$$s_{ij}^{typ} = \begin{cases} 1, & G_{m,i} = G_{m,j} \\ 0.5, & G_{m,i} \text{ and } G_{m,j} \text{ share only 1 allele} \\ 0, & \text{o.w} \end{cases}$$

We can see the difference between two measurements is how they consider the case when the genotypes from two individuals are the same but heterozygote. If the information for the pahsed haplotype is known, then we can tell whether the alleles from individual i is identical to the Aa from individual j. Since the haplotype is unknown, we can only have a guess. Average IBS takes the average value 0.5 while Typical IBS treats the two individuals always share the same haplotype.

The trait similarity $Z_{ij}$is computed by

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

where we assume $\mu_i = E(Y_i|X_i, H_i) = X_i\gamma$ 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

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

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

1.2 The Joint Test

The joint test tests the hypothesis H_0: $\tau_A = \tau_B = \phi = 0$. It is hard to directly derive the test for the testing. However, we find the score test by taking advantage of the connection between the similarity model and the variance component model which is:

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

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 is

$$T_{joint} = \frac{1}{2}\mathbf{Y}' P_0 S P_0 \mathbf{Y}$$

where the $P_0 = \sigma^{-2}(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \sigma^{-2}Q$ and $S = S^A + S^B + S^{AB}$. It is easy seen when H_0 is not true, T_{joint} which is as strictly increasing function. Therefore larger values of T_{joint} provides stronger evidence against H_0. This suggests that the testing procedure should be one sided.

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

$$T_{joint} = \frac{1}{2\sigma^4}\mathbf{Y}' QSQ\mathbf{Y} = \frac{1}{2\sigma^4}(\mathbf{Y} - \mathbf{X}\mu)' QSQ(\mathbf{Y} - \mathbf{X}\mu) \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

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

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^*: $\phi = 0$

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 using score test T_{epi} for the interaction testing, where

$$T_{Epi} = \frac{1}{2}\mathbf{Y}' P S^{AB} P \mathbf{Y}$$

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

$$T_{Epi} = \frac{1}{2}\mathbf{Y}' P S^{AB} P \mathbf{Y} \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

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

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 adjusted with Gene B, the corresponding hypothesis is H_0^1: $\tau_A = 0$ (vice versa for the Gene B main effect test). Similar to previous two tests, the following score test can be applied:

$$T_{OneGene} = \frac{1}{2}\mathbf{Y}' P_B S^A P_B \mathbf{Y}$$

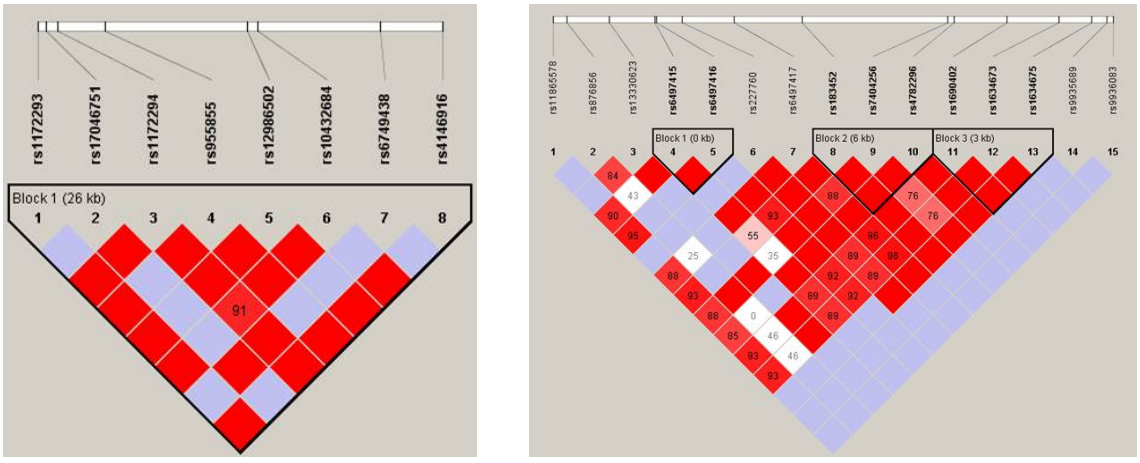
where $P = \mathbf{V}_B^{-1} - \mathbf{V}_B^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}_B^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_B^{-1}$ and $\mathbf{V}_B = \hat{\tau_B} S^B + \hat{\sigma} I$, and we can also derive that the distribution of $T_{OneGene}$ is also weighted χ^2 distribution, that is:

$$T_{OneGene} = \frac{1}{2}\mathbf{Y}' P_B S^A P_B \mathbf{Y} \sim \sum_{i=1}^{c} \lambda_i \chi_1^2$$

