The Korean Journal of Applied Statistics (2011) DOI: http://dx.doi.org/10.5351/KJAS.2011.24.4.609
24(4), 609–622

A Finite Mixture Model for Gene Expression and Methylation Profiles in a Bayesian Framework

Jaesik Jeong[1]

[1]Department of Biostatistics, Indiana University

(Received March 2011; accepted June 2011)

Abstract

The pattern of methylation draws significant attention from cancer researchers because it is believed that DNA methylation and gene expression have a causal relationship. As the interest in the role of methylation patterns in cancer studies (especially drug resistant cancers) increases, many studies have been done investigating the association between gene expression and methylation. However, a model-based approach is still in urgent need. We developed a finite mixture model in the Bayesian framework to find a possible relationship between gene expression and methylation. For inference, we employ Expectation-Maximization(EM) algorithm to deal with latent (unobserved) variable, producing estimates of parameters in the model. Then we validated our model through simulation study and then applied the method to real data: wild type and hydroxytamoxifen(OHT) resistant MCF7 breast cancer cell lines.

Keywords: Expectation-Maximization, hierarchical statistical model, latent variable, methylation, mixture model.

1. Introduction

Epigenetic events (that include DNA methylation) are involved in complex processes of biological interactions that results in the regulation of gene expression (Herman, 1999; Hinshelwood and Clark, 2008; Jeong *et al.*, 2010). DNA methylation is the process that add a methyl group to the 5 position of the cytosine pyrimidine ring, that can be inherited from cell divisions. In mammals, DNA methylation is a crucial part of normal organismal development and cellular differentiation and stably alters the gene expression pattern in cells. In addition, DNA methylation plays a key role in the development of almost all types of cancer (Baylin and Herman, 2000; Bird, 2002; Herman and Baylin, 2003; Dwivedi *et al.*, 2003).

In the clinical setting, drug resistance is a critical issue for cancer treatment. Sometimes drug resistance happens from the beginning of treatment. However, some people do respond first and stop responding after a course of treatment of the same drug that is known as acquired drug resistance. Reasons for such resistance possibly result from epigenetic events such as DNA methylation and chromatin modification. Especially, epigenetic alteration such as DNA methylation plays a key role in acquired drug resistance (Jones and Laird, 1999; Jones and Baylin, 2007).

[1]Post-doctor, Department of Biostatistics, Indiana University, 410 West 10th Street, Suite 3000, Indianapolis, IN 46202, USA. E-mail: jeongjae@iupui.edu

As the interest in the role of methylation pattern in cancer study increases, many studies have investigated the association between gene expression and methylation (Ahuja *et al.*, 1997; Muller *et al.*, 2001; Esteller *et al.*, 1999; Hui *et al.*, 2000; Wang *et al.*, 2009). In a classic approach, fold change is used for status call and χ^2 test (Fisher's exact test for small sample) is used to test for association (Ahuja *et al.*, 1997; Muller *et al.*, 2001; Hui *et al.*, 2000). Such methods provide a kind of global view (category-specific) on the interplay between gene expression and methylation. However, there is increasing evidence that such association is gene-specific. Model-based approaches (especially an empirical Bayes model) have been developed to illustrate the evidence of a local view (gene-specific) other than a global view (Jeong *et al.*, 2010).

A hierarchical statistical model in the Bayesian framework (hierarchical Bayes model) was developed by Jeong *et al.* (2010) and George (1985). They assume normality on log-transformed fold change and specify normal prior on the mean vector. Inference is based on the posterior distribution of the mean vector of each gene. The estimated prior covariance provides a global view and the covariance estimate from the posterior distribution of each gene gives a local view on associations. This method improves the classic χ^2 test because both the global and the local view of association are provided. It, however, still requires an artificial constant choice to specify the window that is used to call the status of each gene that can be removed in the nine component mixture model.

Since a general understanding about the relationship between gene expression and methylation is that hypomethylated genes are more likely upregulated, people are interested in a specific category, for example, category of hypomethylation and upregulation (Ahuja *et al.*, 1997; Muller *et al.*, 2001; Esteller *et al.*, 1999; Hui *et al.*, 2000; Wang *et al.*, 2009). However, the method of Jeong *et al.* (2010) is not category-specific because all (nine) combinatorial categories are explained with only one normal distribution even though it provides a global and local (gene-specific) view on association. Thus, we developed a nine component normal mixture model in the Bayesian framework to obtain a clear insight on the global association in each category (Day, 1969; Figueiredo and Jain, 2002; Xu and Jordan, 1996). In a sense, our method has nine global views and each view corresponds to each category, respectively. The main difference from the method of Jeong *et al.* (2010) is that our method assigns a normal distribution to each category that provides a gene-specific view plus category-specific view. In addition, we do not need to divide sample space for gene assignment, implying that the choice of constant for space categorization is not required. We employ Expectation-Maximization(EM) algorithm to deal with latent (unobserved) variable, which produces estimates of parameters in the model.

The remainder of the paper is consisted as follows. In Section 2, we describe the model. Estimation is described in Section 3. Then, we validate our model through simulation study in Section 4 and real data analysis is given in Section 5. We conclude the paper in Section 6.

2. The Model

We have two data sets: gene expression(GE) and DNA methylation data. Both GE and methylation data have three different classes, respectively (GE: up-regulation, no change, down-regulation; Methylation: hypomethylation, no change, hypermethylation). Since we work on the cointegrated data, it is very natural to consider nine component mixture model and each component is assigned to each combinatorial classes. We look at marginal model first and move on joint model. Such transition helps us to understand the complex model.

