Gene expression

Comments on ‘Bayesian hierarchical error model for analysis of gene expression data’

Xiao-Lin Wu1,†, Larry J. Forney1,* and Paul Joyce2

1Department of Biological Sciences, University of Idaho, Moscow ID 83844-3051, USA and
2Department of Mathematics and Statistics, University of Idaho, Moscow ID 83844-1103, USA

Received on January 30, 2006; revised on April 3, 2006; accepted on April 29, 2006
Advance Access publication May 26, 2006
Associate Editor: John Quackenbush

Cho and Lee (2004) proposed a Bayesian hierarchical error model (HEM) to account for heterogeneous error variability in oligonucleotide microarray experiments. They estimated the parameters of their model using Markov Chain Monte Carlo (MCMC) and proposed an F-like summary statistic to identify differentially expressed genes under multiple conditions. Their HEM is one of the emerging Bayesian hierarchical modeling tools that have been developed for the analysis of multiple-level data structures and variation in microarray gene expression data (Broet et al., 2002; Tadesse and Ibrahim, 2004; Cho and Lee, 2004). In this letter, we first discuss the significance of the HEM developed by Cho and Lee. Then, we re-derive the fully conditional distributions for gene and conditional effects, since we think that these two fully conditional distributions were not presented properly in their paper. Finally, we expand the HEM to deal with biological or/and experimental correlations in gene expression data. A FORTRAN 90 program was developed to implement our extended method and it is available from the authors upon request.

Significance of HEM. Many aspects of the HEM presented in the article of Cho and Lee (2004) are significant. The following are the two aspects that we believe are important in the analysis of microarray gene expression data. First, error variability of gene expression data is decomposed into two components, biological and experimental. The former reflects variation owing to biological sample heterogeneity that affects gene expression, while the latter largely comes from variation in sample preparation and sample analysis. Estimating these two sources of errors separately will not only improve estimation of model parameters but also facilitate a better understanding of patterns of over all gene expression.

Second, and more importantly, the use of a Bayesian HEM is an effective way to represent complex relationships that have many parameters, because it allows the parameters to be modeled separately at several levels and provides a link (S) between the data (D) and the unobserved parameters (θ) of interest. In multilevel modeling, the top layer is the data, p(D | S, θ), which is modeled by an appropriate likelihood function. The data are assumed to be generated by some process that depends on certain unknown parameters. Modeling the link or process is the middle layer, where the art of statistical modeling takes place. On the bottom layer are the prior distributions that represent a priori beliefs about these parameters. Following the Bayesian inference, unknown parameters are inferred from the joint distribution of the parameters given the data

\[
p(θ, S | D) \propto p(D | S, θ)p(S | θ)p(θ)
\]

Thus, Bayesian hierarchical modeling based on multivariate normal distributions effectively handles complex relationships that are governed by many parameters, and provides a tool for evaluating correlated gene expression data. For example, the model of Cho and Lee (2004) can be extended to allow for biological or/and experimental correlations. These will be discussed in more detail in the third section of this letter. Statistically, the estimation of these variables separately is biased when they are correlated (Johnston, 1972; Gianola and Sorensen, 2004). From a practical standpoint, ignoring correlations in gene expression data would result in high variability of statistical estimators and cause the reproducibility to deteriorate (Qiu et al., 2005).

Fully conditional distributions of gene (gi) and condition (ci) effects. We claim that Cho and Lee (2004) made a few mistakes when they derived the fully conditional distributions for the gene (gi) and condition (ci) effects. Let y be a vector containing all the data (xijk) , x a vector containing all the link variables (xijk) that correspond to expression intensities free from experimental errors, g a vector containing all the gene effects (gij) , c a vector containing all the condition effects (cij) , r a vector containing all the interaction effects (rij) of genes and conditions, and Δ a variance–covariance matrix with diagonal elements or variance (σ2ij) and zero off-diagonal elements or covariance. Following their model specification and prior distributions, the posterior of unknown parameters, given the data (y), is

\[
p(θ, S | D) \propto p(y | x, σ2)p(x | μ, g, c, r, Δ)p(μ)p(σ2)p(Δ)p(σ2)
\]

