SCIE 3500 Final Report

Author
Sun Dajun
ID 20492572
dsunad@connect.ust.hk

Supervisor
Prof Yang Can
macyang@ust.hk

January 2020

Abstract

In this report, I'll summarize my work in this semester on cell-type-specific resolution epigenetics. It will first briefly explain why cell-type-specific resolution epigenetics are needed and then explain two popular models from the papers I've studied in detail. More attention will be paid to the HIRE(high-resolution EWAS) and the core part of this method, which is about the expectation maximization algorithm(EM algorithm). Additionally, I'll derive a formula for the shared variance case to deal with the problem of over-parameterization and also talk about some of my ideas to reduce computation time. At last, I'll design a simulated dataset to test the performance of the shared variance model.

1 Introduction

Epigenome-wide association studies (EWAS) aim to identify cytosine-phosphate-guanine (CpG) sites associated with phenotypes of interest, such as disease status, smoking history and age[3]. However, as the samples in EWAS are measured at bulk level rather than single-cell level, the observed methylation level for each sample is aggregated from distinct cell types, which makes it difficult to reveal the original single-cell level methylation data. As a result, most existing methods for EWAS can only detect CpG sites that are associated with phenotypes at aggregate level. It is notable that each cell type in our body performs some specific functions. Thus, disruption of cellular processes in particular cell types may lead to phenotypic alterations or diseases[2]. Therefore, in order to reveal cellular mechanisms affecting disease status of phenotypes, it is important to perform cell-type-specific studies.

Although developments in genomic profiling technologies have led to the availability of large bulk data sets with hundreds or thousands of samples it is still difficult to get large scale cell-type-specific data sets due to limitations in cell sorting and single-cell techniques[2]. It is therefore of significant importance to develop new statistical methods that can restore single-cell methylation profiles from currently available bulk data. In this report, I will summarize two currently states of art single-cell-resolution approaches for EWAS, namely the tensor composition analysis method(TCA) and the high-resolution EWAS method(HIRE).

2 HIRE Model[3]

Let's start by introducing some notations and assumptions. Let m be the number of CpG sites observed in the data set and n be the number of samples. For each sample i, $\mathbf{O_i} = \{O_{1i}, O_{2i}, ..., O_{mi}\}^T$ is the observed methylation levels of the m CpG sites. The cellular compositions $\mathbf{p_i} = \{p_{1i}, p_{2i}, ..., p_{Ki}\}^T$ where K is the total number of cell types. As cellular compositions may be affected by phenotypes and disease status, for each sample i we model a separate copy of $\mathbf{p_i}$. Also we have $\mathbf{x_i} = \{x_{i1}, x_{i1}, ..., p_{iq}\}^T$ to be the phenotype vector for ith sample where q is the number of considered phenotypes. In the HIRE model, we assume that for each sample i and each cell type k, its methylation comes from two aspects, one is the cell-type-specific baseline $\mu_{\mathbf{k}}$, which remain the same for that cell type among all samples. Another source of methylation comes from the phenotype effect, to model the phenotype effect, we define a matrix $\mathbf{B}$ of size $m \times K \times q$ representing the effect size of each phenotype ℓ on each cell type k and each CpG site j. To be more specific, the kth column of $\mathbf{B}_\ell$ which is $\mathbf{B}_{k\ell} = (\beta_{1k\ell}, \ldots, \beta_{mk\ell})^T$ reflects the association of phenotype ℓ with each of the m CpG sites in cell type k. Therefore in cell type k, sample i's cell-type-specific methylation profile, $\mathbf{u_{ik}}$ can be expressed as the summation of the to factors. $\mathbf{u}_{ik} = \boldsymbol{\mu}_k + \sum_{l=1}^{q} \mathbf{B}_{k\ell} x_{i\ell}$. Then,by considering the cell composition of individual i, for the observed methylation level of sample i at the m observation sites, we have

$$\mathbf{O_i} = \textstyle\sum_{\ell=1}^{q} \mathbf{B}_\ell \mathbf{x_{i\ell}} \mathbf{p_i} + \mu \mathbf{p_i},$$

where $\mu = (\mu_{\mathbf{1}}, ..., \mu_{\mathbf{k}})$.

In the HIRE model, we assume that for each specific cell type and under certain phenotype, the methylation level at each site is normally distributed around the mean u_{ijk}. This assumption is reasonable as the expression level of each gene is not uniform even for the same type of cells in one individual. Then we'll have

$$u_{ijk} = \mu_{jk} + \textstyle\sum_{\ell=1}^{q} \beta_{jk\ell} x_{i\ell} + \epsilon_{jk}$$

We further assume that for the observation of each sample at each site, there exists a normally distributed observation error ϵ_{ji}, then we have

$$O_{ji} = \sum_{k=1}^{K} \mu_{jk} p_{ki} + \epsilon_{ji},$$

To be more specific, we assume

$$u_{ijk} \sim N\left(\mu_{jk} + \sum_{\ell=1}^{q} \beta_{jk\ell} x_{i\ell}, \sigma_{jk}^2\right). \tag{1}$$

$$O_{ji} \sim N\left(\sum_{k=1}^{K} u_{ijk} p_{ki}, \sigma_{\epsilon j}^2\right). \tag{2}$$

We adopted a generalized EM algorithm to estimate the parameters, and the details about the EM algorithm are shown in next section.

2.1 EM Algorithm[1],[3]

The expectation–maximization (EM) algorithm is an iterative approach to find maximum likelihood or maximum a posterior probability (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. Basically it alternates between the expectation(E) step and the maximization(M) step. In the E step, we calculate the expectation of the log-likelihood on the latent variable based on current estimation of the parameters. In the M step, we re-estimate the parameters to maximize the expectation we got in E step, maximization methods generally include taking derivatives and set derivatives to zero or solving QP problem when some restrictions needed to be applied. It can be proved by using Jensen's inequality that this method will finally converge.