2.1. Marginal model on gene expression

In gene expression data, we have two groups to compare the wild type(WT) and OHT resistant group. Each group has 4 replicates and there are different number of probes within each gene. The marginal model we use here is:

$$G_{ijkl} = \mu_{il} + b_{ij} + \epsilon_{ijkl}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J_i,\ k = 1, \ldots, K,\ l = 1, 2,$$

where G_{ijkl} is gene expression of probe j within gene i at k^{th} replicate in group l. Note that μ_{il} is average expression of gene i in group l, b_{ij} added effect of probe j for gene i and ϵ_{ijkl} is error term. For three different status for gene expression, we consider three component mixture model in which each component corresponds to the each status: up-regulated, not differentially expressed, and down-regulated. We assume normality on each component in the model:

$$b_{ij} \sim N\left(0, \sigma^2\right), \qquad \epsilon_{ijkl} \sim N\left(0, \delta^2\right)$$

$$\mu_i \equiv (\mu_{i1}, \mu_{i2}) \stackrel{d}{=} \begin{cases} N(\lambda_1, \Sigma_1), & \text{if gene } i \text{ is down-regulated } (X_i = 1), \\ N(\lambda_0, \Sigma_0), & \text{No change } (X_i = 0), \\ N(\lambda_{-1}, \Sigma_{-1}), & \text{if gene } i \text{ is up-regulated } (X_i = -1). \end{cases}$$

We can re-express the model in the linear form: for gene i,

$$G_i = \triangle_{i1}\beta_i + \epsilon_i,$$

where

$$\triangle_{i1} = \left[\begin{array}{cc|ccc} 1_4 & 0 & 1_4 & & \\ \vdots & \vdots & & \ddots & \\ 1_4 & 0 & & & 1_4 \\ \hline 0 & 1_4 & 1_4 & & \\ \vdots & \vdots & & \ddots & \\ 0 & 1_4 & & & 1_4 \end{array}\right], \quad \beta_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ b_1 \\ \vdots \\ b_{J_i} \end{pmatrix}, \quad 1_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

and

$$G_i = \left(G_{i111}, G_{i121}, G_{i131}, G_{i141}, G_{i211}, G_{i221}, G_{i231}, G_{i241}, \ldots, G_{iJ_i11}, \ldots, G_{iJ_i41}, \ldots\right)^t.$$

In the case that gene i is down-regulated, each component in our model has the following distribution:

$$\beta_{i|X_i=1} \sim N(\mu_1{}^*, \Sigma_1{}^*), \quad \mu_1{}^* = (\lambda_1, 0)^t,\ \Sigma_1{}^* = \begin{pmatrix} \Sigma_1 & o \\ 0 & \sigma^2 I_{J_i} \end{pmatrix}$$

and

$$G_{i|X_i=1,\beta_i} \sim N\left(\triangle_{i1}\beta_i,\ \delta^2 I_{n_i}\right)$$

where $\epsilon_i \sim N(0, \delta^2 I_{n_i})$ and $n_i = 2K \times J_i$.

2.2. Marginal model on methylation

Similarly, marginal model for methylation is given:

$$M_{ijl} = \eta_{il} + a_{ij} + d_{ijl}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J_i,\ l = 1, 2,$$

where M_{ijl} is methylation data of probe j within gene i in group l. Note that η_{il} is mean effect of gene i in group l, a_{ij} added effect of probe j for gene i, and d_{ijl} error term. The only difference is the number of replicates: no replicate here. Again, we consider a three component mixture model for three different status of methylation: hypomethylated, unmethylated, and hypermethylated. We consider following distributions for each component:

$$a_{ij} \sim N\left(0, \omega^2\right), \qquad d_{ijl} \sim N\left(0, \tau^2\right)$$

$$\eta_i \equiv (\eta_{i1}, \eta_{i2}) \stackrel{d}{=} \begin{cases} N(\xi_1, \Omega_1), & \text{if gene } i \text{ is hypomethylated } (Y_i = 1), \\ N(\xi_0, \Omega_0), & \text{No change } (Y_i = 0), \\ N(\xi_{-1}, \Omega_{-1}), & \text{if gene } i \text{ is hypermethylated } (Y_i = -1). \end{cases}$$

Thus, the rearranged model in the linear form is given: for gene i,

$$M_i = \triangledown_{i1}\alpha_i + d_i,$$

where

$$\triangledown_{i1} = \left[\begin{array}{cc|c|c} 1 & 0 & 1 & \\ \vdots & \vdots & & \ddots \\ 1 & 0 & & 1 \\ \hline 0 & 1 & 1 & \\ \vdots & \vdots & & \ddots \\ 0 & 1 & & 1 \end{array}\right], \quad \alpha_i = \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ a_1 \\ \vdots \\ a_{J_i} \end{pmatrix}$$

and

$$M_i = (M_{i11}, \ldots, M_{iJ_i1}, M_{i12}, \ldots, M_{iJ_i2})^t.$$

As an illustration, when gene i is hypermethylated, each component in our model follows the distribution below:

$$\alpha_{i\,|Y_i=1} \sim N(\eta_1{}^*, \Omega_1{}^*), \quad \eta_1{}^* = (\xi_1, 0)^t,\ \Omega_1{}^* = \begin{pmatrix} \Omega_1 & o \\ 0 & \omega^2 I_{J_i} \end{pmatrix}$$

and

$$M_{i\,|Y_i=1,\alpha_i} \sim N\left(\triangledown_{i1}\alpha_i, \tau^2 I_{m_i}\right),$$

where $d_i \sim N(0, \tau^2 I_{m_i})$ and $m_i = 2 \times J_i$.