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

2 Simulation study

We studied the performance of our method using simulated data. 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.



(a) Gene RBJ

(b) Gene GPRC5B

Figure 1: LD patten 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

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

(a) Gene RBJ

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

(b) Gene GPRC5B

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

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 binary data: SNP is defined to have high LD (H) with other SNPs if its average LD > 0.18, otherwise its LD is low(L). A SNP labeled with C mean it is a common SNP whose MAF > 0.2 and R means its MAF ≤ 0.2.

Table 2: Simulation models

Main effect	Epi effect	Joint test			Epi test	One Gene test
		our method	PCA	PLS	our method	
(0,0)	0	0.051	0.052	0.051		

Table 3: Simulation models

Model 1	$Y = \beta \times (SNP_1^A \times SNP_1^B + SNP_2^A \times SNP_2^B) + e$
Model 2	$Y = \beta \times (SNP_1^A + SNP_1^B) + e$
Model 3	$Y = \beta \times (SNP_1^A \times SNP_2^A + SNP_1^B \times SNP_2^B) + e$

In the simulation studies, we try different models to test the performance of our method and other methods. Following is how we generate the simulated data. First we download the genotype of RBJ and GPRC5B from HapMap(http://hapmap.ncbi.nlm.nih.gov/) as a gene bank. Then we simulated the genotype person by person. For the *ith* individual, we random picked one person's genotype in the gene bank as the *ith* individual's genotype. After that, we picked one or two

SNPs(depends on the model) to generate the phenotype according to different models(shown in Table 2). Here, SNP_j^i stands for the jth causal SNP from Gene i, e is an error term following a normal distribution with mean 0 and variance 1. β is a scale parameter which controls the power. We tune β so that the power of different SNP sets are distinguishable.

Here are the results from our simulation study. First is the model 1, it is a typical interaction model. And the interaction only occurs between genes. For each Gene, I will pick 2 SNPs as the causal SNPs, and assign a interaction effect to the product of two casual SNPs from different genes, then I add the two interaction effect to generate the phenotype. In the figure, the x axis is the method, we compare our method with widely used PCA and PLS, the y axis is the power, the higher, the better. From the result, we can see that in most cases, our method is better than PCA and PLS. And the bigger the minor allele freq. that is more common SNP show up, the larger the power, and the higher LD, the larger the power is. But when the causal SNP is in Low LD, and the minor allele freq is quite small, our method does not perform as good as the other two methods.We also simulate some other models like additive model and within gene interaction model, both the results show that our method has a better ability to detect the interaction, i.e., has more power. And also, just like the conclusion from model 1, when the causal SNP is in Low LD and the minor allele freq is low, our method did not perform well.

3 Appendix

3.1 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.

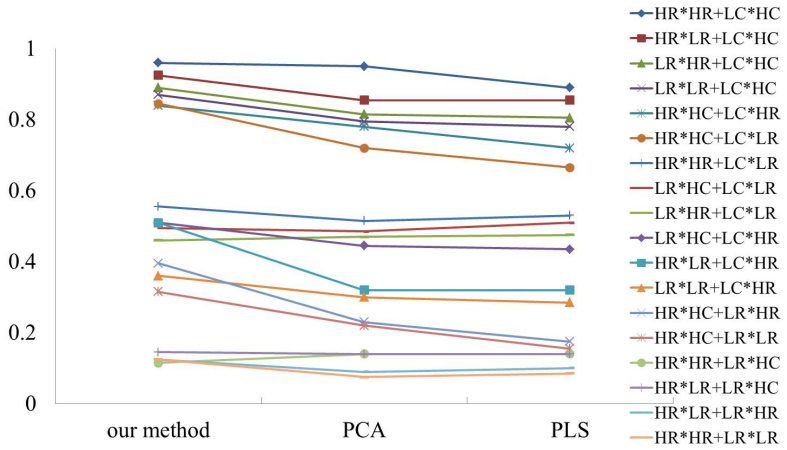
Simple algebra shows that the score statistic under $H_0 : \tau_A = \tau_B = \phi = 0$ is

