Heterogeneous Error Model (HEM) for Identification of Differentially Expressed Genes Under Multiple Conditions

HyungJun Cho and Jae K. Lee

May 3, 2016

Contents

1 Introduction

The *HEM* package fits an error model with heterogenous experimental and/or biological error variances for analyzing microarray data. It enables identification of differentially-expressed genes by correctly and simultaneously estimating a large number of heterogeneous variance parameters. Some features of the *HEM* package are

- Bayesian error modeling with one or two layers of error,
- simultaneous estimation of all parameters by Markov Chain Monte Carlo (MCMC),
- estimation of decomposed experimental and biological variances,
- consideration of heterogeneity of error variances for all genes and under multiple biological conditions, and
- parametric or nonparametric Empirical Bayes (EB) prior specification for variances.

To use the *hem* package in R, install the distributed *HEM* package and call it as follows

```
> library(HEM)
```

The *HEM* package consists of three main functions (`hem`, `hem.eb.prior`, and `hem.fdr`) and several additional functions. Package information and examples can be obtained by typing `help(hem)`, `help(hem.eb.prior)`, and `help(hem.fdr)`.

The remaining sections are organized as explained below.Section 2 describes a Bayesian hierarchical error model (HEM) with experimental and biological error variances and demonstrates the use of the `hem` function by an example. Section 3 shows how to use Empirical Bayes (EB) prior specification, which is useful particularly for a small number of replicates. Section 4 explains false discovery rate (FDR) evaluation to determine a theshold value. Section 5 describes HEM when one of either biological or experimental replicates is not available. Section 6 explains the arguments and values of three functions: `hem`, `hem.eb.prior`, and `hem.fdr`.

2 Heterogeneous Error Model (HEM)

This section explains the use of two-layer EM for microarray data with both experimental and biological replicates. One of either biological or experimental replicates, however, is often unavailable; this case is described in Section 4. We assume that data are already preprocessed as normalization and log-transformation.

Suppose that $y_{i,j,k,l}$ is the l-th experimentally (or technically) replicated gene expression value of the i-th gene for a particular k-th individual (*e.g.* cell line) with the j-th biological condition (*e.g.* tissue sample), where $i = 1, \ldots, G; j = 1, \ldots, C; k = 1, \ldots, m_{i,j}; l = 1, \ldots, n_{i,j,k}$. The two-layer EM first separates the experimental error $e_{i,j,k,l}$ from the observed expression value $y_{i,j,k,l}$, so as to obtain the expression value $x_{i,j,k}$ free of the experimental error. The first layer, thus, is defined as

$$y_{i,j,k,l} | \{x_{i,j,k}, \sigma^2_{e_{i,j,k}}\} = x_{i,j,k} + e_{i,j,k,l}. \tag{1}$$

The experimental error term $e_{i,j,k,l}$ is assumed to be $i.i.d.N(0, \sigma^2_{e_{i,j,k}})$, where the experimental variance is defined to be a function of $x_{i,j,k}$, *i.e.*, $\sigma^2_{e_{i,j,k}} = \sigma^2_e(x_{i,j,k})$, since it varies on expression intensity levels. In the subsequent layer, expression intensity $x_{i,j,k}$ is decomposed into additive effects of gene, condition, and interaction, as follows

$$x_{i,j,k} | \{\mu_{i,j}, \sigma^2_{b_{i,j}}\} = \mu_{i,j} + b_{i,j,k} = \mu + g_i + c_j + r_{i,j} + b_{i,j,k}, \tag{2}$$

where μ is the parameter for the grand mean; g_i and c_j are the parameters for the gene and condition effects respectively; $r_{i,j}$ is the parameter for the interaction effect of gene and condition; and $b_{i,j,k}$ is the error term for the biological variation, assuming $i.i.d.N(0, \sigma^2_{b_{i,j}})$. The biological variance parameter $\sigma^2_{b_{i,j}}$ is allowed to be heterogeneous for each combination of gene i and condition j because a gene has its inherent biological variation under a condition.

