Master's Paper of the Department of Statistics, the University of Chicago
(Internal document only, not for circulation)

Joint analysis of differential gene expression in multiple studies using Bayesian method

Zhiwei Ma

Advisor(s): Matthew Stephens

Approved _____

Date _____

February 15, 2018

Abstract

We introduce a new Empirical Bayes approach for jointly analyzing differential gene expression in multiple studies or under multiple experimental conditions. Our method searches for a small number of latent prior distribution vectors to capture the dependence among multiple studies. Compared with independent analysis of each study, our joint analysis approach improves both power to detect significant effects and effect-size estimation. Model fitting procedures are implemented using expectation maximization (EM) algorithm.

Contents

1 Introduction

Detecting differentially expressed genes across multiple conditions or studies is an important task in the analysis of gene expression data. Examples including detecting the genes which are specific to MTB (Mycobacterium tuberculosis) infection [Blischak et al., 2015]; or studying vertebrate sonic hedgehog (SHH) pathway [Wei et al., 2015]; or identifying eQTLs in multiple cell-type or tissues [Urbut et al., 2017].

The simplest and the most common solution to this problem is to analyze data in different studies one at a time, and then identify the overlap of significant results in multiple studies. Methods for single study including *limma* [Smyth, 2004] or *DESeq* [Anders and Huber, 2010] to estimate effect sizes, followed by methods like *qvalue* [Storey, 2003] or *ash* [Stephens, 2017] to estimate false discovery rates. However when data from multiple related gene expression studies are available, separately analyzing each study is not ideal as it may fail to detect important genes with consistent but relatively weak differential signals in multiple studies.

To address the deficiencies of study-by-study analyses, several methods have been developed for joint analysis of effects under multiple studies. *Cormotif* [Wei et al., 2015] used a small number of latent probability vectors called "correlation motifs" to model the major correlation patterns among the studies. This approach can handle all possible study-specific differential patterns in a computationally-tractable way. However, its simplicity substantially reduce flexibility: firstly, *Cormotif* adopts a two-stage design and uses moderated t-statistics produced by *limma* [Smyth, 2004]. The restrictive assumptions to the estimated effect sizes greatly weakens the adaptability in different situations; secondly, *Cormotif* focus only on testing for the significant effects in each condition, but not on estimating effect sizes, which might be interesting in assessing heterogeneity of effects among conditions.

By applying an Empirical Bayesian method, *mash* [Urbut et al., 2017] builds a hierarchical model for jointly analyzing effect sizes. *mash* is an extension of *ash* by setting the priors of effect sizes as a mixture of multivariate normal distribution with both data-driven and canonical covariance matrices. However, *mash* fails to include some common situations, such as when two studies have both positive or negative effects at the same time for some genes. In addition, simple normal distribution may not be able to capture the prior distributions for the effect sizes well.

Here we introduce a more flexible method that combines the attractive features of existing methods. Specifically, we extend the models in *Cormotif* [Wei et al., 2015] by allowing general unimodal distributions for the effect size distributions in each component. These general unimodal distributions are modeled using finite mixtures as in *ash* [Stephens, 2017]. Compared with *Cormotif*, our approach can handle both effect size estimation and testing for significant effects by computing the posterior distributions of the effect sizes. Compared with *mash*, our model includes the situation when two studies are conditional independent and we give more freedom for choosing the prior distribution. However, unlike *mash* we do not allow for correlations in effect sizes across studies within each mixture component.

Another feature of our approach is that our method is adaptive to both the amount of signal in the data and the measurement precision of each observation.

Our approach is a general method and can be applied to various problems in analyzing differential gene expression in multiple studies. We provide implementation in an R package, `miximash`, available at `https://github.com/stephenslab/miximash`. Code and instructions for reproducing analyses and figures in this paper are at `https://github.com/stephenslab/cormotif`.

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 (log) expression of gene j under study r in 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 $j = 1, 2, \ldots, n$. Similar for $\hat{\beta}_j$ and $\hat{s}_j$.