2.3. Joint model

In this section, we consider joint model on merged data set, *i.e.*, $D_i = (G_i{}^t, M_i{}^t)^t$. We apply nine component mixture model to the data.

Hyper	C3	C6	C9
M	C2	C5	C8
Hypo	C1	C4	C7
	Down regulation	GE	Up regulation

Figure 2.1. Nine categories; X-axis presents three different gene expression status: down-regulation, no change, up-regulation. Y-axis presents three different methylation status: hypomethylation, no change, hypermethylation.

$$\begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \eta_{i1} \\ \eta_{i2} \end{bmatrix} \stackrel{d}{=} \begin{cases} N(\mu_1, \Sigma_1), & Z_i = 1 \ (X_i = 1 \text{ and } Y_i = 1), \\ N(\mu_2, \Sigma_2), & Z_i = 2 \ (X_i = 1 \text{ and } Y_i = 0), \\ N(\mu_3, \Sigma_3), & Z_i = 3 \ (X_i = 1 \text{ and } Y_i = -1), \\ N(\mu_4, \Sigma_4), & Z_i = 4 \ (X_i = 0 \text{ and } Y_i = 1), \\ N(\mu_5, \Sigma_5), & Z_i = 5 \ (X_i = 0 \text{ and } Y_i = 0), \\ N(\mu_6, \Sigma_6), & Z_i = 6 \ (X_i = 0 \text{ and } Y_i = -1), \\ N(\mu_7, \Sigma_7), & Z_i = 7 \ (X_i = -1 \text{ and } Y_i = 1), \\ N(\mu_8, \Sigma_8), & Z_i = 8 \ (X_i = -1 \text{ and } Y_i = 0), \\ N(\mu_9, \Sigma_9), & Z_i = 9 \ (X_i = -1 \text{ and } Y_i = -1). \end{cases}$$

As an illustration, nine categories are plotted in Figure 2.1. The value of Z_i presents the nine combinatorial classes. For example, the class of $Z_i = 1$ corresponds to C1 in the figure

For specific gene i, the linear model we use here is:

$$D_i = \triangle_i{}^\star \beta_i{}^\star + \epsilon_i{}^\star,$$

where

$$\triangle_i{}^\star = \left[\begin{array}{c|c} \begin{matrix} 1_4 & & \\ \vdots & & \\ 1_4 & & \\ & 1_4 & \\ & \vdots & \\ & 1_4 & \end{matrix} & \begin{matrix} 1_4 & & \\ & \ddots & \\ & & 1_4 \\ 1_4 & & \\ & \ddots & \\ & & 1_4 \end{matrix} \\ \hline \begin{matrix} 1 & \\ \vdots & 0 \\ 1 & \\ & 1 \\ 0 & \vdots \\ & 1 \end{matrix} & \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \\ 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \end{array}\right], \qquad \beta_i{}^\star = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \eta_{i1} \\ \eta_{i2} \\ b_1 \\ \vdots \\ b_{J_i} \\ a_1 \\ \vdots \\ a_{H_i} \end{pmatrix}.$$

The distribution assumption for each component in our model is given:

$$(\mu_{i1}, \mu_{i2}, \eta_{i1}, \eta_{i2})^t{}_{|Z_i=l} \sim N(\lambda_l, \Sigma_l), \qquad \beta_i{}^\star{}_{|Z_i=l} \sim N(\mu_l{}^\star, \Sigma_l{}^\star)$$

and

$$D_{i|Z_i=l,\beta_i{}^\star} \sim N\left(\triangle_{il}{}^\star\beta_i{}^\star, \begin{pmatrix} \delta^2 I_{n_i} & 0 \\ 0 & \tau^2 I_{m_i} \end{pmatrix}\right),$$

where

$$\mu_l{}^\star = (\mu_l{}^t, 0)^t, \quad \Sigma_l{}^\star = \begin{pmatrix} \Sigma_l & 0 & 0 \\ 0 & \sigma^2 I_{J_i} & 0 \\ 0 & 0 & \omega^2 I_{H_i} \end{pmatrix}.$$

Thus, hierarchical statistical model is given:

$$\begin{aligned} \beta^\star{}_{i|Z_i=l} &\sim N(\mu_l{}^\star, \Sigma_l{}^\star), \\ D_{i|Z_i=l,\beta_i{}^\star} &\sim N\left(\triangle_{il}{}^\star\beta_i{}^\star, \mathrm{diag}\left(\delta^2 I_{n_i}, \tau^2 I_{m_i}\right)\right). \end{aligned}$$

3. Estimation

We introduce Expectation-Maximization(EM) algorithm to handle latent variables. The EM algorithm consists of two steps, Expectation and Maximization (Dempster *et al.*, 1977; McLachlan and Krishnan, 2007; Sundberg, 1974, 1976; Wu, 1983). In the E-step, conditional expectation of complete-data log likelihood given observed data is calculated. Then, parameter estimates are updated in the M-step. These two iterative steps are repeated until convergence of the algorithm is attained. We briefly describe the EM algorithm applied to our case here.

3.1. E-step

In this step, we calculate conditional expectation of complete-data log likelihood given observed data:

$$Q\left(\theta; \theta^{(k)}\right) \equiv E\left[\log L_c(\theta)|D, \theta^{(k-1)}\right],$$

where $\log L_c(\theta)$ is complete-data log likelihood function and θ is the parameter vector. We need to calculate proportion estimates and conditional expectation of the mean vector. The proportion variable for gene i is defined as:

$$(Z_{i1}, \ldots, Z_{i9}) = \begin{cases} (1, 0, \ldots, 0), & \text{if } Z_i = 1, \\ \vdots & \vdots \\ (0, \ldots, 0, 1), & \text{if } Z_i = 9, \end{cases}$$

where $P(Z_{il} = 1) = \rho_{il}$, $l = 1, \ldots, 9$ and $\sum_l \rho_{il} = 1$.