Here we talk about the details of the EM algorithm for our HIRE model. What we want to maximize is the observed data log likelihood, which is

$$L_o(\mathbf{\Theta}|\mathbf{O}) = \sum_{i=1}^{n}\sum_{j=1}^{m} log(N(O_{ji})), \tag{3}$$

where $\mathbf{\Theta}$ is the collection of parameters in the model. As O_{ji} forms the distribution

$$O_{ji} \sim N\left(\mathbf{u_{ij}^T p_i}, \sigma_{\epsilon,\mathbf{j}}^{\mathbf{2}}\right), \tag{4}$$

where $\mathbf{u_{ij}}$ itself is a latent variable. The observed data log likelihood is intraceable without giving any information about $\mathbf{u_{ij}}$. Therefore, we augment the missing data $\mathbf{u} = \{\mathbf{u_{ij}} : 1 \leq i \leq n,\ 1 \leq j \leq m\}$ to the observed data $\mathbf{O}$ and getting the complete data log likelihood is

$$\begin{aligned} L_c(\mathbf{\Theta}|\mathbf{O}, \mathbf{u}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} log(N(O_{ji}|\mathbf{u_{ij}})) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\{-log\sigma_{\epsilon,j} - \frac{(O_{ji} - \mathbf{u_{ij}}^T\mathbf{p_i})^2}{2\sigma_{\epsilon,j}^2} - \sum_{k=1}^{K} log\sigma_{jk} \\ &\qquad - \frac{1}{2}(\mathbf{u_{ij}} - \mu_\mathbf{j} - \mathbf{B}^{(j)}\mathbf{x_i})^T\Sigma_j^{-1}(\mathbf{u_{ij}} - \mu_\mathbf{j} - \mathbf{B}^{(j)}\mathbf{x_i})\} + Constant \end{aligned}$$

Then we need to calculate the distribution of $\mathbf{u_{ij}}$, as space is limited, I show the result directly

$$\mathbf{u_{ij}}|\mathbf{\Theta}^t, O_{ji} \sim N(\mu_\mathbf{ij}^\mathbf{t}, \Sigma_{ij}^t) \tag{5}$$

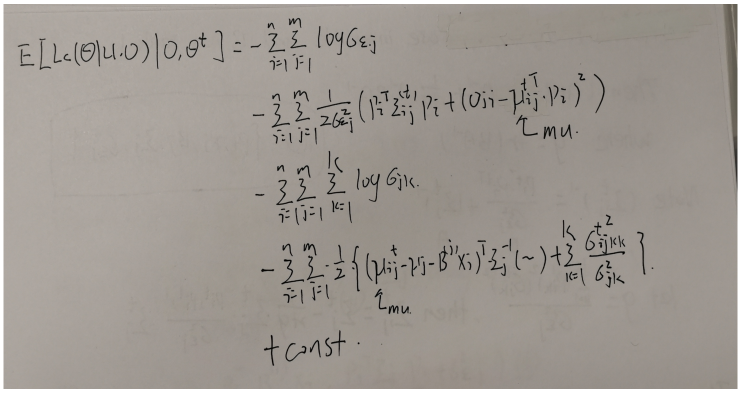
where

$$(\Sigma_{ij}^t)^{-1} = \frac{\mathbf{p_i^t}\mathbf{p_i^t}^T}{\sigma_{\epsilon,j}^2} + (\Sigma_j^t)^{-1}$$

$$(\mathbf{u_{ij}}^t)^T(\Sigma_{ij}^t)^{-1} = (\mu_\mathbf{j}^\mathbf{t} + \mathbf{B^t}^{(j)T}\mathbf{x_i})(\Sigma_j^t)^{-1} + \frac{O_{ji}\mathbf{p_i^t}^T}{\sigma_{\epsilon,j}^2}$$

Therefore the expectation of the complete data log likelihood with respect to $\mathbf{u}$ can be calculated and shown in Figure1.

Then by taking derivatives and quadratic optimization, we can gain the $(t+1)$th estimation for Θ(Figure2).



$$E[L_c(\Theta|u,O)|O,\theta^t] = -\sum_{i=1}^{n}\sum_{j=1}^{m}\log\sigma_{\varepsilon j}$$

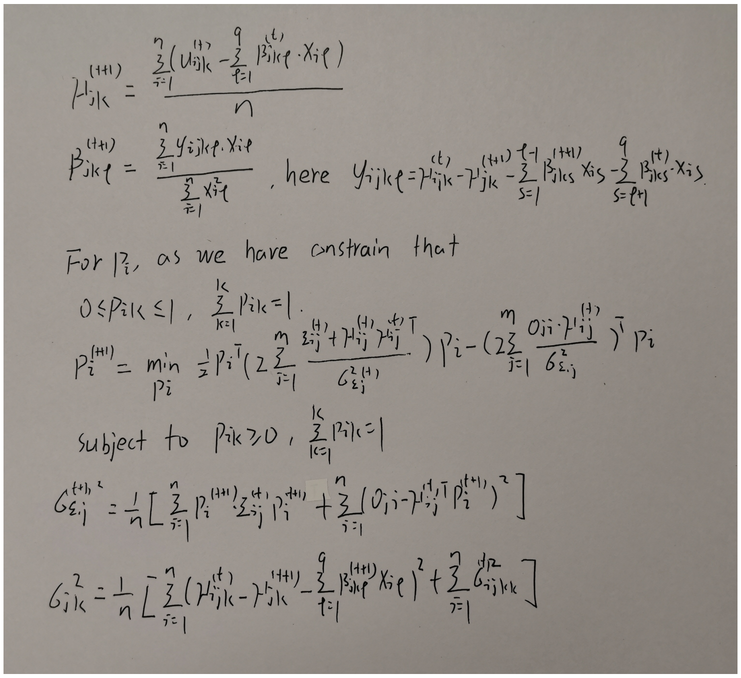
$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2\sigma_{\varepsilon j}^2}\left(p_i^T\Sigma_{ij}^{(t)}p_i+(O_{ji}-\underset{\uparrow\text{mu.}}{\mu_{ij}^{tT}}\cdot p_i)^2\right)$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K}\log\sigma_{jk}.$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}\left\{(\underset{\uparrow\text{mu.}}{\mu_{ij}^{t}}-\mu_j-\beta^{j)}x_i)^T\Sigma_j^{-1}(\sim)+\sum_{k=1}^{K}\frac{\sigma_{ijkk}^{t2}}{\sigma_{jk}^2}\right\}.$$

$$+\text{const}.$$

Figure 1: The expectation of the complete data log likelihood with respect to **u**



$$\mu_{jk}^{(t+1)}=\frac{\sum_{i=1}^{n}(u_{ijk}^{(t)}-\sum_{\ell=1}^{q}\beta_{jk\ell}^{(t)}\cdot x_{i\ell})}{n}$$

$$\beta_{jk\ell}^{(t+1)}=\frac{\sum_{i=1}^{n}y_{ijk\ell}\cdot x_{i\ell}}{\sum_{i=1}^{n}x_{i\ell}^2},\ \text{here } y_{ijk\ell}=\mu_{ijk}^{(t)}-\mu_{jk}^{(t+1)}-\sum_{s=1}^{\ell-1}\beta_{jks}^{(t+1)}x_{is}-\sum_{s=\ell+1}^{q}\beta_{jks}^{(t)}\cdot x_{is}.$$