Similar to the original *Cormotif* [Wei et al., 2015], we assume that all units fall into K different classes to reduce the complexity of parameter space. Specifically, we have the following key assumptions:

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, more embellishments can be found in Detailed Method. 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 $\{g_{kr}|k = 1, \ldots, K,\ r = 1, \ldots, R\}$, the set of unimodal distributions. 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 [Stephens, 2017]:

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

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. By using a sufficiently large L the finite mixture (2) can approximate any scale mixture of zero-centered normals to arbitrary accuracy.

$\sigma_1, \sigma_2, \ldots, \sigma_L$ are built via a data-driven approach–see Implementation Details. The mixture proportions $w^{kr} = (w_0^{kr}, \ldots, w_L^{kr})$ for $k = 1, \ldots, K$, $r = 1, \ldots, R$ and $\pi = (\pi_1, \ldots, \pi_K)$ are the parameters to be estimated.

For likelihood $p(\hat{\beta}|\beta, \hat{s})$, we assume a Normal approximation [Wakefield, 2009]:

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

Together, (1)-(3) 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 $\sigma_0 := 0$ in the formula above.

Actually, model (4) is the extension of the original *Cormotif*: set $L = 1$ and assume for any fixed r, $\hat{s}_{jr}^2$s are identical for all j, (4) is just *Cormotif* model under Normal distribution. The difference here is that by using a larger L we obtain more flexible models than *Cormotif*. Similar to *Cormotif*, our model can capture the dependence 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. Compared with *Cormotif*, another advantage of our generic model is its focus on both testing for significant effects and estimating the effect sizes, which are realized by computing the posterior distribution of effect β_{jr}, $p(\beta_{jr}|\hat{\beta}, \hat{s}, \pi, g)$.

When setting $K = 1$, our model is equivalent to applying *ash* [Stephens, 2017] to each study separately. Compared with *ash*, our model allows dependence among studies. As we show in Simulation, this advantage improves both effect estimation and detection of significant effects in multiple studies.

2.2 Fitting the model

Our method involves two steps:

1. For a large enough K, estimate the parameters π and g by maximizing the likelihood $L(\pi, g)$, given by (4), denote as $\hat{\pi}$ and $\hat{g}$.

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

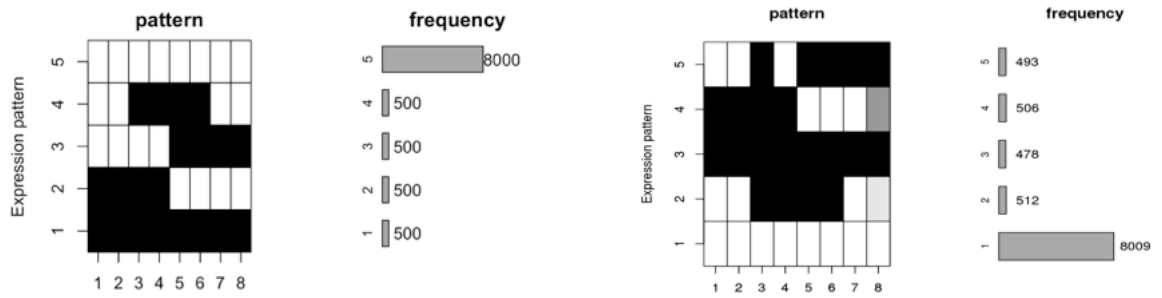
All the details can be found in Implementation Details. For Step 1, we have totally $K \times R \times L + K - 1$ parameters to estimate. We apply an EM algorithm to solve it (Alternatively interior point (IP) method [Koenker and Mizera, 2014] can be applied here as well). The conditional distributions $p(\beta_{jr}|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g})$ in Step 2 are analytically available, each a mixture of a point mass on zero and L normal distribution. With the posterior distributions for the effect, one can estimate the effect size by $E(\beta_{jr}|\hat{\beta}, \hat{s}, \hat{\pi}, \hat{g})$ and test for non-zero effects.

2.3 Local False Sign Rate

To measure "significance" of an effect β_{jr} we use the local false sign rate (*lfsr*) [Stephens, 2017], 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 $lfsr_{jr} \geq lfdr_{jr}$.



(a) True pattern (b) Estimated pattern

Figure 1: Patterns learned in Simulation 1 (constant precision, $\hat{s}_{jr} = 1$).