At iteration k, posterior probability of each proportion variable of belonging to category l is given:

$$E\left(Z_{il}|D_i, \theta^{(k-1)}\right) = P\left(Z_{il} = 1|D_i, \theta^{(k-1)}\right) = \frac{\rho_l{}^{(k-1)}\left[D_i|Z_{il} = 1, \theta^{(k-1)}\right]}{\sum_d \rho_d{}^{(k-1)}\left[D_i|Z_{il} = 1, \theta^{(k-1)}\right]} = \rho_{il}{}^{(k)}.$$

Then, we calculate conditional expectation of mean vector, β_i. Here we need to calculate two things:

$$E\left(\beta_i|D_i, \theta^{(k-1)}, Z_{il} = 1\right), \qquad \mathrm{Cov}\left(\beta_i|D_i, \theta^{(k-1)}, Z_{il} = 1\right).$$

To this end, we derive posterior distribution of parameter given data:

$$\left[\beta_i | D_i, \theta^{(k-1)}, Z_{il} = 1\right] \sim N(K, K^*),$$

where $K^* = ({\triangle_i}^T {\Sigma_e}^{-1} \triangle_i + {\Sigma_{pl}}^{-1})^{-1} \equiv {V_{il}}^{(k)}$ and $K = K^*({\triangle_i}^T {\Sigma_e}^{-1} D_i + {\Sigma_{pl}}^{-1} {\mu_l}^{*(k)}) \equiv {\Lambda_{il}}^{(k)}$.

3.2. M-step

Once E-step is done, we update parameter estimates by maximizing the target function, $Q(\theta; \theta^{(k)})$. The complete-data log likelihood for gene i is given:

$$\log L_c(\theta) = \sum_l Z_{il} \left\{\log \rho_l + \log[D_i, {\beta_i}^\star | {Z_i}^\star, \theta]\right\}$$

and $\sum_l \rho_l = 1$. Thus, maximization with a constraint on ρ_l can be solved by using Lagrange Multipliers:

$$R(\theta, \lambda) = Q\left(\theta; \theta^{(k)}\right) - \lambda \left(\sum_l \rho_l - 1\right).$$

Under the normality assumptions, parameter estimators can be easily calculated with algebra and each parameter in the model has the closed form estimator, respectively. At the k^{th} iteration, estimators in closed form are given:

$$\begin{aligned}
{\rho_l}^{(k+1)} &= \frac{\sum_i {\rho_{il}}^{(k)}}{I}, \\
({\mu_l}^*)^{(k+1)} &= \frac{\sum_i {\rho_{il}}^{(k)} {\Lambda_{il}}^{(k)}}{\sum_i {\rho_{il}}^{(k)}}, \\
\delta^2 &= \frac{1}{N} \sum_i \sum_l {\rho_{il}}^{(k)} \left[\mathrm{tr}\left({\triangle_i}^{(1)T} {\triangle_i}^{(1)} {V_{il}}^{(k)}\right) + \left(D_i - \triangle_i {\Lambda_{il}}^{(k)}\right)^{(1)T} \left(D_i - \triangle_i {\Lambda_{il}}^{(k)}\right)^{(1)}\right], \\
\tau^2 &= \frac{1}{M} \sum_i \sum_l {\rho_{il}}^{(k)} \left[\mathrm{tr}\left({\triangle_i}^{(2)T} {\triangle_i}^{(2)} {V_{il}}^{(k)}\right) + \left(D_i - \triangle_i {\Lambda_{il}}^{(k)}\right)^{(2)T} \left(D_i - \triangle_i {\Lambda_{il}}^{(k)}\right)^{(2)}\right], \\
\sigma^2 &= \frac{8}{N} \sum_i \sum_l {\rho_{il}}^{(k)} \left[\mathrm{tr}\left({V_{il}}^{2(k)}\right) + \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(2T)} \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(2)}\right], \\
\omega^2 &= \frac{2}{M} \sum_i \sum_l {\rho_{il}}^{(k)} \left[\mathrm{tr}\left({V_{il}}^{3(k)}\right) + \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(3T)} \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(3)}\right], \\
\Sigma_l &= \frac{1}{\sum_i {\rho_{il}}^{(k)}} \sum_i {\rho_{il}}^{(k)} \left[{V_{il}}^{1(k)} + \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(1)} \left({\mu_l}^* - {\Lambda_{il}}^{(k)}\right)^{(1T)}\right],
\end{aligned}$$

where I is the number of genes. Note that

$$\triangle_i = \begin{pmatrix} {\triangle_i}^{(1)} \\ {\triangle_i}^{(2)} \end{pmatrix}, \quad V_{il} = \begin{pmatrix} {V_{il}}^{(1)} & \cdot & \cdot \\ \cdot & {V_{il}}^{(2)} & \cdot \\ \cdot & \cdot & {V_{il}}^{(3)} \end{pmatrix} \quad \text{and} \quad \Lambda_{il} = \begin{pmatrix} {\Lambda_{il}}^{(1)} \\ {\Lambda_{il}}^{(2)} \\ {\Lambda_{il}}^{(3)} \end{pmatrix}.$$