The prior distributions are assumed to be a uniform prior on μ and normal priors on g_i, c_j, and $r_{i,j}$ with mean zero and variance σ^2_g, σ^2_c, and σ^2_r, respectively. For variance

parameters $\sigma_{b_{i,j}}^{-2}$ and $\sigma_e^{-2}(x_{i,j,k})$, gamma priors with parameters (α_b, β_b) and (α_e, β_e) are used; however, note that the constant gamma assumptions for variances can be relaxed so as to be more sensitive to heterogeneous (non-constant) variances, which is suitable for microarray data with a small number of replicates particularly, as described in Section 3.

The Gibbs sampling is utilized to estimate such a large number of parameters and unobserved data simulataneously. For discovering differentially expressed genes, the summary statistic using the estimates $(\bar{\mu}, \bar{g}_i, \bar{c}_j, \bar{r}_{i,j}, \bar{\sigma}^2_{b_{i,j}}, \bar{\sigma}^2_{e_{i,j,k}})$ is

$$H_i = \sum_{j=1}^{C} \frac{w_{i,j}(\bar{\mu}_{i,j} - \bar{\bar{\mu}}_i)^2}{(\bar{\sigma}^2_{b_{i,j}} + \sum_{k=1}^{m_{i,j}} \bar{\sigma}^2_{e_{i,j,k}}/m_{i,j})},$$

where $w_{i,j} = m_{i,j}/\sum_j m_{i,j}$, $\bar{\mu}_{i,j} = \bar{\mu}+\bar{g}_i+\bar{c}_j+\bar{r}_{i,j}$, and $\bar{\bar{\mu}}_i = \sum_j \bar{\mu}_{i,j}/C$. More differentially expressed genes have larger H-scores; hence, a desired number of genes with large H-scores are selected. The false discovery rate (FDR) can be used to determine a threshold value of H, as described in Section 4.

Example 2.1 shows how the *HEM* package is used to estimate parameters in the above model. The used primate brain data consists of gene expression from the post-mortem tissue samples of the frozen brains of three humans and three chimpanzees. Two independent tissue samples for each individual were used; thus, for HEM with two layers $m_{i,1} = 3$, $m_{i,2} = 3$, and $n_{i,j,k} = 2$, where $i = 1, \ldots, 12600$, $j = 1, 2$, $k = 1, \ldots, m_{i,j}$, and $l = 1, \ldots, n_{i,j,k}$. The design matrix must be prepared as in Example 2.1. The first column of the design matrix is for conditions, and the other two columns are for biological replicate and experimental (technical) replicate.

The `hem.preproc` function is used to do IQR normalization and log2-transformation; however, other normalization methods can be used. In the `hem` function, the argument *n.layer=2* indicates two-layer HEM, and the missing arguments *method.var.e="gam"* and *method.var.b="gam"* (as the default) represent Gamma(α_e, β_e) and Gamma(α_b, β_b) priors for experimental and biological variance parameters. In addition, the default values (*var.g=1, var.c=1, var.r=1, alpha.e=3, bet.e=.1, alpha.b=3, beta.b=.1*) are used for prior parameters unless other values are given. For estimation, 3000 MCMC samples are used after 1000 burn-ins as the default (*burn-ins=1000, n.samples=3000*).

```
Example 2.1: Two-layer HEM

> library(HEM)

> data(pbrain)  #call pbrain data
> dim(pbrain)
[1] 12600     13

> pbrain[1:5,]
```

```
> cond <- c(1,1,1,1,1,1,2,2,2,2,2,2)
> ind  <- c(1,1,2,2,3,3,1,1,2,2,3,3)
> rep  <- c(1,2,1,2,1,2,1,2,1,2,1,2)
> design <- data.frame(cond,ind,rep) #design matrix
> design
   cond ind rep
1     1   1   1
2     1   1   2
3     1   2   1
4     1   2   2
5     1   3   1
6     1   3   2
7     2   1   1
8     2   1   2
9     2   2   1
10    2   2   2
11    2   3   1
12    2   3   2
```

```
> pbrain.nor <- hem.preproc(pbrain[,2:13])            #preprocessing
> pbrain.nor[1:5,]

> pbrain.hem <- hem(pbrain.nor, n.layer=2, design=design)    #fit HEM
> pbrain.hem$H                                               #H-scores
```