3 Simulation

3.1 Estimation of patterns

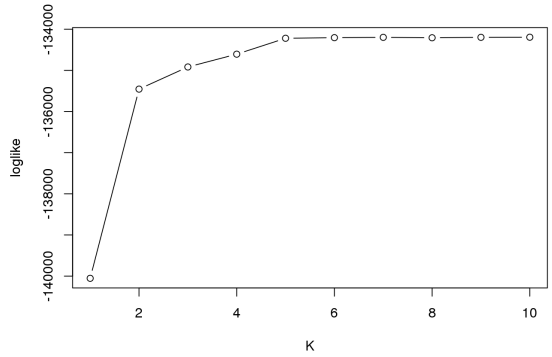
When set $K = 1$ in our model, our model is equivalent to applying *ash* to each study separately. For comparison, we would like to check the performance for $K = 1$ and $K \geq 2$ separately. The increase in the performance can be viewed as the gain by considering the correlation among studies.

In simulation 1, we set unit number $n = 10000$, and study number $R = 8$. Assume every observation has the same standard error, $s_{jr} = 1$. That is, $\hat{\beta}_{jr}|\beta_{jr} \sim N(\beta_{jr}; 0, 1)$. The 10000 units come from 5 patterns ($K = 5$): 8000 units have zero effects in all four studies, that is $\beta_{jr} = 0$, for all $r = 1, 2, 3, 4, 5, 6, 7, 8$; 500 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 1, 2, 3, 4$ and $\beta_{jr} = 0$ for $r = 5, 6, 7, 8$; 500 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 5, 6, 7, 8$ and $\beta_{jr} = 0$ for $r = 1, 2, 3, 4$; 500 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 3, 4, 5, 6$ and $\beta_{jr} = 0$ for $r = 1, 2, 7, 8$; 500 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 1, 2, 3, 4, 5, 6, 7, 8$.

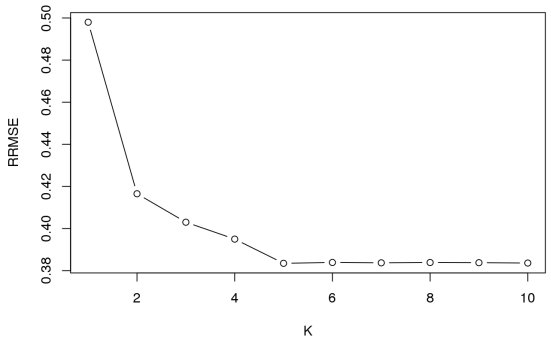
Figure 1a shows the true patterns and its frequencies. In the left plot, the value in k-th row and r-th column is $p(\beta_{jr} \neq 0|z_j = k, g) = 1 - w_0^{kr}$. The right plot shows the frequencies of each pattern. For estimation, we set $K = 5$ and the estimated patterns are displayed in Figure 1b. The estimated pattern are very similar to the true underlying differential patterns in Figure 1a.

3.2 Improved effect size estimates

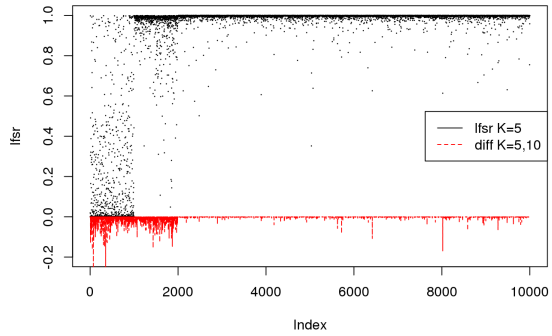
We fit our model for K increasing from 1 to 10. Since our model is nested for K (for larger K, we can always set some π_k to be zero and get the same log likelihood function), the log likelihood function for larger K is no less than smaller ones. Figure 2a shows the log likelihood function in (4) for different K. The log likelihood function has a non-decreasing



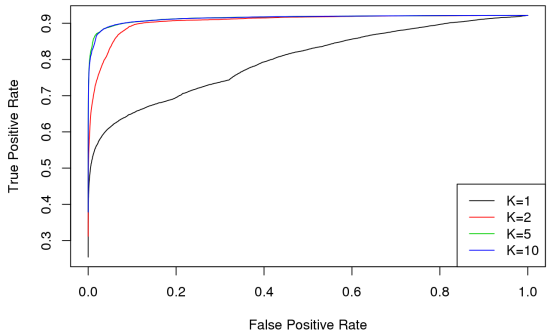
(a) Log likelihood function for different K