3.3. Inference on relationship between GE and methylation

Given the parameter estimates of interest, we can characterize relationship between gene expression and methylation. For local (gene-specific) association of gene i, we use covariance estimate from posterior distribution of gene i:

$$\widehat{\text{corr}}_l(\text{GE}_i, M_i) = \frac{\widehat{\text{cov}}_l(\text{GE}_i, M_i)}{\sqrt{\widehat{\text{var}}_l(\text{GE}_i)}\sqrt{\widehat{\text{var}}_l(M_i)}},$$

where $\text{corr}_l(\text{GE}_i, M_i)$ is local correlation between gene expression and methylation of gene i. For a category-specific association of a category s, we use estimated prior covariance for category s:

$$\widehat{\text{corr}}_s(\text{GE}, M) = \frac{\widehat{\text{cov}}_s(\text{GE}, M)}{\sqrt{\widehat{\text{var}}_s(\text{GE})}\sqrt{\widehat{\text{var}}_s(M)}},$$

where $\text{corr}_s(\text{GE}, M)$ is category-specific correlation between gene expression and methylation of category s.

4. Simulation Study

In this section, we evaluate our method through simulation study. We generate data that mimics the structure of real data as much as possible given in the next section. Structural similarity is summarized as follows: (1) data were generated based on normality assumption, (2) we consider two groups such as case and control, (3) within each group, gene expression has four replication, but no replicate for methylation, and (4) each gene has a couple of probes ranging from 1 to 4.

4.1. Simulation setup

The data are generated jointly by using pre-specified values given below.

- The number of genes: 800.
- Each gene has a few probes: $1 \sim 4$.

The true parameter values are given:

$$\rho = (1,2,1,2,4,2,1,2,1)/16, \quad \sigma^2 = 0.81, \quad \delta^2 = 1.96, \quad \omega^2 = 0.64, \quad \tau^2 = 1.44,$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 9 \\ 3 \\ 9 \\ 3 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 2 & 1.8 \\ 0.1 & 0.1 & 1.8 & 2 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 9 \\ 3 \\ 6 \\ 6 \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1.5 & 1.2 \\ 0.1 & 0.1 & 1.2 & 1.5 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 9 \\ 3 \\ 3 \\ 9 \end{pmatrix}, \quad \Sigma_4 = \begin{pmatrix} 2 & 1.8 & 0.1 & 0.1 \\ 1.8 & 2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{pmatrix}, \quad \mu_4 = \begin{pmatrix} 6 \\ 6 \\ 9 \\ 3 \end{pmatrix},$$

$$\Sigma_5 = \begin{pmatrix} 2 & 1.8 & 0.1 & 0.1 \\ 1.8 & 2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 2 & 1.8 \\ 0.1 & 0.1 & 1.8 & 2 \end{pmatrix}, \quad \mu_5 = \begin{pmatrix} 6 \\ 6 \\ 6 \\ 6 \end{pmatrix}, \quad \Sigma_6 = \begin{pmatrix} 2 & 1.8 & 0.1 & 0.1 \\ 1.8 & 2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1.5 & 1.2 \\ 0.1 & 0.1 & 1.2 & 1.5 \end{pmatrix}, \quad \mu_6 = \begin{pmatrix} 6 \\ 6 \\ 3 \\ 9 \end{pmatrix},$$

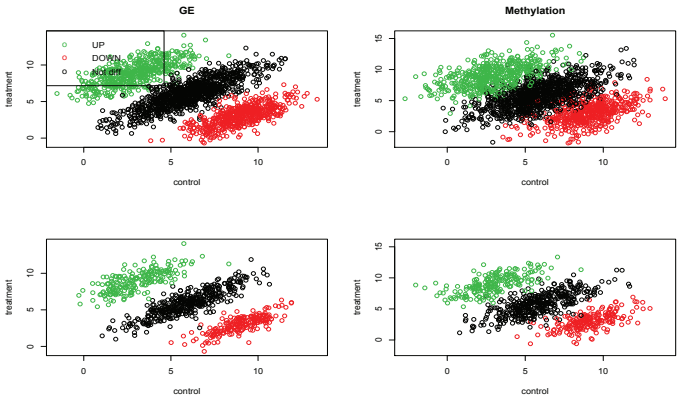

Figure 4.1. Simulated data; top: gene expression(left) and methylation(right) at the probe level; bottom: gene expression(left) and methylation(right) at the gene level

Table 4.1. Simulated data: Parameter estimates

Parameter	True value	Parameter estimates
(ρ_1, ρ_2, ρ_3)	(0.0625, 0.125, 0.0625)	(0.06625, 0.11750, 0.06750)
(ρ_4, ρ_5, ρ_6)	(0.1250, 0.250, 0.1250)	(0.12750, 0.23625, 0.13375)
(ρ_7, ρ_8, ρ_9)	(0.0625, 0.125, 0.0625)	(0.07000, 0.12125, 0.06000)
(δ^2, τ^2)	(1.96, 1.44)	(1.9385, 1.3358)
(σ^2, ω^2)	(0.81, 0.64)	(0.8395, 0.7208)
(μ_{11}, μ_{12})	(9, 3)	(9.3237, 3.2087)
(μ_{41}, μ_{42})	(6, 6)	(5.8922, 5.9481)
(μ_{71}, μ_{72})	(3, 9)	(3.0097, 9.0770)
(η_{11}, η_{12})	(9, 3)	(8.6763, 2.6672)
(η_{21}, η_{22})	(6, 6)	(5.6981, 5.5928)
(η_{31}, η_{32})	(3, 9)	(2.9341, 8.8845)