\[
\times \left[ \left( \sum_{i} \left( \sum_{j} \left( \sum_{k} x_{ijk} \right)^2 \right) \right)^{-1} \exp \left( \frac{-(\sum_{i} \left( \sum_{j} \left( \sum_{k} x_{ijk} \right)^2 \right)^{-1}}{2} \right) \right] \times \left( \prod_{i} \left( \prod_{j} \left( σ_{hi}^2 \right)^{-1} \right) \right) \times \exp \left( -\frac{\sum_{i} \sum_{j} \left( σ_{hi}^2 \right)^{-1} }{2} \right) \times \exp \left( -\frac{\sum_{i} \sum_{j} \left( σ_{hi}^2 \right)^{-1} }{2} \right) \times \exp \left( -\frac{\sum_{i} \sum_{j} \left( σ_{hi}^2 \right)^{-1} }{2} \right) \times \left( \prod_{i} \left( \prod_{j} \left( \left( σ_{hi}^2 \right)^{-(n-1)} \exp(-β_0 σ_{hi}^2) \right) \right) \right) \times \left( \left( σ_{hi}^2 \right)^{-(n-1)} \exp(-β_0 σ_{hi}^2) \right),
\]

(2)
Bayesian hierarchical error model for analysis of gene expression data

where \( \theta_L = \{ \mu, \sigma_r, \sigma_e, \Delta \} \) is a collection of unknown parameters determining the link \( \kappa \), with \( \mu, m_{ij}, \sigma_{e,ij}, \sigma_r^2, \sigma_e^2 \) and \( \sigma_e^2 \) being defined the same as \( \mu, m_{ij}, \sigma_{e,ij}, \sigma_r^2, \sigma_e^2 \) and \( \sigma_e^2 \) respectively, in the paper of Cho and Lee (2004). To ensure that our results are comparable with those of Cho and Lee (2004), we made the same assumptions regarding the prior distribution. That is, we also applied a flat prior to the grand mean \( \mu \) such that \( p(\mu) \) cancels out in (2) as a constant. [See Cho and Lee (2004) for a complete discussion of the implications of using flat priors.]

To derive the fully conditional distribution for an unknown parameter, we simply pick from (2) the component or components that are related to the parameter of interest. First, consider the grand mean. Since

\[
p(\mu | \cdot) \propto \exp \left( -\sum_i \sum_j \sum_k \alpha_i^{(j,k)} \left( \frac{1}{2} \left( x_{ijk} - g_i - c_j - r_k \right)^2 \right) - \frac{1}{2} \sum_i \sum_j \left( m_i \sigma_{e,ij}^2 \right) \right)
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_i \sum_j \left( m_i \sigma_{e,ij}^2 \right) \frac{1}{2} \sum_i \sum_j \left( m_i \sigma_{e,ij}^2 \right) \right)
\]

the fully conditional distribution of the overall mean, given the data and all unknown parameters, is

\[
\mu | \cdot \sim N \left( \frac{\sum_i \sum_j \sum_k \sigma_{e,ij}^2 (x_{ijk} - g_i - c_j - r_k)}{\sum_i \sum_j \left( m_i \sigma_{e,ij}^2 \right)}, \frac{1}{\sum_j \left( m_i \sigma_{e,ij}^2 \right)} \right)
\]

\[
\sim N \left( \frac{\sum_i \sum_j \sum_k (x_{ijk} - g_i - c_j - r_k)}{\Lambda}, \frac{1}{\Lambda} \right), \tag{3}
\]

where \( \Lambda = \sum_i \sum_j \left( m_i \sigma_{e,ij}^2 \right) \). In (3) and throughout this letter, we use \( ' | ' \) to mean 'conditional on the data and all unknown parameters except the one currently under consideration'. For the effect that corresponds to the \( i \)-th gene, we pick components in (2) that are related to \( g_i \), which leads to

\[
p(g_i | \cdot) \propto \exp \left( -\frac{1}{2} \sum_i \sum_j \left( \sigma_{e,ij}^2 \right) \frac{1}{2} \sum_i \sum_j \left( \sigma_{e,ij}^2 \right) \right) \times \exp \left( -\frac{1}{2} \sum_i \sum_j \sigma_{e,ij}^2 \right)
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_i \sum_j \sigma_{e,ij}^2 \right)
\]