For p_i, as we have constrain that

$$0\le p_{ik}\le 1,\ \sum_{k=1}^{K}p_{ik}=1.$$

$$p_i^{(t+1)}=\min_{p_i}\frac{1}{2}p_i^T\left(2\sum_{j=1}^{m}\frac{\Sigma_{ij}^{(t)}+\mu_{ij}^{(t)}\mu_{ij}^{(t)T}}{\sigma_{\varepsilon j}^{2(t)}}\right)p_i-\left(2\sum_{j=1}^{m}\frac{O_{ji}\cdot\mu_{ij}^{(t)}}{\sigma_{\varepsilon j}^2}\right)^T p_i$$

subject to $p_{ik}\ge 0,\ \sum_{k=1}^{K}p_{ik}=1$

$$\sigma_{\varepsilon j}^{(t+1),2}=\frac{1}{n}\left[\sum_{i=1}^{n}p_i^{(t+1)}\Sigma_{ij}^{(t)}p_i^{(t+1)}+\sum_{i=1}^{n}(O_{ji}-\mu_{ij}^{(t)T}p_i^{(t+1)})^2\right]$$

$$\sigma_{jk}^2=\frac{1}{n}\left[\sum_{i=1}^{n}(\mu_{ijk}^{(t)}-\mu_{jk}^{(t+1)}-\sum_{\ell=1}^{q}\beta_{jk\ell}^{(t+1)}x_{i\ell})^2+\sum_{i=1}^{n}\sigma_{ijkk}^{t2}\right]$$

Figure 2: The optimizers of the expectation

2.2 Possible Improvements to HIRE

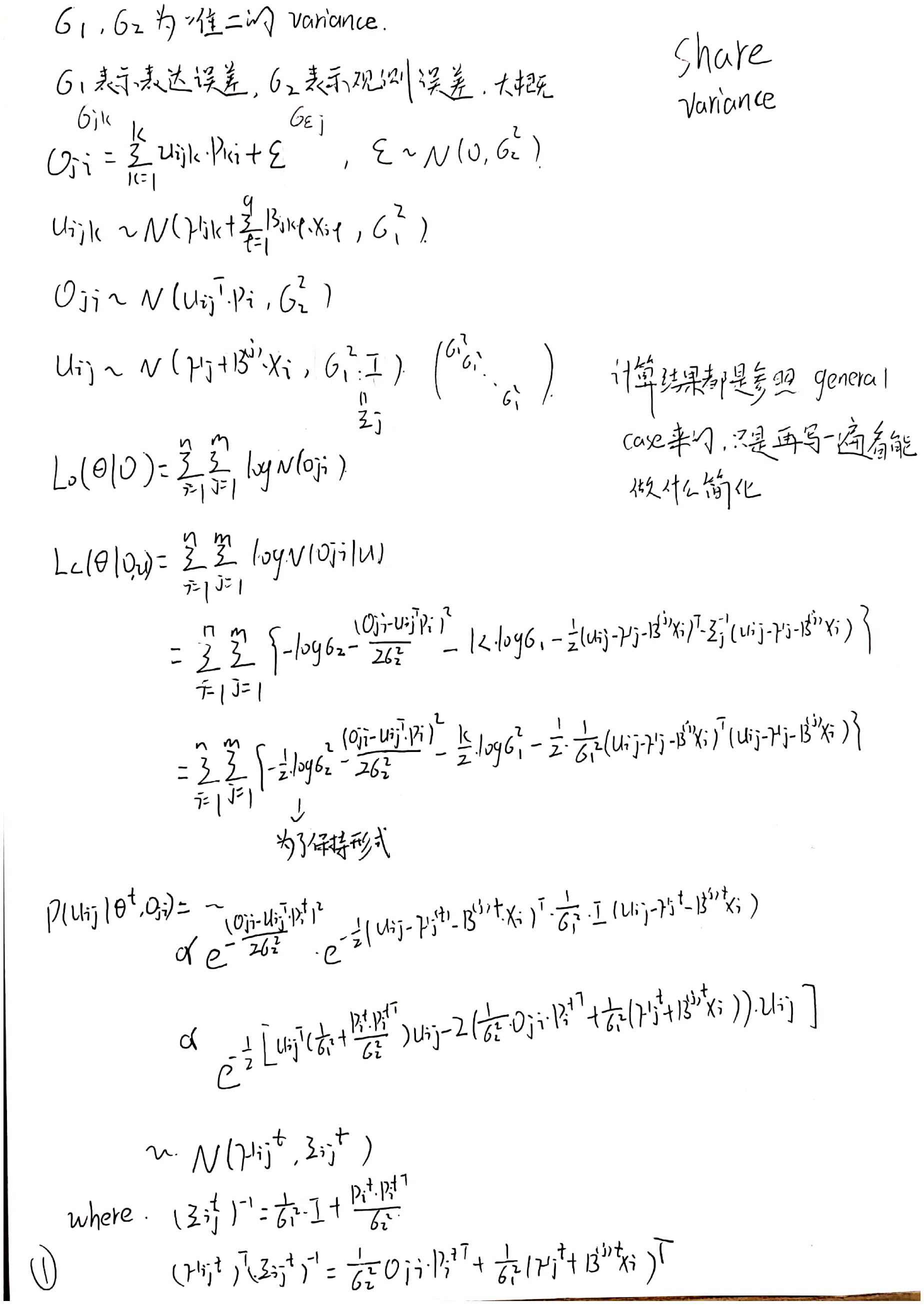
As found by Prof Yang in some experiments, there are two major problems in the current HIRE model. First, the algorithm is not very efficient and it converges very slowly, making it difficult to be applied to large data sets containing the methylation level of tens of thousands of CpG sites. Second, there exists the problem of overparameterization since we have too many parameters. For these problems, I've proposed some possible solutions. However, as time is limited, I haven't tried any of them. In this report, I'll only describe my ideas without giving any experiment results.