$$\Sigma_7 = \begin{pmatrix} 1.5 & 1.2 & 0.1 & 0.1 \\ 1.2 & 1.5 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{pmatrix}, \quad \mu_7 = \begin{pmatrix} 3 \\ 9 \\ 9 \\ 3 \end{pmatrix}, \quad \Sigma_8 = \begin{pmatrix} 1.5 & 1.2 & 0.1 & 0.1 \\ 1.2 & 1.5 & 0.1 & 0.1 \\ 0.1 & 0.1 & 2 & 1.8 \\ 0.1 & 0.1 & 1.8 & 2 \end{pmatrix}, \quad \mu_8 = \begin{pmatrix} 3 \\ 9 \\ 6 \\ 6 \end{pmatrix},$$

$$\Sigma_9 = \begin{pmatrix} 1.5 & 1.2 & 0.1 & 0.1 \\ 1.2 & 1.5 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1.5 & 1.2 \\ 0.1 & 0.1 & 1.2 & 1.5 \end{pmatrix}, \quad \mu_9 = \begin{pmatrix} 3 \\ 9 \\ 3 \\ 9 \end{pmatrix}.$$

The simulated data obtained by using the values above are given in Figure 4.1.

Here we consider the iteration size of 300. Parameter estimates are summarized in Table 4.1 in order to check the accuracy of our method

Since our estimates are very close to the true value of each parameter, we are sure that our method is working very well under the condition that our assumption is correct. In addition, we checked the accuracy of the assignment. To this end, we apply a cutoff value to the posterior probability that each gene belongs to each category. Here we consider a stringent cutoff value of 0.9. Assignment results are summarized in the Table 4.2. As we can see, our method correctly assigns 769 genes out of 800, *i.e.*, accuracy is more than 96%(796/800 = 0.96125).

Table 4.2. Simulated data: The results of category assignment; NG: number of genes(true value), NAG: number of assigned genes, NCA: number of correctly assigned genes, NIA: number of incorrectly assigned genes

Cutoff = 0.9	C1	C2	C3	C4	C5	C6	C7	C8	C9	Tot
NG	50	100	50	100	200	100	50	100	50	800
NAG	53	94	54	102	189	107	56	97	48	800
NCA	50	94	50	97	186	99	50	95	48	769
NIA	3	0	4	5	3	8	6	2	0	31

Table 5.1. GE data structure

		Group 1				Group 2			
		$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
Gene 1	b_{11}	G_{1111}	G_{1121}	G_{1131}	G_{1141}	G_{1112}	G_{1122}	G_{1132}	G_{1142}
	b_{12}	G_{1211}	G_{1221}	G_{1231}	G_{1241}	G_{1212}	G_{1222}	G_{1232}	G_{1242}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	b_{1J_1}	G_{1J_111}	G_{1J_121}	G_{1J_131}	G_{1J_141}	G_{1J_112}	G_{1J_122}	G_{1J_132}	G_{1J_142}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene I	b_{I1}	G_{I111}	G_{I121}	G_{I131}	G_{I141}	G_{I112}	G_{I122}	G_{I132}	G_{I142}
	b_{I2}	G_{I211}	G_{I221}	G_{I231}	G_{I241}	G_{I212}	G_{I222}	G_{I232}	G_{I242}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	b_{IJ_I}	G_{IJ_I11}	G_{IJ_I21}	G_{IJ_I31}	G_{IJ_I41}	G_{IJ_I12}	G_{IJ_I22}	G_{IJ_I32}	G_{IJ_I42}

5. Application to Real Data

In the previous section, we noticed that our method is working well when our assumption is correct. In this section, we apply our method to real data: wild type and OHT resistant MCF7 breast cancer cell lines.

5.1. Data description (OHT versus WT)

For gene expression analysis, the Human Genome U133A 2.0 Array was used; in addition, differential methylation hybridization(DMH) was done using Affymetrix oligonucleotide microarrays. Microarray Analysis Suite(MAS) version 5.0 was used for preprocessing. Experimental details were described in (Fan *et al.*, 2006).

We compare two groups, wild type and OHT resistant cell line. Each group in gene expression has four replicates while each group in methylation has no replicate. GE data structure is given in the Table 5.1. We, however, restrict our attention to genes with at least two "present call" for gene expression and our focus to DNA methylation in the promoter region. Then, we select common genes existing in both data sets. As a result, gene expression data have 11286 probes at the probe level and 4078 genes at the gene level while methylation data have 10223 probes and 4078 genes. The raw data plot is given in Figure 5.1.

5.2. Results

As an initial value for long chain, we use parameter estimates obtained from pilot study with iteration size of 300. Based on the result of the parameter estimates, we estimate that the iteration size of 2000 is enough to make sure the convergence of EM algorithm. As an illustration, trace plot of four variance estimates are given in Figure 5.2.

