Generic Cormotif

Zhiwei Ma

September 2017

1 Introduction

We introduce a new Empirical Bayes approach for jointly analyzing differential gene expression in multiple studies or under multiple experimental conditions (e.g., measured in different cell types), which allows for correlations among studies. Compared with existing methods, the new approach improves power and effect-size estimation. Model fitting procedures are implemented using expectation maximization (EM).

2 Methods

2.1 Model Outline

Our method is designed to estimate the effects of multiple units in multiple studies. Suppose there are totally n units and R studies. Let $\beta = [\beta_{jr}]_{J\times R}$ denote "effects" of interest. For instance, β_{jr} could be the difference in the mean expression of gene j at study r under two conditions.

Assume that the available data are estimates $\hat{\beta} = [\hat{\beta}_{jr}]_{J\times R}$ of the effects, and the corresponding estimated standard errors $\hat{s} = [\hat{s}_{jr}]_{J\times R}$. Let $\beta_j := (\beta_{j1}, \ldots, \beta_{jr})'$, where $j = 1, 2, \ldots, n$. Similarly for $\hat{\beta}_j$ and $\hat{s}_j$,

Similar to the original *Cormotif* [**?**], we assume that all units fall into K different classes. Besides, we have the following assumptions, which distinguish our model from existing ones:

Assumption 1 *Each units j is randomly and independently assigned to a class label z_j according to probability $\pi = (\pi_1, \ldots, \pi_K)$. Here $\pi_k = P(z_j = k)$ is the prior probability that a unit belongs to class k. We have $\sum_k \pi_k = 1$.*

Assumption 2 *Given unit's class label $z_j = k$, the effects $\beta_{j1}, \beta_{j2}, \ldots, \beta_{jR}$ are independent from unimodal distribution g_{kr}, that is $p(\beta_j | z_j = k, \hat{s}) = \prod_{r=1}^{R} p(\beta_{jr} | z_j = k, \hat{s}) = \prod_{r=1}^{R} g_{kr}(\beta_{jr}; \hat{s})$*

Here we describe the simplest version of our model. First we assume the effect β_{jr} is independent of their standard errors $\hat{s}_{jr}$. Then by Assumption 1 and 2, we have

$$p(\beta | \pi, g, \hat{s}) = \prod_{j=1}^{n} p(\beta_j | \pi, g, \hat{s}) = \prod_{j=1}^{n} \left[\sum_{k=1}^{K} \pi_k \prod_{r=1}^{R} g_{kr}(\beta_{jr}) \right] \tag{1}$$

where g represents $\{g_{kr} | k = 1, \ldots, K,\ r = 1, \ldots, R\}$. A simple way to implement the unimodal assumption (UA) is to assume that g_{kr} is a mixture of a point mass at 0 and a mixture of *zero-mean* normal distribution [**?**]:

$$g_{kr}(\cdot) = w_0^{kr} \delta_0(\cdot) + \sum_{l=1}^{L} w_l^{kr} N(\cdot; 0, \sigma_l^2), \tag{2}$$

where $\delta_0(\cdot)$ denotes a point mass on 0, and $N(\cdot; \mu, \sigma^2)$ denotes the density of Normal distribution with mean μ and variance σ^2. Here we assume $\sigma_1, \sigma_2, \ldots, \sigma_L$ are known and fixed positive numbers forming a wide and dense grid, which are built via a data-driven approach.

For likelihood $p(\hat{\beta}|\beta, \hat{s})$, we assume a Normal approximation:

$$p(\hat{\beta}|\beta, \hat{s}) = \prod_{j=1}^{n} p(\hat{\beta}_j|\beta_j, \hat{s}) = \prod_{j=1}^{n}\prod_{r=1}^{R} N(\hat{\beta}_{jr}; \beta_{jr}, \hat{s}_{jr}^2). \tag{3}$$

Together, (**??**)-(**??**) imply that

$$\begin{aligned} p(\hat{\beta}|\pi, g, \hat{s}) &= \prod_{j=1}^{n}\left[\sum_{k=1}^{K}\pi_k \prod_{r=1}^{R}(g_{kr} * N_{jr})(\hat{\beta}_{jr})\right] \\ &= \prod_{j=1}^{n}\left\{\sum_{k=1}^{K}\pi_k \prod_{r=1}^{R}\left[\sum_{l=0}^{L} w_l^{kr} N(\hat{\beta}_{jr}; 0, \sigma_l^2 + \hat{s}_{jr}^2)\right]\right\}. \end{aligned} \tag{4}$$

Here N_{jr} denotes $N(\cdot; 0, \hat{s}_{jr}^2)$ and $*$ means the convolution of two functions. We define $\sigma_0 := 0$.