H-scores are saved into `pbrain.hem$H` in the order of original data, and estimates of $x_{i,j,k}, \mu_{i,j}, \sigma^2_{e_{i,j,k}}$ and $\sigma^2_{b_{i,j}}$ are contained in **m.x, m.mu, m.var.e**, and **n.var.b** under `pbrain.hem`.

3 Empirical Bayes (EB) Prior Specification

In the previous section we used normal, gamma, or uniform priors for the model parameters. However, constant gamma priors, Gamma(α_e, β_e) and Gamma(α_b, β_b), for variances are not enough to capture heterogeneity of variances with a small number of replicates. Prior specification can be defined differently to improve the performance of variance estimation. Biological variance parameter, $\sigma^{-2}_{b_{i,j}}$, can be defined to follow a gamma prior of gene-condition specific parameters $(\alpha_b, \beta_{b_{i,j}})$, where it is important to specify appropriate values of the paramemeters. To do so, we need to employ a baseline

variance estimator such as LPE. The algorithm of parametric EB prior specification for biological variance, $\sigma^2_{b_{i,j}}$, is summarized as follows

(1) Assume that $\sigma^{-2}_{b_{i,j}}$ is distributed as Gamma$(\alpha_b, \beta_{b_{i,j}})$.

(2) Employ LPE to obtain the prior estimate of $\sigma^2_{b_{i,j}}$.

(3) Use Gibbs sampling to obtain the posterior estimate of $\sigma^2_{b_{i,j}}$.

Prior specification for experimental variance can be defined similarly. Experimental variance, however, can precisely be estimated by LPE and a re-sampling technique without assuming any parametric distribution because the variance depends on the expression intensity as a function of $x_{i,j,k}$. Experimental variance can thus be estimated as follows

(1) Assume that $\sigma^{-2}_e(x_{i,j,k})$ is distributed as an unknown distribution $h(.)$.

(2) Employ LPE and bootstrapping to estimate $h(.)$.

(3) Use the Metropolis-Hasting algorithm to obtain the posterior estimate of $\sigma^2_{e_{i,j,k}}$.

Example 3.1 shows how to apply the above EB prior specification to the primate brain data, using `hem` and `hem.eb.prior` functions. The arguments, *method.var.e="neb"* and *method.var.b="peb"*, represent a nonparametric EB prior for experimental variance and parametric EB prior for biological variance, respectively. The matrices of the variance estimates from `hem.eb.prior` are plugged into `hem`. The important (missing) arguments are q and B, which control the count of genes in a bin for pooling and the number of iterations for re-sampling respectively. The default *q=0.01* partitions 12600 genes into 100 groups based on their expression intensity levels; hence, 126 genes are used to compute a pooled variance in each bin. The default *B=25* generates 100 variance estimates for each intensity level by re-sampling.

```
Example 3.1: Two-layer HEM with EB

> pbrain.eb  <- hem.eb.prior(pbrain.nor, n.layer=2,  design=design,
                             method.var.e="neb", method.var.b="peb")
> pbrain.hem <- hem(pbrain.nor, n.layer=2,  design=design,
                    method.var.e="neb", method.var.b="peb",
                    var.e=pbrain.eb$var.e, var.b=pbrain.eb$var.b)
```

