Solving Correlation Motifs via EM algorithm

Zhiwei Ma

October 17, 2017

1 Introduction

This article contains introduction about CorMotif model by *Wei and Li* (2015) and detail implementation of EM algorithm in CorMotif. The original CorMotif is based on *limma* (*Smyth*, 2004). In this article, we expand CorMotif to any distribution. For Gaussian model, we introduce the iterative formula to estimate its parameters.

2 Correlation Motif Model

Suppose there are n genes and R studies for each gene. Let x_{ir} denote the statistics of gene i in study r, where $i = 1, 2, \ldots, n$ and $r = 1, 2, \ldots, R$. The collection of all observed data is

$$X = \{x_{ir} | i = 1, 2, \ldots, n; r = 1, 2, \ldots, R\}.$$

For each study, the statistics x_{ir} may follow two different distribution: one from the *null* hypothesis, the other from the alternative hypothesis. Let $y_{ir} = 1$ or 0 denote whether

x_{ir} comes from the *alternative* hypothesis. Given the state y_{ir}, assume x_{ir} follows

$$f_{r0}(x_{ir}) := f_r(x_{ir}|y_{ir} = 0), \tag{1}$$

$$f_{r1}(x_{ir}) := f_r(x_{ir}|y_{ir} = 1). \tag{2}$$

The states of gene i can be expressed as $Y_i = (y_{i1}, \ldots, y_{iR})^T$. For R studies, there are totally 2^R possible configurations. One way to study the correlation between different studies is to find the frequencies of each configuration among n genes. However, we need at least $O(2^R)$ samples to calculate each frequency, which will increase rapidly as R increases. Here we introduce the Correlation Motifs.

CorMotif adopts a hierarchical mixture model. It assumes that all genes fall into K classes. In addtion, it assumes

Assumption 1 *Each gene i is assigned to a class label z_i, here $z_i \in \{1, 2, \ldots, K\}$. The prior distribution for z_i is $Pr(z_i = k) = \pi_k, k = 1, 2, \ldots, K$. Thus, we have $\sum_k \pi_k = 1$ and denote $\pi = (\pi_1, \ldots, \pi_K)$.*

Assumption 2 *Given gene's class label z_i, gene's state y_{ir} are independent, following $Pr(y_{ir} = 1|z_i = k) = q_{kr}$. For the k^{th} class, denote $q_k = (q_{k1}, \ldots, q_{kR})^T$.*

Assumption 3 *Given the gene's state y_{ir}, the statistic x_{ir} are independently following (1) and (2).*

Let $Z = (z_1, \ldots, z_n)$ denote the class membership and $Q = (q_1, \ldots, q_K)^T$ is the $K \times R$ matrix. For gene i and study r we have

$$p(x_{ir}, y_{ir}|z_i = k, \pi, Q) = [q_{kr} f_{r1}(x_{ir})]^{y_{ir}} [(1 - q_{kr}) f_{r0}(x_{ir})]^{1-y_{ir}}.$$

Thus,

$$p(X_i, Y_i|z_i = k, \pi, Q) = \prod_{r=1}^{R} [q_{kr} f_{r1}(x_{ir})]^{y_{ir}} [(1 - q_{kr}) f_{r0}(x_{ir})]^{1-y_{ir}},$$

here $X_i = (x_{i1}, \ldots, x_{iR})^T$. We can get

$$p(X_i, Y_i, z_i|\pi, Q) = \prod_{k=1}^{K} \{\pi_k \prod_{r=1}^{R} [q_{kr} f_{r1}(x_{ir})]^{y_{ir}} [(1 - q_{kr}) f_{r0}(x_{ir})]^{1-y_{ir}}\}^{\mathbb{I}(z_i=k)},$$

where $\mathbb{I}$ is an indicator function. Therefore, based on above formulas, the joint probability distribution of X, Y and Z condtional on π and Q is

$$p(X, Y, Z|\pi, Q) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k \prod_{r=1}^{R} [q_{kr} f_{r1}(x_{ir})]^{y_{ir}} [(1 - q_{kr}) f_{r0}(x_{ir})]^{1-y_{ir}}\}^{\mathbb{I}(z_i=k)}. \quad (3)$$