(b) Accuracy of effect estimates (RRMSE)



(c) *lfsr* for the first study



(d) Detection of non-null effects (ROC curves)

Figure 2: Results for Simulation 1

trend. From $K = 1$ to 2, the log likelihood function has the most significant increase. And for $K \geq 5$, the log likelihood function remains approximately constant. The increasing trend shows that our hierarchical model fits better than applying `ash` independently across studies (K=1). Then trend of log likelihood values after $K = 5$ is due to that only five $\hat{\pi}_k$ get significant non-zeros values, while all others are estimated to be almost zeros.

Figure 2b compares the accuracy of effect size estimates by relative root mean squared error (RRMSE). Denote estimates for β_{jr} is $\tilde{\beta}_{jr}$ by our method. The RRMSE for estimates $\tilde{\beta}_{jr}$ is computed as

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{NR}\sum_{j=1}^{N}\sum_{r=1}^{R}(\tilde{\beta}_{jr} - \beta_{jr})}}{\sqrt{\frac{1}{NR}\sum_{j=1}^{N}\sum_{r=1}^{R}(\hat{\beta}_{jr} - \beta_{jr})}}, \tag{6}$$

which is the RMSE of estimates $\tilde{\beta}_{jr}$ divided by the RMSE achieved by simply using the original observed estimates $\hat{\beta}_{jr}$. Therefore, RRMSE less than 1 indicates that the estimates is more accurate than original observation $\hat{\beta}_{jr}$. From Figure 2b, we can find that even $K = 1$ gives approximate 0.5 RRMSE. In addition, the RRMSE has a decline trend as K increases and reaches approximate constance for $K \geq 5$. Compared with *ash* ($K = 1$), our method has a substantial improvement in accuracy for large enough K.

3.3 Improved detection of significant effects

In addition to effect estimates, our method also provides a measure of significance for each effect by *lfsr*defined in (5). We used the same simulation as above to illustrate the gains in power to detect significant effects coming from our method. Figure 2c shows the *lfsr* values computed by our method. The black dot represent the estimated *lfsr* in the first study for $K = 5$. Notice that in the first study, we only set the first 1000 units to own significant effects ($\beta_{jr} \sim N(0, 4^2)$) out of totally 10000 units. For the first 1000 units, the estimated *lfsr* tend to gather to 0, which means we are confident that they are nonzero. The following units have *lfsr* gathering at 1, which means they have zero effects. The red line shows the negative absolutely values of the differences of *lfsr* between $K = 5$ and $K = 10$. Except for the first 2000 units, the differences are very small in each unit, and almost all differences are less than 0.1. This again illustrate that large enough K will give similar results as the exact one.

Figure 2d shows the trade-off between false positive and true positive discoveries for different K (ROC curves). The True Positive Rate and False Positive Rate are computed at any given threshold γ as

$$\text{True Positive Rate} := \frac{|CS \cap S|}{|T|} \tag{7}$$

$$\text{False Positive Rate} := \frac{|N \cap S|}{|N|} \tag{8}$$

where S is the set of significant effects at threshold γ, CS is the set of correctly signed effects, T is the set of true (non-zeros) effects and N is the set of null (zero) effects:

$$
\begin{aligned}
S &:= \{(j,r)|lfsr_{jr} \leq \gamma\} && (9)\\
CS &:= \{(j,r)|E(\beta_{jr}|D) \times \beta_{jr} > 0\} && (10)\\
N &:= \{(j,r)|\beta_{jr} = 0\} && (11)\\
T &:= \{(j,r)|\beta_{jr} \neq 0\} && (12)
\end{aligned}
$$

A true positive needs to be correctly signed, not only significant. The performance precisely mirrors the RRMSE results: our method outperforms *ash* and the performance is similar for large enough K.

Using a similar approach, we performed simulations 2, which involved different study number and differential expression patterns, see Appendix Figure 3. The conclusion is similar to simulation 1.