While running the HIRE software on real data sets, I found that the error terms, which are σ_{jk}s and $\sigma_{\epsilon j}$s are generally very small and their values are quite similar. Therefore, it is a natural idea to try to merge them into only two variables. To be more specific, we can use σ_1 to represent all σ_{jk}s and use σ_2 to stand for all $\sigma_{\epsilon j}$s. By doing so we can reduce the number of parameters by $m(K+1)-2$ which is a considerable amount as m is generally very large. Additionally, computation time can be shortened. I've derived the formula for the EM algorithm for this shared variance case, in the next step, I'll try to implement the code and see whether it will work. Basically, I just need to switch all σ_{jk}s to σ_1 and switch all $\sigma_{\epsilon,j}$s to σ_2. Note that in this case, the $\Sigma_{ij}^{(t)}$ can be expressed simply as

$$\Sigma_{ij}^{(t)} = (\sigma_1)^2\mathbf{I} - \frac{1}{(\sigma_2)^2 + (\sigma_1)^2\sum_{k=1}^{K} p_{ik}^2}\mathbf{p_i}\mathbf{p_i}^T, \qquad (6)$$

which does not involve complicate matrix computations. Some other computations can also be simplified. From this aspect of view, I developed an EM algorithm for the sgared variance version.

Let σ_1 denotes all σ_{jk}s in previous model and σ_2 denotes all $\sigma_{\epsilon j}^2$s, all other notations and settings remain unchanged. The detailed calculations of the shared variance version are shown in the following images.



σ_1, σ_2 为唯二的 variance.

Share variance

σ_1表示表达误差, σ_2表示观测误差. 大概

$$O_{ji} = \sum_{k=1}^{K} \overset{\sigma_{jk}}{u_{ijk}} \cdot p_{ki} + \overset{\sigma_{\varepsilon j}}{\varepsilon}, \quad \varepsilon \sim N(0, \sigma_2^2)$$

$$u_{ijk} \sim N(\mu_{jk} + \sum_{t=1}^{q} \beta_{jkt} \cdot x_{it}, \sigma_1^2)$$

$$O_{ji} \sim N(u_{ij}^T \cdot p_i, \sigma_2^2)$$

$$u_{ij} \sim N(\mu_j + \beta^{(j)} \cdot x_i, \underset{\Sigma_j}{\underset{\parallel}{\sigma_1^2 \cdot I}}) \quad \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_1 \end{pmatrix}$$

计算结果都是参照 general case来的, 只是再写一遍看能做什么简化

$$L_o(\theta|O) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log N(O_{ji})$$

$$L_c(\theta|O,u) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log N(O_{ji}|u)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ -\log \sigma_2 - \frac{(O_{ji} - u_{ij}^T p_i)^2}{2\sigma_2^2} - K \cdot \log \sigma_1 - \frac{1}{2}(u_{ij} - \mu_j - \beta^{(j)} x_i)^T \cdot \Sigma_j^{-1} (u_{ij} - \mu_j - \beta^{(j)} x_i) \right\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \underset{\text{为了保持形式}}{\underset{\downarrow}{-\frac{1}{2} \cdot \log \sigma_2^2}} - \frac{(O_{ji} - u_{ij}^T \cdot p_i)^2}{2\sigma_2^2} - \frac{K}{2} \cdot \log \sigma_1^2 - \frac{1}{2} \cdot \frac{1}{\sigma_1^2} (u_{ij} - \mu_j - \beta^{(j)} x_i)^T (u_{ij} - \mu_j - \beta^{(j)} x_i) \right\}$$

$$p(u_{ij}|\theta^t, O_{ji}) = \sim$$

$$\propto e^{-\frac{(O_{ji} - u_{ij}^T \cdot p_i^t)^2}{2\sigma_2^2}} \cdot e^{-\frac{1}{2}(u_{ij} - \mu_j^{(t)} - \beta^{(j)t} \cdot x_i)^T \cdot \frac{1}{\sigma_1^2} \cdot I (u_{ij} - \mu_j^t - \beta^{(j)t} x_i)}$$

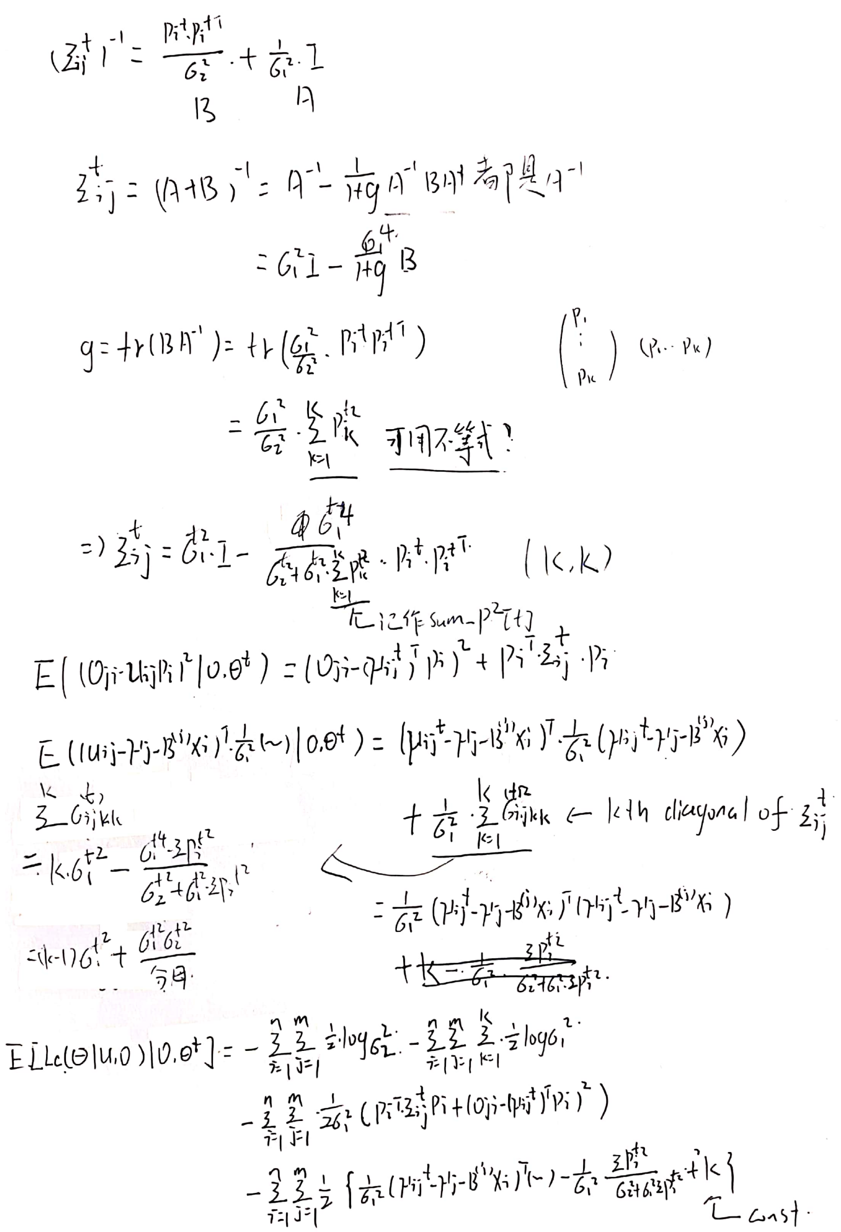
$$\propto e^{-\frac{1}{2}\left[u_{ij}^T \left(\frac{1}{\sigma_1^2} + \frac{p_i^t \cdot p_i^{tT}}{\sigma_2^2} \right) u_{ij} - 2\left(\frac{1}{\sigma_2^2} \cdot O_{ji} \cdot p_i^{tT} + \frac{1}{\sigma_1^2} (\mu_j^t + \beta^{(j)t} x_i) \right) \cdot u_{ij} \right]}$$