4 False Discovery Rate (FDR) Thesholding

The *HEM* package provides resampling-based FDR estimates, which can be used to determine a theshold value of F. For resampling-based estimation, it is essential to generate null data assimilating real microarray data. A within-gene permutation has a limited number of distinct permutations in a microarray experiment with a small number of replicates. In null data from across-gene (or full) permutation, variances are homogeneous over all intensity levels, which is not the pattern of practical microarray data; moreover, there is no distinction between both biological and experimental replicates that exist in raw data since full permutation does not preserve chip identities. To obtain null data assimilating practical microarray data, we resample raw data preserving gene and condition identities, *i.e.*, all of $y_{i,j,1,1}, \ldots, y_{i,j,m_{i,j},n_{i,j,k}}$ for gene i and condition j are sampled together for a generated gene under a condition.

Suppose H-statistics are computed from raw data, and H^0-statistics from generated null data. Generation of null data is repeated B times independently. Given a critical value Δ, the estimate of FDR is defined as

$$\widehat{FDR}(\Delta) = \frac{\hat{\pi}_0(\lambda)\bar{R}^0(\Delta)}{R(\Delta)}, \tag{3}$$

where $\bar{R}^0(\Delta) = \#\{H^0_{ib} | H^0_{ib} \geq \Delta,\ i = 1, \ldots, G, b = 1, \ldots, B\}/B$ is the average number of significant genes in null data, and $R(\Delta) = \#\{H_i | H_i \geq \Delta,\ i = 1, \ldots, G\}$ is the number of significant genes in raw data. The estimate of a correction factor with the λ-quantile m_λ of H^0_{ib} is $\hat{\pi}_0(\lambda) = \#\{H_i | H_i \leq m_\lambda\}/\#\{H^0_{ib} | H^0_{ib} \leq m_\lambda\}$, which is required because of the different numbers of true insignificant genes in raw data and null data.

The function `hem.fdr` provides FDRs at all H values and H critical values for given target FDRs. Example 3.1 shows how to obtain an FDR estimate using the `hem.fdr` function for the primate brain data.

```
Example 3.1 (Continued):

> pbrain.fdr <- hem.fdr(pbrain, n.layer=2,  design=design,
                        hem.out=pbrain.hem, eb.out=pbrain.eb)

> plot(pbrain.fdr$fdr)

> pbrain.fdr$targets
```

5 When One of either Biological or Experimental Replicates is Unavailble

We have described the two-layer EM for microarray data with both biological and experiemental replicates in Section 2. A biologically-replicated experiment does not have experimental replicates and an experimentally-replicated experiment does not have biological replicates, *i.e.*, one of either biological or experimental replicates is unavailable. In such cases, biological and experimental errors are confounded; hence, EM is reduced into a model with one layer as follows

$$y_{i,j,k}|\{\mu, g_i, c_j, r_{i,j}, \sigma^2_{\epsilon_{i,j}}\} = \mu + g_i + c_j + r_{i,j} + \epsilon_{i,j,k}, \tag{4}$$

where $\epsilon_{i,j,k}$ is the error term for the biological and experimental error variation, assuming $i.i.d.N(0, \sigma^2_{\epsilon_{i,j}})$. The other parameters are the same as those in the two-layer model and the $l-$subscript is suppressed in this model.

For variance parameter $\sigma^{-2}_{\epsilon_{i,j}}$, use non-contant gamma prior with hyper-parameters $(\alpha_\epsilon, \beta_{\epsilon_{i,j}})$. The parametric prior distribution can also be relaxed for variance parameters, as described in Section 3. The summary statistic is defined as $H_i = \sum_{j=1}^{C} w_{i,j}(\bar{\mu}_{i,j} - \bar{\bar{\mu}}_i)^2/\bar{\sigma}^2_{\epsilon_{i,j}}$.