In the joint probability distribution above, only X is observed, Y and Z are missing values or *latent* data. π and Q are unknown parameters. The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data

$$\begin{aligned} L(\pi, Q; X) &= p(X|\pi, Q) = \sum_{Y} \sum_{Z} p(X, Y, Z|\pi, Q) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{r=1}^{R} [q_{kr} f_{r1}(x_{ir}) + (1 - q_{kr}) f_{r0}(x_{ir})]. \end{aligned} \quad (4)$$

However, it is unrealistic to obtain the optimal values for π and Q by maximizing above formula directly. Instead, we will apply the EM algorithm to handle this problem.

3 Estimation Method

3.1 Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm seeks to find the MLE of the marginal likelihood by iteratively apply these two steps:

- **E-step:** Calculate the expected value of log-likelihood function, with respect to the conditional distribution of Y, Z given X under the current estimate of the parameters $(\pi^{(t)}, Q^{(t)})$:

$$Q(\pi, Q|\pi^{(t)}, Q^{(t)}) = E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\log L(\pi, Q; X, Y, Z)].$$

- **M-step:** Find the parameter that maximizes this quantity:

$$(\pi^{(t+1)}, Q^{(t+1)}) = \underset{(\pi,Q)}{\operatorname{argmax}}\, Q(\pi, Q|\pi^{(t)}, Q^{(t)}).$$

In the E-step, we have

$$\begin{aligned}
& \log L(\pi, Q; X, Y, Z) = \log p(X, Y, Z|\pi, Q) \\
= & \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(z_i = k)\{\log \pi_k + \sum_{r=1}^{R} y_{ir}[\log q_{kr} + \log f_{r1}(x_{ir})] \\
+ & \sum_{r=1}^{R}(1 - y_{ir})[\log(1 - q_{kr}) + \log f_{r0}(x_{ir})]\} \\
= & \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(z_i = k)\log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(z_i = k)\{\sum_{r=1}^{R} y_{ir}[\log q_{kr} + \log f_{r1}(x_{ir})] \\
+ & \sum_{r=1}^{R}(1 - y_{ir})[\log(1 - q_{kr}) + \log f_{r0}(x_{ir})]\}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
& Q(\pi, Q|\pi^{(t)}, Q^{(t)}) = E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\log L(\pi, Q; X, Y, Z)] \\
= & \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R} p_{ikr1}[\log q_{kr} + \log f_{r1}(x_{ir})] \\
+ & \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R} p_{ikr0}[\log(1 - q_{kr}) + \log f_{r0}(x_{ir})], \qquad (5)
\end{aligned}
$$

where we denote

$$
\begin{aligned}
p_{ik} &= E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\mathbb{I}(z_i = k)], \\
p_{ikr1} &= E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\mathbb{I}(z_i = k) y_{ir}], \\
p_{ikr0} &= E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\mathbb{I}(z_i = k)(1 - y_{ir})].
\end{aligned}
$$

It is easy to see that $p_{ik} = p_{ikr1} + p_{ikr1}$. We can compute p_{ik} and p_{ikr1} by

$$
\begin{aligned}
p_{ik} &= E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\mathbb{I}(z_i = k)] = Pr(z_i = k|X_i, \pi^{(t)}, Q^{(t)}) \\
&= \frac{p(z_i = k, X_i|\pi^{(t)}, Q^{(t)})}{p(X_i|\pi^{(t)}, Q^{(t)})} \\
&= \frac{\pi_k^{(t)} \prod_{r=1}^{R}[q_{kr}^{(t)} f_{r1}(x_{ir}) + (1 - q_{kr}^{(t)}) f_{r0}(x_{ir})]}{\sum_{l=1}^{K} \pi_l^{(t)} \prod_{r=1}^{R}[q_{lr}^{(t)} f_{r1}(x_{ir}) + (1 - q_{lr}^{(t)}) f_{r0}(x_{ir})]}, \qquad (6)
\end{aligned}
$$