$$\sim N(\mu_{ij}^t, \Sigma_{ij}^t)$$

where. $(\Sigma_{ij}^t)^{-1} = \frac{1}{\sigma_1^2} \cdot I + \frac{p_i^t \cdot p_i^{tT}}{\sigma_2^2}$

① $(\mu_{ij}^t)^T (\Sigma_{ij}^t)^{-1} = \frac{1}{\sigma_2^2} O_{ji} \cdot p_i^{tT} + \frac{1}{\sigma_1^2} (\mu_j^t + \beta^{(j)t} x_i)^T$

Figure 3: Shared variance EM-p1



$$(\Sigma_{ij}^{t})^{-1} = \underbrace{\frac{p_i^{t} p_i^{tT}}{\sigma_2^2}}_{B} + \underbrace{\frac{1}{\sigma_1^2} I}_{A}$$

$$\Sigma_{ij}^{t} = (A+B)^{-1} = A^{-1} - \frac{1}{1+g} \underline{A^{-1}} B A^{-1}$$ 都是 A^{-1}

$$= \sigma_1^2 I - \frac{\sigma_1^4}{1+g} B$$

$$g = tr(BA^{-1}) = tr\left(\frac{\sigma_1^2}{\sigma_2^2} \cdot p_i^{t} p_i^{tT}\right) \qquad \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix} (p_1 \cdots p_K)$$

$$= \frac{\sigma_1^2}{\sigma_2^2} \cdot \underline{\sum_{k=1}^{K} p_k^{t2}}$$ <u>可用不等式!</u>

$$\Rightarrow \Sigma_{ij}^{t} = \sigma_1^{t2} \cdot I - \frac{\sigma_1^{t4}}{\sigma_2^{t2} + \sigma_1^{t2} \cdot \underbrace{\sum_{k=1}^{K} p_k^{t2}}_{\text{记作 sum\_p}^2[t]}} \cdot p_i^{t} \cdot p_i^{tT} \qquad (K, K)$$

$$E\left((O_{ji} - u_{ij} p_i)^2 \mid O, \theta^t\right) = (O_{ji} - (\mu_{ij}^{t})^T p_i)^2 + p_i^T \Sigma_{ij}^{t} \cdot p_i$$

$$E\left((u_{ij} - \mu_j - B^{(j)} x_i)^T \frac{1}{\sigma_1^2} (\sim) \mid O, \theta^t\right) = (\mu_{ij}^{t} - \mu_j - B^{(j)} x_i)^T \cdot \frac{1}{\sigma_1^2} (\mu_{ij}^{t} - \mu_j - B^{(j)} x_i)$$

$$+ \frac{1}{\sigma_1^2} \cdot \underline{\sum_{k=1}^{K} \sigma_{ijkk}^{t2}} \leftarrow \text{kth diagonal of } \Sigma_{ij}^{t}$$

$$\sum^{K} \sigma_{ijkk}^{(t)}$$

$$= K \cdot \sigma_1^{t2} - \frac{\sigma_1^{t4} \cdot \sum p_i^{t2}}{\sigma_2^{t2} + \sigma_1^{t2} \sum p_i^{t2}}$$

$$= (K-1)\sigma_1^{t2} + \frac{\sigma_1^{t2} \sigma_2^{t2}}{\text{分母}}$$

$$= \frac{1}{\sigma_1^2} (\mu_{ij}^{t} - \mu_j - B^{(j)} x_i)^T (\mu_{ij}^{t} - \mu_j - B^{(j)} x_i)$$