4 Detailed Method

4.1 Embellishments

We can extend g_{kr} in (2) to other symmetric or even asymmetric unimodal distributions. For a more general case, suppose

$$g_{kr}(\cdot; w) = \sum_{l=0}^{L} w_l^{kr} f_l(\cdot), \tag{13}$$

where f_0 is a point mass at 0, and f_l $(l = 1, \ldots, L)$ are component distribution with one of the following forms:

(i) $f_l(\cdot) = N(\cdot; 0, \sigma_l^2)$,

(ii) $f_l(\cdot) = U[\cdot; -a_l, a_l]$,

(iii) $f_l(\cdot) = U[\cdot; -a_l, 0]$ and/or $U[\cdot; 0, a_l]$,

where $U[\cdot; a, b]$ denotes the density of a uniform distribution on $[a, b]$. Actually, with dense enough a_l, (iii) can approximate any unimodal distribution about 0. To better approximate the unimodal distribution, larger K should be better. However in practice, a moderate K can provide reasonable performance. The detailed method for estimating σ_l^2 and a_l form data can be found in the Implementation Details in [Stephens, 2017].

4.2 Implementation Details

4.2.1 Likelihood

Using the prior $\beta_{jr}|z_j = k \sim \sum_{l=0}^{L} w_l^{kr} f_l(\beta_{jr})$ given by (13) and the normal likelihood in (3), integrating over β_{jr} yields

$$p(\hat{\beta}_{jr}|z_j = k, \pi, g, \hat{s}) = \sum_{l=0}^{L} w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr}), \tag{14}$$

where

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

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

These convolutions are straightforward to evaluate when f_l is a normal or uniform density. Specifically,

$$\tilde{f}_l(\hat{\beta}_{jr}) = \begin{cases} N(\hat{\beta}_{jr}; 0, \hat{s}_{jr}^2 + \sigma_l^2) & \text{if } f_l(\cdot) = N(\cdot; 0, \sigma_l^2), \\ \frac{\Phi((\hat{\beta}_{jr}-a_l)/\hat{s}_{jr}) - \Phi((\hat{\beta}_{jr}-b_l)/\hat{s}_{jr})}{b_l - a_l} & \text{if } f_l(\cdot) = U(\cdot; a_l, b_l), \end{cases} \tag{16}$$

where Φ denotes the cumulative distribution function (c.d.f.) of the standard normal distribution.

4.2.2 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 $z = (z_1, \ldots, z_n)'$:

$$\begin{aligned} & \log p(\hat{\beta}, z|\pi, g, \hat{s}) = \sum_{j=1}^{n} \log p(\hat{\beta}_j|z_j, \pi, g, \hat{s}) + \sum_{j=1}^{n} \log p(z_j|\pi) \\ = & \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log p(\hat{\beta}_j|z_j = k, \pi, g, \hat{s}) + \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log p(z_j = k|\pi) \\ = & \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \sum_{r=1}^{R} \log \left[p(\hat{\beta}_{jr}|z_j = k, \pi, g, \hat{s}) \right] + \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbb{I}(z_j = k) \log \pi_k. \end{aligned} \tag{17}$$

Here z is a latent variable. The EM algorithm seeks to find the MLE of the marginal likelihood (4) 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

$$\begin{aligned} & Q(\pi, g|\pi^{(t)}, g^{(t)}) = E_{z|\hat{\beta},\hat{s},\pi^{(t)},g^{(t)}} \left[\log p(\hat{\beta}, z|\pi, g, \hat{s}) \right] \\ = & \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R} p_{jk} \log \left[p(\hat{\beta}_{jr}|z_j = k, \pi, g, \hat{s}) \right] + \sum_{j=1}^{n}\sum_{k=1}^{K} p_{jk} \log \pi_k, \end{aligned} \tag{18}$$

where we denote

$$\begin{aligned} p_{jk} &= E_{z|\hat{\beta},\hat{s},\pi^{(t)},g^{(t)}}[\mathbb{I}(z_j = k)] = p(z_j = k|\hat{\beta}_j, \hat{s}, \pi^{(t)}, g^{(t)}) \\ &= \frac{p(\hat{\beta}_j, z_j = k|\hat{s}, \pi^{(t)}, g^{(t)})}{p(\hat{\beta}_j|\hat{s}, \pi^{(t)}, g^{(t)})} \\ &= \frac{\pi_k^{(t)} \prod_{r=1}^{R} p(\hat{\beta}_{jr}|z_j = k, \pi^{(t)}, g^{(t)}, \hat{s})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \prod_{r=1}^{R} p(\hat{\beta}_{jr}|z_j = k, \pi^{(t)}, g^{(t)}, \hat{s})} \end{aligned} \tag{19}$$