Model (**??**) is the extension of original *Cormotif*: set $L = 1$ and assume for any fixed r, $\hat{s}_{jr}^2$s are identical for all j, (**??**) is just Cormotif under Normal distribution. Similar to Cormotif, our model can capture the correlation among multiple studies. To see this, consider the likelihood for unit j. Based on our model, $p(\hat{\beta}_j|\pi, g, \hat{s}) = \sum_{k=1}^{K}\pi_k \prod_{r=1}^{R}(g_{kr} * N_{jr})(\hat{\beta}_{jr})$. The distribution for unit j under study r is $p(\hat{\beta}_{jr}|\pi, g, \hat{s}) = \sum_{k=1}^{K}\pi_k (g_{kr} * N_{jr})(\hat{\beta}_{jr})$. It is clear that $p(\hat{\beta}_j|\pi, g, \hat{s}) \neq \prod_{r=1}^{R} p(\hat{\beta}_{jr}|\pi, g, \hat{s})$, so different studies are dependent.

When setting $K = 1$, our model is equivalent to applying `ash` [**?**] to each study separately. The advantage of our model is that it allows correlations among studies. This advantage leads to higher accuracy for effect estimation in multiple studies.

Another advantage of our generic model is that we could estimate the posterior distribution for effects β_{jr}, that is $p(\beta_{jr}|\hat{\beta}, \hat{s}, \pi, g)$.

2.2 Fitting the model

Our method involves three steps:

1. For a given K, estimate the parameters π and g by maximizing the likelihood $L(\pi, g)$, given by (**??**), denote as $\hat{\pi}$ and $\hat{g}$.

2. Choose the best K by minimizing the Bayesian Information Criterion (BIC).

3. Compute, for each j and r, the posterior distribution $p(\beta_{jr}|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g})$.

All the details can be found in Section (**??**). For Step 1 we apply an EM algorithm to solve it. Step 2 is straightforward. The conditional distributions $p(\beta_{jr}|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g})$ in Step 3 are analytically available, each a mixture of a point mass on zero and L normal distribution.

2.3 Local False Sign Rate

To measure "significance" of an effect β_{jr} we use the local false sign rate (*lfsr*), which is defined as:

$$lfsr_{jr} := \min\{p(\beta_{jr} \geq 0|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g}), p(\beta_{jr} \leq 0|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g})\} \tag{5}$$

Intuitively, $lfsr_{jr}$ is the probability that we would get the sign of effect β_{jr} incorrect if we were to use our best guess of the sign. Therefore, a small *lfsr* indicates high confidence in determining the sign of an effect. Notice that *lfsr* is more conservative than the local false discovery rate (*lfdr*), since we can infer $lfsr_{jr} \geq lfdr_{jr}$ from the definition.

3 Detailed Method

3.1 Embellishments

3.2 Implementation Details

3.2.1 Optimization

This section We presents the EM algorithm used to estimate both π and g. First compute the log likelihood function for $\hat{\beta}$ and group label $z = (z_1, \ldots, z_n)'$:

$$\begin{aligned}
& \log p(\hat{\beta}, z|\pi, g, \hat{s}) = \sum_{j=1}^{n} \log p(\hat{\beta}_j|z_j, \pi, g, \hat{s}) + \sum_{j=1}^{n} \log p(z_j|\pi) \\
= & \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log p(\hat{\beta}_j|z_j = k, \pi, g, \hat{s}) + \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log p(z_j = k|\pi) \\
= & \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \sum_{r=1}^{R} \log \left[(g_{kr} * N_{jr})(\hat{\beta}_{jr}) \right] + \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log \pi_k. \qquad (6)
\end{aligned}$$

Here z is a latent variable. The EM algorithm seeks to find the MLE of the marginal likelihood (**??**) by iteratively applying the E-step and the M-step.

In the E-step, one evaluates the Q-function $Q(\pi, g|\pi^{(t)}, g^{(t)})$, here $(\pi^{(t)}, g^{(t)})$ is the current estimation. We have

$$\begin{aligned}
& Q(\pi, g|\pi^{(t)}, g^{(t)}) = E_{z|\hat{\beta},\hat{s},\pi^{(t)},g^{(t)}} \left[\log p(\hat{\beta}, z|\pi, g, \hat{s}) \right] \\
= & \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R} p_{jk} \log \left[(g_{kr} * N_{jr})(\hat{\beta}_{jr}) \right] + \sum_{j=1}^{n}\sum_{k=1}^{K} p_{jk} \log \pi_k, \qquad (7)
\end{aligned}$$

where we denote

$$\begin{aligned} p_{jk} &= E_{z|\hat{\beta},\hat{s},\pi^{(t)},g^{(t)}}[\mathbb{I}(z_j = k)] = p(z_j = k|\hat{\beta}_j, \hat{s}, \pi^{(t)}, g^{(t)}) \\ &= \frac{p(\hat{\beta}_j, z_j = k|\hat{s}, \pi^{(t)}, g^{(t)})}{p(\hat{\beta}_j|\hat{s}, \pi^{(t)}, g^{(t)})} \\ &= \frac{\pi_k^{(t)} \prod_{r=1}^{R} (g_{kr}^{(t)} * N_{jr})(\hat{\beta}_{jr})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \prod_{r=1}^{R} (g_{k'r}^{(t)} * N_{jr})(\hat{\beta}_{jr})} \end{aligned} \tag{8}$$