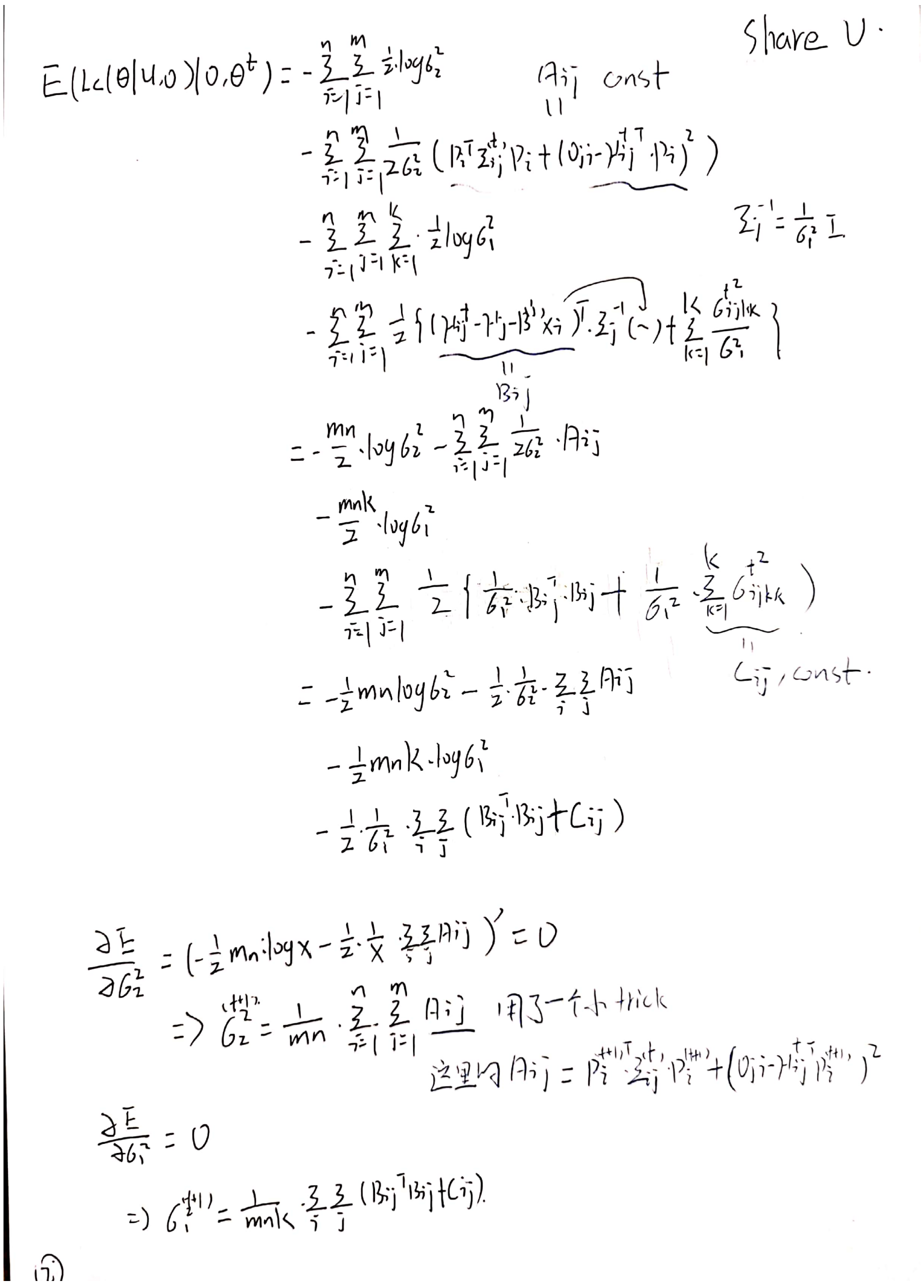
$$+ \cancel{K - \frac{1}{\sigma_1^2} \cdot \frac{\sum p_i^{t2}}{\sigma_2^2 + \sigma_1^2 \sum p_i^{t2}}}$$

$$E[L_c(\Theta \mid U, O) \mid O, \theta^t] = -\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{2} \log \sigma_2^2 - \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{K} \frac{1}{2} \log \sigma_1^2$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{2\sigma_1^2} \left(p_i^T \Sigma_{ij}^{t} p_i + (O_{ji} - (\mu_{ij}^{t})^T p_i)^2\right)$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{2} \left\{ \frac{1}{\sigma_1^2} (\mu_{ij}^{t} - \mu_j - B^{(j)} x_i)^T (\sim) - \frac{1}{\sigma_1^2} \frac{\sum p_i^{t2}}{\sigma_2^2 + \sigma_1^2 \sum p_i^{t2}} + \underbrace{K}_{\text{const}} \right\}$$

Figure 4: Shared variance EM-p2



Share U.

$$E(L_c(\theta|U,O)|O,\theta^t) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}\log\sigma_2^2$$

$$\underset{\parallel}{A_{ij}} \text{ const}$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2\sigma_2^2}\left(V_i^T\Sigma_{ij}^{(t)}V_i+(O_{ij}-\mu_{ij}^{t\,T}\cdot V_i)^2\right)$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K}\cdot\frac{1}{2}\log\sigma_1^2$$

$$\Sigma_j^{-1}=\frac{1}{\sigma_1^2}I$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}\Big\{\underbrace{(\mu_{ij}^t-\mu_j^t-B^{(t)}X_i)}_{\substack{\parallel\\ B_{ij}}}{}^T\cdot\Sigma_j^{-1}(\sim)+\sum_{k=1}^{K}\frac{\sigma_{ijkk}^{t\,2}}{\sigma_1^2}\Big\}$$

$$=-\frac{mn}{2}\cdot\log\sigma_2^2-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2\sigma_2^2}\cdot A_{ij}$$

$$-\frac{mnK}{2}\cdot\log\sigma_1^2$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}\Big\{\frac{1}{\sigma_1^2}\cdot B_{ij}^T\cdot B_{ij}+\frac{1}{\sigma_1^2}\cdot\underbrace{\sum_{k=1}^{K}\sigma_{ijkk}^{t\,2}}_{\substack{\parallel\\ C_{ij},\text{ const.}}}\Big)$$

$$=-\frac{1}{2}mn\log\sigma_2^2-\frac{1}{2}\cdot\frac{1}{\sigma_2^2}\cdot\sum_i\sum_j A_{ij}$$

$$-\frac{1}{2}mnK\cdot\log\sigma_1^2$$

$$-\frac{1}{2}\cdot\frac{1}{\sigma_1^2}\cdot\sum_i\sum_j\left(B_{ij}^T\cdot B_{ij}+C_{ij}\right)$$

$$\frac{\partial E}{\partial\sigma_2^2}=\left(-\frac{1}{2}mn\cdot\log x-\frac{1}{2}\cdot\frac{1}{x}\cdot\sum_i\sum_j A_{ij}\right)'=0$$

$$\Rightarrow \sigma_2^{2\,(t+1)}=\frac{1}{mn}\cdot\sum_{i=1}^{n}\sum_{j=1}^{m}\underline{A_{ij}}$$ 用了一个小 trick