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

$$(\pi^{(t+1)}, g^{(t+1)}) = \underset{(\pi,g)}{\operatorname{argmax}}\, Q(\pi, g|\pi^{(t)}, g^{(t)}). \tag{20}$$

For $\pi^{(t+1)}$, we could optimize it from (18) directly and get

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} p_{jk}. \tag{21}$$

Notice in (18), we could separately optimize g_{kr} for fixed k and r, that is

$$\begin{aligned} g_{kr}^{(t+1)} &= \underset{g_{kr}}{\operatorname{argmax}} \sum_{j=1}^{n} p_{jk} \log\left[p(\hat{\beta}_{jr}|z_j = k, \pi, g, \hat{s})\right] \\ &= \underset{w^{kr}}{\operatorname{argmax}} \sum_{j=1}^{n} p_{jk} \log\left[\sum_{l=0}^{L} w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr})\right]. \end{aligned} \tag{22}$$

Optimizing (22) is a convex problem, which we solve using an EM algorithm, accelerated using R package `SQUAREM`. A simple interior point (IP) methods could also be applied here using R package `REBayes`.

4.2.3 Posterior distribution

By Bayes theorem,

$$\begin{aligned} & p(\beta_{jr}|\hat{\beta}, \hat{s}, \pi, g) = p(\beta_{jr}|\hat{\beta}_j, \hat{s}_j, \pi, g) \\ = & \sum_{k=1}^{K} p(\beta_{jr}|\hat{\beta}_j, \hat{s}_j, \pi, g, z_j = k) p(z_j = k|\hat{\beta}_j, \hat{s}_j, \pi, g) \\ = & \sum_{k=1}^{K} p(\beta_{jr}|\hat{\beta}_{jr}, \hat{s}_j, \pi, g, z_j = k) p_{jk} \\ = & \sum_{k=1}^{K} \frac{p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr}) g_{kr}(\beta_{jr})}{\int p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr}) g_{kr}(\beta_{jr}) d\beta_{jr}} \times p_{jk} \\ = & \sum_{k=1}^{K} \frac{\sum_{l=0}^{L} w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr}) h_l(\beta_{jr})}{\sum_{l'=0}^{L} w_{l'}^{kr} \tilde{f}_{l'}(\hat{\beta}_{jr})} \times p_{jk} \\ = & \sum_{l=0}^{L} \theta_{ljr} h_l(\beta_{jr}). \end{aligned} \tag{23}$$

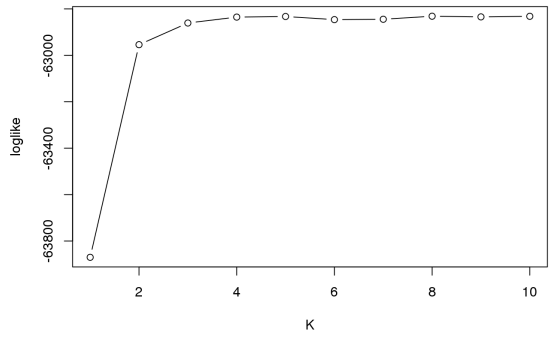
Here the posterior mixture component h_l is the posterior on β_{jr} that would be obtained using prior $f_l(\beta_{jr})$ and likelihood $p(\hat{\beta}_{jr}|\beta_{jr}, \hat{s}_{jr})$, and the mixture weights θ_{ljr} is

$$\theta_{ljr} = \sum_{k=1}^{K} \frac{w_l^{kr} \tilde{f}_l(\hat{\beta}_{jr}) p_{jk}}{\sum_{l'=0}^{L} w_{l'}^{kr} \tilde{f}_{l'}(\hat{\beta}_{jr})}. \tag{24}$$