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

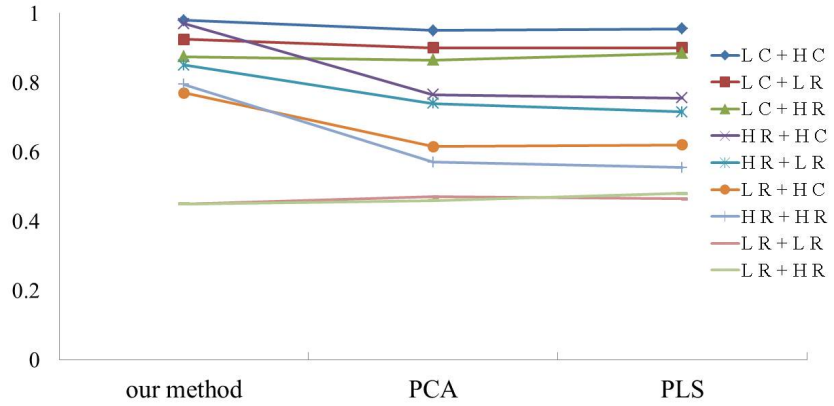
where the $P_0 = \sigma_e^{-2}(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \sigma_e^{-2}Q$. It is easy seen that $E(U_{\sigma_b^2}) = 0$ under H_0^* and when $\sigma_b^2 > 0$, $E(U_{\sigma_b^2}) = \sigma_b^2 \cdot tr(QSQS)/(2\sigma_e^4)$ which is a strictly increasing function. Therefore 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):

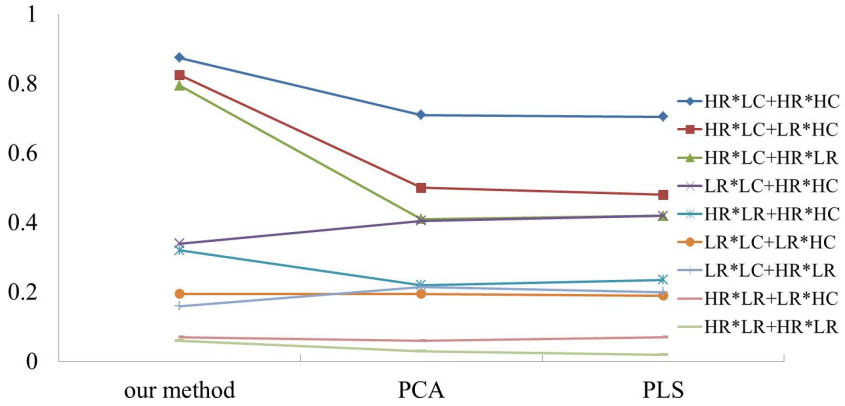
$$T = \frac{1}{2}\mathbf{Y}'P_0SP_0\mathbf{Y} = \frac{1}{2\sigma_e^4}\mathbf{Y}'QSQ\mathbf{Y}$$



(a) model 1



(b) model 2



(c) model 3

Figure 2: result for models

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

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

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

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

$$\begin{aligned} Cov(U, G_A) =& Cov(A^T X\gamma + A^T G_A + A^T G_B + A^T e, G_A) \\ =& Cov(A^T G_A, G_A) \\ =& A^T Cov(G_A, G_A) \\ =& \tau_A A^T S_A \end{aligned}$$

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

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

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 $Cov(G_A, G_B) = 0$

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

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

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

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

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

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,

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

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,

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

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

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

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

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

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

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

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

where $Y^*=Y-\tau_A^{(t-1)}S_APY-\tau_B^{(t-1)}S_BPY$