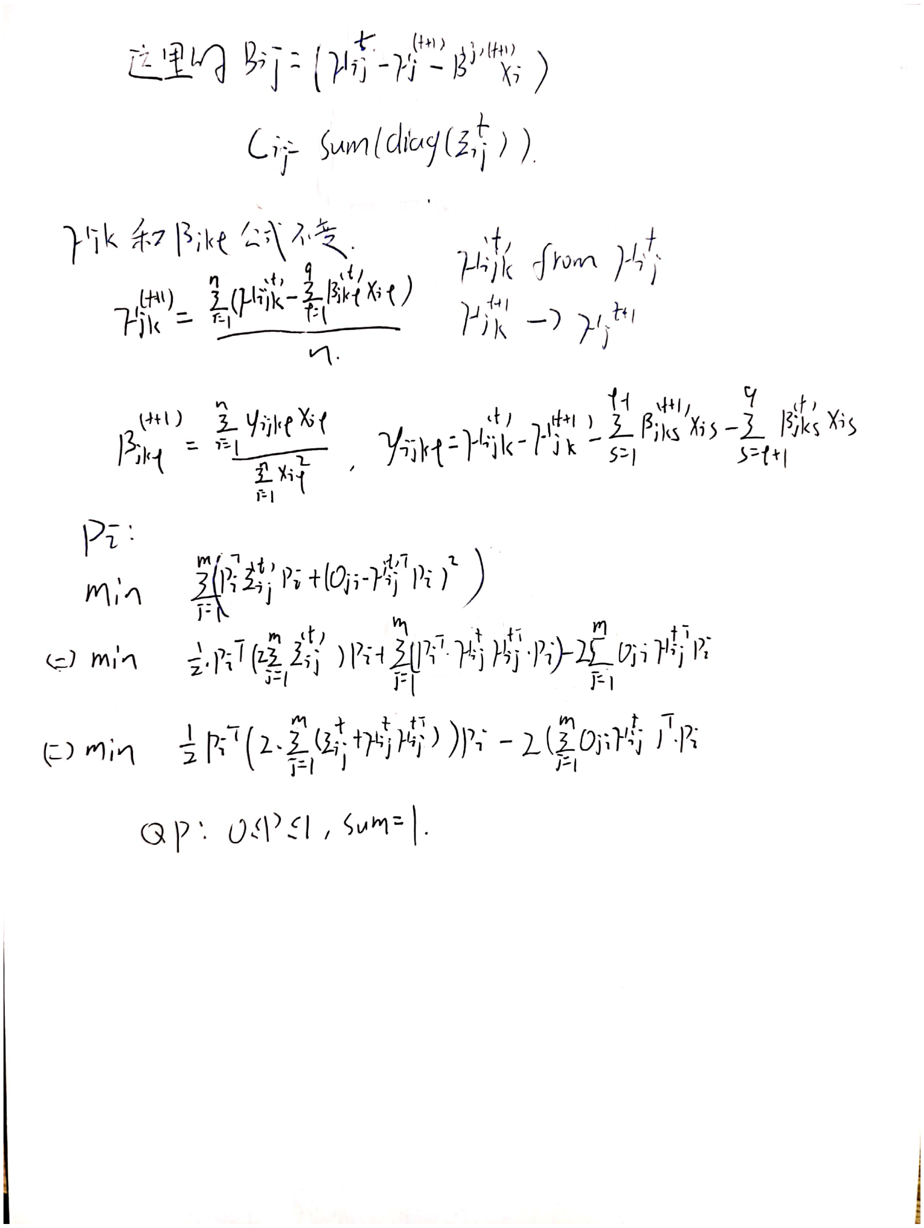
这里的 $A_{ij}=V_i^{(t+1)T}\Sigma_{ij}^{(t)}V_i^{(t+1)}+(O_{ij}-\mu_{ij}^{t\,T}V_i^{(t+1)})^2$

$$\frac{\partial E}{\partial\sigma_1^2}=0$$

$$\Rightarrow \sigma_1^{2\,(t+1)}=\frac{1}{mnK}\cdot\sum_i\sum_j\left(B_{ij}^T B_{ij}+C_{ij}\right).$$

(7)

Figure 5: Shared variance EM-p3



这里的 $B_{ij} = (\mu_{ij}^{t} - \mu_j^{(t+1)} - \beta^{j,(t+1)} x_i)$

C_{ij} sum(diag(Σ_{ij}^{t})).

μ_{jk} 和 β_{jkl} 公式不变.

$$\mu_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n}(\mu_{ijk}^{(t)} - \sum_{l=1}^{q}\beta_{jkl}^{(t)} x_{il})}{n}.$$

$\mu_{ijk}^{(t)}$ from μ_{ij}^{t}

$\mu_{jk}^{t+1} \to \mu_j^{t+1}$

$$\beta_{jkl}^{(t+1)} = \frac{\sum_{i=1}^{n} y_{ijkl} x_{il}}{\sum_{i=1}^{n} x_{il}^{2}}, \quad y_{ijkl} = \mu_{ijk}^{(t)} - \mu_{jk}^{(t+1)} - \sum_{s=1}^{l-1}\beta_{jks}^{(t+1)} x_{is} - \sum_{s=l+1}^{q}\beta_{jks}^{(t)} x_{is}$$

p_i:

$$\min \sum_{j=1}^{m}\left(p_i^{T}\Sigma_{ij}^{(t)}p_i + (O_{ji} - \mu_{ij}^{(t),T}p_i)^2\right)$$

$$\Leftrightarrow \min \frac{1}{2}\cdot p_i^{T}\left(2\sum_{j=1}^{m}\Sigma_{ij}^{(t)}\right)p_i + \sum_{j=1}^{m}(p_i^{T}\cdot\mu_{ij}^{t}\mu_{ij}^{tT}\cdot p_i) - 2\sum_{j=1}^{m}O_{ji}\mu_{ij}^{tT}p_i$$

$$\Leftrightarrow \min \frac{1}{2}p_i^{T}\left(2\cdot\sum_{j=1}^{m}(\Sigma_{ij}^{t} + \mu_{ij}^{t}\mu_{ij}^{tT})\right)p_i - 2\left(\sum_{j=1}^{m}O_{ji}\mu_{ij}^{t}\right)^{T}\cdot p_i$$

QP: $0 \le p \le 1$, sum=1.