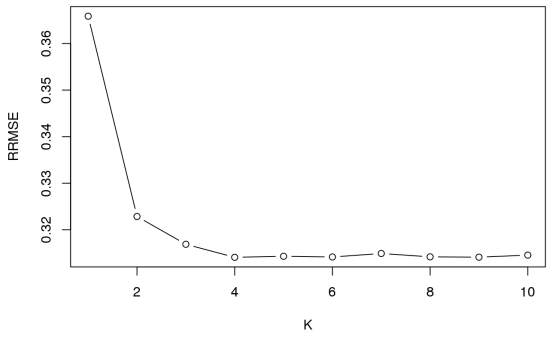
5 Conclusions

In this paper, we introduce an Empirical Bayesian method for jointly analyzing differential gene expression in multiple studies. To capture the dependence among different studies, we introduce a small number of latent patterns for prior distributions of effect sizes. Our approach focus on both estimation of effect sizes and testing for significant effects. Simulation results show that our method outperforms the single study approach *ash* in both effect-size estimation and detecting significant effects. In addition, in simple settings our approach can accurately estimate the true patterns in the data.

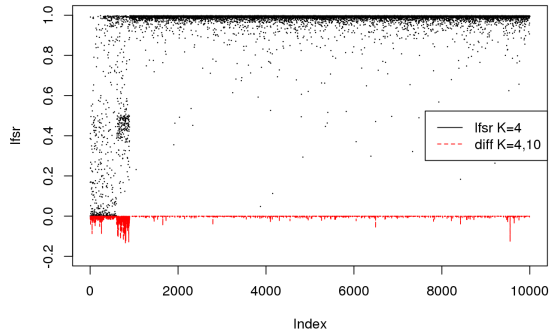
A Appendix



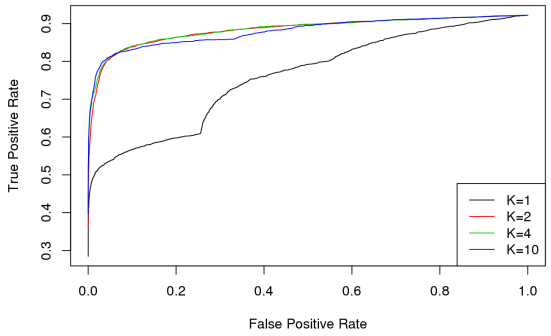
(a) Log likelihood function for different K



(b) Accuracy of effect estimates (RRMSE)



(c) *lfsr* for the first study



(d) Detection of non-null effects (ROC curves)

Figure 3: Results for Simulation 2 (constant precision, $\hat{s}_{jr} = 1$).

Note: In simulation 2, we set unit number $n = 10000$, and study number $R = 4$. Assume every observation has the same standard error $\hat{s}_{jr} = 1$. That is, $\hat{\beta}_{jr}|\beta_{jr} \sim N(\beta_{jr}; 0, 1)$. The 10000 units come from 4 patterns ($K = 4$): 9100 units have zero effects in all four studies, that is $\beta_{jr} = 0$, for $r = 1, 2, 3, 4$; 300 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 1, 2$ and $\beta_{jr} = 0$ for $r = 3, 4$; 300 units have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 3, 4$ and $\beta_{jr} = 0$ for $r = 1, 2$; 300 unitss have effect $\beta_{jr} \sim N(0, 4^2)$ for $r = 1, 2, 3, 4$.

References

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

[Blischak et al., 2015] Blischak, J. D., Tailleux, L., Mitrano, A., Barreiro, L. B., and Gilad, Y. (2015). Mycobacterial infection induces a specific human innate immune response. *Scientific Reports*, 5(1):16882.

[Koenker and Mizera, 2014] Koenker, R. and Mizera, I. (2014). Convex Optimization in R. *JSS Journal of Statistical Software*, 60(5).

[Smyth, 2004] Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

[Stephens, 2017] Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18(2):275–294.

[Storey, 2003] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q -value. *The Annals of Statistics*, 31(6):2013–2035.

[Urbut et al., 2017] Urbut, S. M., Wang, G., and Stephens, M. (2017). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*, page 096552.

[Wakefield, 2009] Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with ¡i¿P¡/i¿ -values. *Genetic Epidemiology*, 33(1):79–86.

[Wei et al., 2015] Wei, Y., Tenzen, T., and Ji, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, 16(1):31–46.