In the M-step, one finds π and g that maximize the Q-function $Q(\pi, g|\pi^{(t)}, g^{(t)})$, and denote them as $\pi^{(t+1)}$ and $g^{(t+1)}$, that is

$$(\pi^{(t+1)}, g^{(t+1)}) = \underset{(\pi,g)}{\operatorname{argmax}}\, Q(\pi, g|\pi^{(t)}, g^{(t)}). \tag{9}$$

For $\pi^{(t+1)}$, we could optimize it from (**??**) directly and get

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} p_{jk}. \tag{10}$$

Notice in (**??**), we could separately optimize g_{kr} for fixed k and r, that is

$$g_{kr}^{(t+1)} = \underset{g_{kr}}{\operatorname{argmax}} \sum_{j=1}^{n} p_{jk} \log \left[(g_{kr} * N_{jr})(\hat{\beta}_{jr}) \right]. \tag{11}$$

Optimizing (**??**) is a convex problem, which we solve using an EM algorithm, accelerated using R package `SQUAREM` .

3.2.2 Model Selection: Bayesian Information Criterion

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

$$\begin{aligned} \mathrm{BIC}(K) &= -2\log p(\hat{\beta}|\pi, g, \hat{s}) + (K \times R \times L + K - 1) \times \log n \\ &= -2\sum_{j=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \prod_{r=1}^{R} (g_{kr} * N_{jr})(\hat{\beta}_{jr}) \right] \\ &+ (K \times R \times L + K - 1) \times \log n. \end{aligned} \tag{12}$$

Here $K-1$ is the number of parameters for π, $K \times R \times L$ is the number of parameters involved in g and n is the unit number. We choose the K with the smallest BIC, that is

$$\hat{K} = \underset{K \geq 1}{\operatorname{argmin}}\, \mathrm{BIC}(K). \tag{13}$$

3.2.3 Posterior distribution

For a more general case, suppose

$$g_{kr}(\cdot; w) = \sum_{l=0}^{L} w_l^{kr} f_l(\cdot), \tag{14}$$

Denote

$$\tilde{f}_l(\hat{\beta}_{jr}) := \int f_l(\beta_{jr}) p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr}) d\beta_{jr}, \tag{15}$$

as the likelihood of $\hat{\beta}_{jr}$ if the prior of β_{jr} follows $f_l(\beta_{jr})$.

By Bayes theorem,

$$\begin{aligned}
& p(\beta_{jr}|\hat{\beta}, \hat{s}, \pi, g) = p(\beta_{jr}|\hat{\beta}_j, \hat{s}_j, \pi, g) \\
= & \sum_{k=1}^{K} p(\beta_{jr}|\hat{\beta}_j, \hat{s}_j, \pi, g, z_j = k) p(z_j = k|\hat{\beta}_j, \hat{s}_j, \pi, g) \\
= & \sum_{k=1}^{K} p(\beta_{jr}|\hat{\beta}_{jr}, \hat{s}_j, \pi, g, z_j = k) p_{jk} \\
= & \sum_{k=1}^{K} \frac{p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr}) g_{kr}(\beta_{jr})}{\int p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr}) g_{kr}(\beta_{jr}) d\beta_{jr}} \times p_{jk} \\
= & \sum_{k=1}^{K} \frac{\sum_{l=0}^{L} w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr}) h_l(\beta_{jr})}{\sum_{l'=0}^{L} w_{l'}^{kr} \tilde{f}_{l'}(\hat{\beta}_{jr})} \times p_{jk} \\
= & \sum_{l=0}^{L} \theta_{ljr} h_l(\beta_{jr}). \qquad (16)
\end{aligned}$$

Here the posterior mixture component h_l is the posterior on β_{jr} that would be obtained using prior $f_l(\beta_{jr})$ and likelihood $p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr})$, and the mixture weights θ_{ljr} is

$$\theta_{ljr} = \sum_{k=1}^{K} \frac{w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr}) p_{jk}}{\sum_{l'=0}^{L} \eta_{l'r} \tilde{f}_{l'}(\hat{\beta}_{jr})}. \qquad (17)$$

References

[1] Wei, Y. and Ji, H. (2014). Joint analysis of differential gene expression in multiple studies using correlation motif. *Biostatistics*, doi:10.1093/biostatistics/ kxu038.

[2] Stephens, M. (2016). False discovery rate: A new deal. *Biostatistics* 18 (2): 275-294

[3] Urbut, S.M., Wang, G. and Stephens, M. (2017). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*, doi:10.1101/096552.