Figure 6: Shared variance EM-p4

3 TCA Model[2]

The tensor composition analysis(TCA) model is not significantly different from the HIRE model, except for the fact that it divides the influence of phenotypes into two levels: cell-type specific level and global level. Let C^1 be a $p_1 \times n$ matrix of p_1 covariates that may potentially affect methylation levels in a cell-type-specific manner. And C^2 be a $p_2 \times n$ matrix of p_2 global covariates that affect the observed methylation level. Let Z^i_{kj} and X_{ij} denotes the methylation level of individual i in cell type k at site j and the observed methylation level of the i-th individual in cell type j respectively. Then we have:

$$Z^i_{kj} = \mu_{kj} + (c_i^{(1)})^T \beta_h^j + \epsilon^i_{kj} \tag{7}$$

$$X_i j = (c_i^{(2)})^T \delta_j + \sum_{k=1}^{K} w_{ki} Z^i_{kj} + \epsilon_{ij} \tag{8}$$

Where ϵ_{ij} and ϵ^i_{kj} denotes the randomness, β and δ denotes the influence factor and w_{ij} are the weights(compositions). Note that this model setting is similar to the test data I designed later.

4 Designing of test data

In the model, we assume that the contribution of each phenotype to the observed methylation level of each CpG site determined and the observed methylation level is only determined by the cell type average value and the observed phenotype plus a random variance. However, we should notice that in reality, we are not able to observed all phenotype and therefore there should be a specific term representing the error caused by the non-observed phenotypes and this influence is independent from the random error. To be speific, let n be the number of cells, m be the number of CpG sites, k be the number of cell types, p1 be the number of observed phenotypes and p2 be the number of non-observed phenotypes. The relationship between the observed methylation level of cell i at site j and those phenotypes should be:

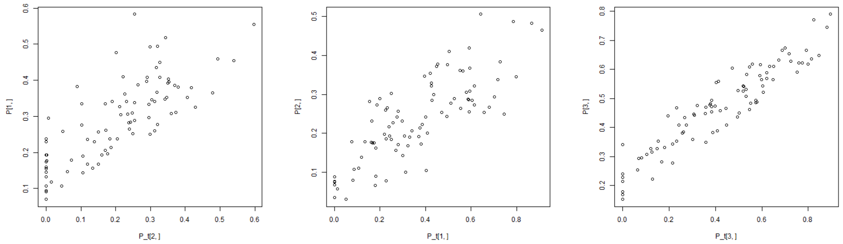
$$u_{ijk} = \mu_{jk} + \textstyle\sum_{\ell=1}^{p1} \beta_{jk\ell} x_{i\ell} + \sum_{\ell=1}^{p2} \beta^{'}_{jk\ell} x_{i\ell} + \epsilon_{jk}$$

$$O_{ij} = \textstyle\sum_{k=1}^{K} u_{ijk} p_{ki} + \epsilon$$

where β and $\beta^{'}$ denotes the influence factor and ϵ is the random error. This data set together with the real data set will be used to test the performance of the shared variance model.

5 Experiments

To test the performance of the shared-variance model, I designed the test data described above using similar setting as in [3]. To be specific, let n=180, n1=60(control), n2=120(test), m=2000, p1=2, p2=2, k=3 and initialize μ_{jk}s randomly. The observed



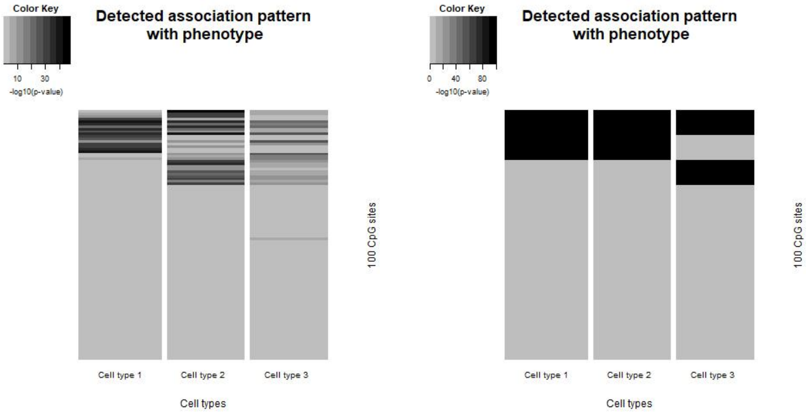
(a) P1

(b) P2

(c) P3

Figure 7: P1-P3 shows the estimated cell type compositions compared with the true composition. You can see that it almost lies on a straight line so the estimation is relatively efficient.

methylation matrix O can therefore be computed and then fed into the model. Some experiment results are shown and explained in Figure 7-Figure 9.



(a) Test

(b) True

Figure 8: These two images compare the estimated and true CpG sites that corresponds to phenotype, again you can see the estimation is relatively accurate.

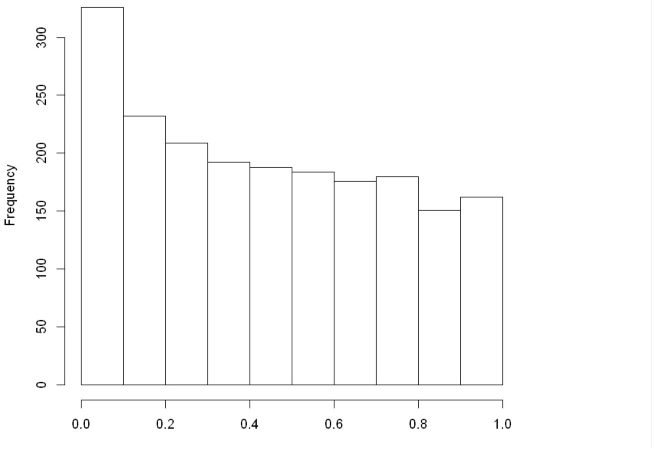


Figure 9: This picture shows the histogram of calculated P-values. It follows a nice distribution.

References

[1] Nan M Laird Donald B Rubin Arthur P Dempster. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society*, 1977.

[2] Brooke Rhead Lindsey A. Criswell Lisa F. Barcellos Eleazar Eskin Saharon Rosset Sriram Sankararaman Eran Halperin Elior Rahmani, Regev Schweiger. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. In *Nature Communications*, 2019.

[3] Can Yang Yingying Wei Xiangyu Luo. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. In *Nature Communications*, 2019.