The use of one-layer HEM is demonstrated with the mouse B-cell development data set, which consists of gene expression of the five consecutive stages (pre-B1, large pre-B2, small pre-B2, immature B, and mature B cells) of mouse B-cell development. The data was obtained with high-density oligonucleotide arrays, Affymetrix Mu11k GeneChip™, from flow-cytometrically purified cells. Each of six sample replicates for pre-B1 cell and each of the five sample replicates for the other conditions were hybridized on a chip, but there was no replicate for an identical sample condition (*i.e.*, $m_{i,1} = 6$, $m_{i,2} = m_{i,3} = m_{i,4} = m_{i,5} = 5$).

Example 5.1 shows how to fit one-layer HEM with Gamma$(\alpha_\epsilon, \beta_\epsilon)$ prior for variance parameter, $\sigma^{-2}_{\epsilon_{i,j}}$. Note that the design matrix consists of two columns: one for condition and the other for (individual or experimental) replicate. The argument *n.layer=1* represents one-layer HEM and the missing argument *method.var.t=1* represents Gamma$(\alpha_\epsilon, \beta_\epsilon)$ prior for variance parameter, $\sigma^{-2}_{\epsilon_{i,j}}$.

```
Example 5.1: One-layer HEM

> data(mubcp)
> dim(mubcp) #call MuBCP data
[1] 13027  26

> cond <- c(rep(1,6),rep(2,5),rep(3,5),rep(4,5),rep(5,5))
> ind <- c(1:6,rep((1:5),4))
```

```
> design <- data.frame(cond,ind)
> design
  cond ind
1    1   1
2    1   2
3    1   3

.    .   .
.    .   .
.    .   .

24   5   3
25   5   4
26   5   5

> mubcp.nor <- hem.preproc(mubcp)                  #preprocessing


> mubcp.hem <- hem(mubcp.nor, n.layer=1,design=design)       #fit HEM
> mubcp.fdr <- hem.fdr(mubcp.nor, n.layer=1, design=design,  #get FDR
                       hem.out=mubcp.hem)
```

In Example 5.2, the arguments *method.var.t="neb"* represent nonparametric EB prior specification. The prior estimate of total error variance from `hem.eb.prior` is plugged into *var.t* of `hem`.

```
Example 5.2: One-layer HEM with EB

> mubcp.eb  <- hem.eb.prior(mubcp.nor, n.layer=1, design=design,     #EB
                            method.var.t="neb")
> mubcp.hem <- hem(mubcp.nor, n.layer=1, design=design,              #fit HEM
                   method.var.t="neb", var.t=mubcp.eb$var.t)
> mubcp.fdr <- hem.fdr(mubcp.nor, n.layer=1, design=design,          #get FDR
                       hem.out=mubcp.hem, eb.out=mubcp.eb)
```

6 Explanation of Arguments and Values

This section explains the arguments and values of three main functions (`hem`, `hem.eb.prior`, and `hem.fdr`) in the *HEM* package.

```
hem(x, tr=" ", n.layer, design, burn.ins=1000, n.samples=3000,
    method.var.e="gam", method.var.b="gam", method.var.t="gam",
    var.e=NULL, var.b=NULL, var.t=NULL, var.g=1, var.c=1, var.r=1,
    alpha.e=3, beta.e=.1, alpha.b=3, beta.b=.1, alpha.t=3, beta.t=.2,
    n.digits=10, print.message.on.screen=TRUE)
```

- x: data.
- *tr*: if "log2", "log10", or "loge", then log-transformation (with base 2, 10, or e respectively) is taken.
- *n.layer*: number of layers: 1=one-layer EM; 2=two-layer EM.
- *design*: the design matrix-the first column is for conditions: if n.layer=2, the second and third columns are for biological and experimental replicates; if n.layer=1, then the second column is for one of either biological or experimental replicates.
- *burn.ins, n.samples*: numbers of burn-ins and samples for MCMC.
- *method.var.e*: prior specification method for experimental variance; "gam"=Gamma(alpha,beta), "peb"=parametric EB prior specification, "neb"=nonparametric EB prior specification (n.layer=2).
- *method.var.b*: prior specification method for biological variance; "gam"=Gamma(alpha,beta), "peb"=parametric EB prior specification (n.layer=2).
- *method.var.t*: prior specification method for total variance; "gam"=Gamma(alpha,beta), "peb"=parametric EB prior (n.layer=1).
- *var.e, var.b*: prior estimate matrices for experimental and biological error variances (n.layer=2).
- *var.t*: prior estimate matrix for total error variance (n.layer=1).
- *var.g, var.c, var.r*: variance prior parameters for the effects of gene, condition, or their interaction, *i.e.*, N(0, σ_g^2), N(0, σ_c^2), or N(0, σ_r^2).
- *alpha.e, beta.e, alpha.b, beta.b*: prior parameters for inverses of experimental and biological error variance (n.layer=2), *i.e.*, Gamma(α_e, β_e) and Gamma(α_b, β_b).