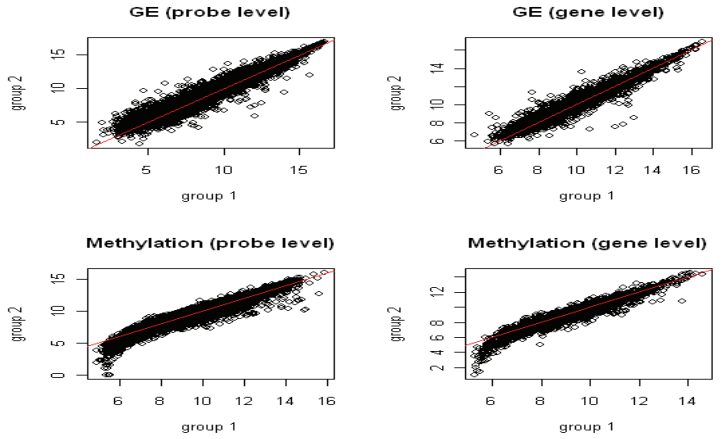

Figure 5.1. Raw data plot

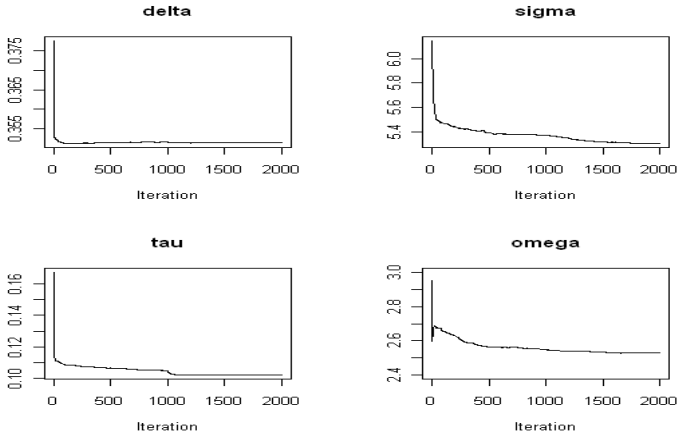

Figure 5.2. Four variance estimates

All parameter estimates are summarized in Table 5.2. Here $\rho_1, \ldots, \rho_9$ are proportion estimates corresponding to each category, respectively. The majority of the genes (30%) belong to Category 5 (C5).

At the category level, global association is summarized in Table 5.3.

In Table 5.3, it is clear that association is category-specific. For example, let us focus on the first row of the Table 5.3, correlation between wild type gene expression and wild type methylation. Each category has totally different correlation estimates ranging from -0.61 to 0.95, implying that correlation is category-specific.

To examine the association at the gene level, we selected one category (say, Category 3) and calculated gene-specific association three different ways as done in the global case: (1) correlation between WT GE and WT Methylation, (2) correlation between resistant GE and resistant Methylation, (3) correlation between difference in the two groups of GE and that of Methylation. In Figure 5.3, three different histograms that show the distribution of local (gene-specific) associations

Table 5.2. Real data: Parameter estimates

Parameter	Parameter estimates
(ρ_1, ρ_2, ρ_3)	(0.1135, 0.0780, 0.0329)
(ρ_4, ρ_5, ρ_6)	(0.1089, 0.2990, 0.0933)
(ρ_7, ρ_8, ρ_9)	(0.1275, 0.1017, 0.0453)
(δ^2, τ^2)	(0.3514, 0.1024)
(σ^2, ω^2)	(5.3047, 2.5286)
(μ_{11}, μ_{12})	(9.9819, 9.6381)
(μ_{41}, μ_{42})	(11.1259, 11.1306)
(μ_{71}, μ_{72})	(8.6259, 9.0194)
(η_{11}, η_{12})	(8.1152, 7.6924)
(η_{21}, η_{22})	(10.0419, 9.8902)
(η_{31}, η_{32})	(8.7607, 9.2403)

Table 5.3. Estimated global correlation for nine category

Corr	C1	C2	C3	C4	C5	C6	C7	C8	C9
$\mathrm{Corr}(\mathrm{GE}_W, M_W)$	0.176	0.130	−0.081	0.264	0.379	0.947	−0.299	−0.607	−0.210
$\mathrm{Corr}(\mathrm{GE}_R, M_R)$	0.228	0.635	0.299	0.366	0.427	0.890	−0.292	−0.680	−0.209
$\mathrm{Corr}(\mathrm{GE}_d, M_d)$	0.237	0.041	−0.636	−0.554	−0.516	0.918	−0.930	−0.720	−0.256

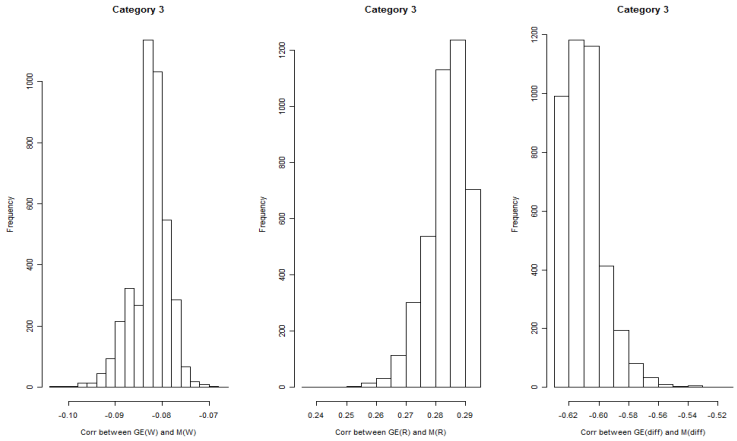

Figure 5.3. Histogram of local correlation for category 3: left(correlation between gene expression wild type and methylation wild type); center(correlation between gene expression resistant and methylation resistant); right(correlation between gene expression difference and methylation difference)

are given. As we can see, the local correlations from each gene are widely distributed, implying that correlation is gene-specific.

As an illustrating example, we selected gene CDH3 that is very crucial in breast cancer studies. It is well known that the CDH3 gene, which act as a tumor suppressor gene, is hypomethylated in breast cancer. Our method assigned the gene to Category 4 (no change in GE and hypomethylation). Estimated mean of the gene is $\hat{\mu} = (10.338, 10.417, 6.754, 5.113)$ where $\mu = (\mu_{GE_W}, \mu_{GE_R}, \eta_{M_W}, \eta_{M_R})$. Furthermore, at the category level, $\mathrm{Corr}_4(\mathrm{GE}_d, M_d) = -0.554$ and local association for the gene CDH3, $\mathrm{Corr}_{\mathrm{CDH3}}(\mathrm{GE}_d, M_d)$ was -0.272. Even though there is weak signal, our results are consistent with the general understanding that hypomethylated genes are more likely to be upregulated.