$$
\begin{aligned}
p_{ikr1} &= E_{Y,Z|X,\pi^{(t)},Q^{(t)}}[\mathbb{I}(z_i = k) y_{ir}] = Pr(z_i = k, y_{ir} = 1|X_i, \pi^{(t)}, Q^{(t)}) \\
&= Pr(y_{ir} = 1|z_i = k, X_i, \pi^{(t)}, Q^{(t)}) \times Pr(z_i = k|X_i, \pi^{(t)}, Q^{(t)}) \\
&= \frac{p(y_{ir} = 1, x_{ir}|z_i = k, \pi^{(t)}, Q^{(t)})}{p(x_{ir}|z_i = k, \pi^{(t)}, Q^{(t)})} \times p_{ik} \\
&= \frac{q_{kr}^{(t)} f_{r1}(x_{ir})}{q_{kr}^{(t)} f_{r1}(x_{ir}) + (1 - q_{kr}^{(t)}) f_{r0}(x_{ir})} \times p_{ik}. \qquad (7)
\end{aligned}
$$

Take (6) and (7) into (5), we can obtain $Q(\pi, Q|\pi^{(t)}, Q^{(t)})$.

In the M-step, we find $\pi^{(t+1)}$ and $Q^{(t+1)}$ by maximize $Q(\pi, Q|\pi^{(t)}, Q^{(t)})$. Notice that $\sum_k \pi_k = 1$, we write the Lagrangian of the problem

$$L(\pi, Q) = Q(\pi, Q|\pi^{(t)}, Q^{(t)}) + \lambda(\sum_k \pi_k - 1).$$

By solving

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= 0, \\ \frac{\partial L}{\partial \pi_k} &= 0, \\ \frac{\partial L}{\partial q_{kr}} &= 0, \end{aligned}$$

we have

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ik}, \tag{8}$$

$$q_{kr}^{(t+1)} = \frac{\sum_{i=1}^{n} p_{ikr1}}{\sum_{i=1}^{n} p_{ik}}. \tag{9}$$

Therefore, we could iteratively use the EM algorithm and obtain the estimation for π and Q.

3.2 Model Selection: Bayesian Information Criterion

To determine the motif number of K , we use Bayesian Information Criterion(BIC). The BIC in our setting is written as

$$\begin{aligned} \mathrm{BIC}(K) &= -2\log L(\hat{\pi}, \hat{Q}; X) + (K \times R + K - 1 + N_f) \times \log n \\ &= -2\sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \hat{\pi}_k \prod_{r=1}^{R} [\hat{q}_{kr} f_{r1}(x_{ir}) + (1 - \hat{q}_{kr}) f_{r0}(x_{ir})] \right\} \\ &+ (K \times R + K - 1 + N_f) \times \log n. \end{aligned} \tag{10}$$

Here N_f is the number of parameters in f_{r0} and f_{r1}, $r = 1, \ldots, R$. We choose the K with the smallest BIC, that is

$$\hat{K} = \underset{K \geq 1}{\operatorname{argmin}}\, \mathrm{BIC}(K). \tag{11}$$

3.3 Example

In this section, we will introduce a trivial example applying CorMotif method. Under our setting, suppose

$$\begin{aligned} f_{r0}(x_{ir}) &= N(x_{ir}; 0, 1), \\ f_{r1}(x_{ir}) &= N(x_{ir}; 0, 1 + \sigma_r^2). \end{aligned}$$

We have

$$\log f_{r1}(x_{ir}) = -\frac{1}{2}\log(1 + \sigma_r^2) - \frac{x_{ir}^2}{2(1 + \sigma_r^2)} + constant.$$

Solving

$$\frac{\partial Q}{\partial \sigma_r^2} = 0,$$

we get

$$\sigma_r^{2(t+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}(x_{ir}^2 - 1)p_{ikr1}}{\sum_{i=1}^{n}\sum_{k=1}^{K}p_{ikr1}}.$$

References

[1] Joint analysis of differential gene expression in multiple studies using correlation motifs, Ji HK and Wong WH, Biostatistics, 2015, doi: 10.1093/biostatistics/kxu038

[2] Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Smyth GK, Statistical Applications in Genetics and Molecular Biology 3, 3.