- *alpha.t, beta.t*: prior parameters for inverse of total error variance (n.layer=1), *i.e.*, Gamma($\alpha_\epsilon, \beta_\epsilon$).
- *n.digits*: number of digits.
- *print.message.on.screen*: if TRUE, process status is shown on screen.

The `hem` function provides the following values

- *n.gene, n.chip, n.cond*: numbers of genes, chips, and conditions.
- *design*: the design matrix.
- *burn.ins, n.samples*: numbers of burn-ins and samples for MCMC.
- *priors*: given prior parameters, σ_g^2, σ_c^2, σ_r^2, α_e, β_e, α_b, β_b (n.layer=2), or σ_g^2, σ_c^2, σ_r^2, $\alpha_\epsilon, \beta_\epsilon$ (n.layer=1).
- *m.mu*: estimated mean expression intensity, $\mu_{i,j} = \mu + g_i + c_j + r_{i,j}$.
- *m.x*: estimated unobserved expression intensity, $x_{i,j,k}$ (n.layer=2).
- *m.var.b*: estimated biological variances, $\sigma^2_{b_{i,j}}$ (n.layer=2).
- *m.var.e*: estimated experimental variances, $\sigma^2_{e_{i,j}}$ (n.layer=2).
- *m.var.t*: estimated total variances, $\sigma^2_{\epsilon_{i,j}}$ (n.layer=1).
- H: H-scores.

```
hem.eb.prior(x, tr=" ",  n.layer, design,
             method.var.e="neb", method.var.b="peb", method.var.t="neb",
             q=0.01, B=25, n.digits=10, print.message.on.screen=TRUE)
```

- q: quantile for partitioning genes based on expression.
- B: number of iterations for re-sampling

The `hem.eb.prior` function provides the following values

- *var.b*: prior estimate matrix for biological variances (n.layer=2).
- *var.e*: prior estimate matrix for experimental variances (n.layer=2).

- *var.t*: prior estimate matrix for total variances (n.layer=1).

In the prior estimate matrices, if parametric EB prior specification is used, the columns and rows of the matrices are for conditions and genes respectively. If nonparametric EB prior specification is used, the columns are for quantiles and the rows are for resampling iterations and expression quantiles, *i.e.*, the matrix size is $(B+1)\times$(number of conditions) by $(1/q)$.

```
hem.fdr(x, tr=" ", n.layer, design, hem.out, eb.out=NULL, n.iter=5, q.trim=0.9,
        target.fdr=c(0.001,0.005,0.01,0.05,0.1,0.15,0.20,0.30,0.40,0.50),
        n.digits=10, print.message.on.screen=TRUE)
```

- *hem.out*: output from `hem` function.
- *eb.out*: output from `hem.eb.prior` function.
- *n.iter*: number of iterations.
- *q.trim*: quantile used for estimtaing the proportion of true negatives (pi0).
- *target.fdr*: target FDRs

The `hem.fdr` function provides the following values

- *fdr*: H-scores and corresponding FDRs.
- *pi0*: estimated proportion of true negatives.
- *H.null*: H-scores from null data.
- *targets*: given target FDRs; corrsponding critical values and numbers of significant genes are provided.