6. Conclusion

In this article, we constructed hierarchical Bayes model to get clear insight on the interplay between gene expression and DNA methylation in promoter region. Our model provides a global (category-specific) and local (gene-specific) view on association, and such rich information might be used for an understanding of epigenetic therapy, leading to the important part of drug discovery. For example, through the restoration of DNA methylation patterns, we may make cancer cells respond back to the treatment.

Our results show that category-specific (global) correlation varies from category to category (for example, $\mathrm{Corr}(\mathrm{GE}_W, M_W)$ ranges from -0.6 to 0.9 in Table 5.3). In addition, within each category (say, Category 3), the distributions of three different types of local correlations are different. Collectively, our results show that association between gene expression and DNA methylation is category-specific and gene-specific as well.

Applying a cutoff value to the posterior probability of each gene, our method assigns each gene to one of nine categories. After such gene assignment, we may focus on subset of genes that are assigned to a category of interest, especially, hypomethylated categories (say, Category $1, 4, 7$) for breast cancer studies. Then those genes can be used for further study such as gene set analysis and gene pathway analysis. Since tens of thousands of genes are overwhelming for such analysis, our method plays a key role of narrowing the number of genes to produce an appropriate number of genes for following studies.

Acknowledgements

I am grateful to the Changyu Shen for helpful comments as well as to Kenneth Nephew for the data.

References

Ahuja, N., Mohan, A. L., Li, Q., Stolker, J. M., Herman, J. G., Hamilton, S. R., Baylin, S. B. and Issa, J. J. (1997). Association between CpG island methylation and microsatellite instability in colorectal cancer, *Cancer Research*, **57**, 3370.

Baylin, S. B. and Herman, J. G. (2000). DNA hypermethylation in tumorigenesis: Epigenetics joins genetics, *Trends Genetics*, **16**, 168–174.

Bird, A. (2002). DNA methylation patterns and epigenetic memory, *Gene Development*, **16**, 6–21.

Day, N. E. (1969). Estimating the components of a mixture of two normal distributions, *Biometrika*, **56**, 463–474.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society B*, **39**, 1–38.

Dwivedi, R. S., Qiu, Y. Y., Devine, J. and Mirkin, B. L. (2003). Role of DNA methylation in acquired drug resistance in neuroblastoma tumors, *Proceedings of Indian National Science Academy*, **69**, 111–120.

Esteller, M., Hamilton, S. R., Burger, P. C., Baylin, S. B. and Herman, J. B. (1999). Inactivation of the DNA repair gene O6-Methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia, *Cancer Research*, **59**, 793.

Fan, M., Yan, P. S., Hartman, F. C., Chen, L., Paik, H., Oyer, S. L., Salisbury, J. D., Cheng, A. S., Li, L., Abbosh, P. H., Huang, T. H. and Nephew, K. P. (2006). Diverse gene expression and DNA methylation pro thisfiles correlate with differential adaptation of breast cancer cells to the antiestrogens Tamoxifen and Fulvestrant, *Cancer Research*, **66**, 11954–11966.

Figueiredo, A. T. M. and Jain, A. K. (2002). Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 381–396.

George, C. (1985). An introduction to empirical bayes data analysis, *American Statistician*, **39**, 83–87.

Herman, J. G. (1999). Hypermethylation of tumor suppressor genes in cancer, *Seminars of Cancer Biology*, **9**, 359–367.

Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation, *New England Journal of Medicine*, **349**, 2042–2054.

Hinshelwood, R. A. and Clark, S. J. (2008). Breast cancer epigenetics: normal human mammary epithelial cells as a model system, *Journal of Molecular Medicine*, **86**, 1315–1328.

Hui, R., Macmillan, R. D., Kenny, F. S., Musgrove, E. A., Blamey, R. W., Nicholson, R. I., Robertson, J. F. and Sutherland, R. L. (2000). INK4a gene expression and methylation in primary breast cancer: Overexpression of p16INK4a messenger RNA is a marker of poor prognosis, *Clinical Cancer Research*, **6**, 2777.

Jeong, J., Li, L., Liu, Y., Nephew, K. P., Huang, T. H. and Shen, C. (2010). An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer, *BMC Medical Genomics*, **3**, 55.

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer, *Cell*, **128**, 683–692.

Jones, P. A. and Laird, P. W. (1999). Cancer-epigenetics comes of age, *Nature Genetics*, **21**, 163–167.

McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions*, John Wiley & Sons, New Jersey.

Muller, S., Fong, K. M., Maitra, A., Lam, S., Geradts, J., Ashfaq, R., Virmani, A. K., Milchgrub, S., Gazdar, A. F. and Minna, J. D. (2001). 5' CpG island methylation of the FHIT gene is correlated with loss of gene expression in lung and breast cancer, *Caner Research*, **61**, 3581.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family, *Scandinavian Journal of Statistics*, **1**, 49–58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families, *Communications in Statistics-Simulation and Computation*, **5**, 55-64.

Wang, X., Chao, L., Jin, G., Ma, G., Zang, Y. and Sun, J. (2009). Association between CpG island methylation of the WWOX gene and its expression in breast cancers, *Tumor Biology*, **30**, 8–14.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**, 95–103.

Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures, *Neural Computation*, **8**, 129–151.