Thus, the fully conditional distribution of \( g_i \) given the data and all other parameters is

\[
g_i | \cdot \sim N \left( \frac{\sum_i \sum_j \left( x_{ijk} - \mu - c_j - r_k \right)}{\sigma_g^2 + \sum_j m_j \sigma_{e,ij}^2}, \frac{1}{\sigma_g^2 + \sum_j m_j \sigma_{e,ij}^2} \right)
\]

\[
\sim N \left( \frac{\sum_i \sum_j \left( x_{ijk} - \mu - c_j - r_k \right)}{\Lambda_{ij}}, \left( \sigma_g^2 + \Lambda_{ij} \right)^{-1} \right), \tag{4}
\]

where \( \Lambda_{ij} = \sum_j \left( m_j \sigma_{e,ij}^2 \right) \). Similarly, since

\[
p(c_j | \cdot) \propto \exp \left( -\frac{1}{2} \sum_i \sum_j \left( \sigma_{e,ij}^2 \right) \frac{1}{2} \sum_i \sum_j \left( \sigma_{e,ij}^2 \right) \right) \times \exp \left( -\frac{1}{2} \sum_i \sum_j \sigma_{e,ij}^2 \right)
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_i \sum_j \sigma_{e,ij}^2 \right)
\]

the fully conditional distribution of \( c_j \), the effect that corresponds to the \( j \)-th condition, is

\[
c_j | \cdot \sim N \left( \frac{\sum_i \sum_j \left( x_{ijk} - \mu - g_i - r_k \right)}{\sigma_c^2 + \sum_j \left( m_j \sigma_{e,ij}^2 \right)}, \frac{1}{\sigma_c^2 + \sum_j \left( m_j \sigma_{e,ij}^2 \right)} \right)
\]

\[
\sim N \left( \frac{\sum_i \sum_j \left( x_{ijk} - \mu - g_i - r_k \right)}{\left( \sigma_c^2 + \Lambda_c \right)}, \left( \sigma_c^2 + \Lambda_c \right)^{-1} \right) \tag{5}
\]

where \( \Lambda_c = \sum_j \left( m_j \sigma_{e,ij}^2 \right) \). Comparing (3), (4) and (5) with the first three equations in Appendix 1 of the paper of Cho and Lee (2004), we found that they applied a common quantity \( \Lambda = \sum_j \left( m_j \sigma_{e,ij}^2 \right) \) for the three fully conditional distributions and this introduced systematic biases to the posterior estimates of model parameters. In our simulation study, we found this flaw to be dramatically downward-biased posterior estimations of both condition effects and gene effects, and it consequently affected posterior estimation of interaction effects and biological variances. Figure 1 illustrates how it affected posterior sampling of a randomly chosen gene effect toward zero. In contrast, using the corrected fully conditional distribution as shown in (4), we were able to obtain a posterior estimate for this gene effect that was close to its true value (i.e. 0.7923). This situation was representative of all gene and condition effects. However, we observed no significant influence of this flaw on the posterior estimation of \( \mu \) and \( \sigma_e^2 \). In other words, the posterior estimates of \( \mu \) and \( \sigma_e^2 \) were very close to their true values using either the algorithm of Cho and Lee (2004) or the corrected one that we propose. This is not very surprising if one considers
the restrictions \( \sum_i g_i = 0 \) and \( \sum_j c_j = 0 \) in our data simulation. When these restrictions hold, bias in the posterior estimation of \( g_i \) and \( c_j \) did not significantly affect the posterior sampling of \( a \) and \( \sigma^2_e \). In their paper, the conditional distribution \( x_{ijk} \) was derived with replicates, as shown below:

\[
x_{ijk} \mid \cdot \sim N \left( \frac{\sigma^2_e}{\sigma^2_e + \sigma^2_{n_{ijk}}} \left( \frac{n_{ijk}}{\sigma^2_e} + \frac{1}{\sigma^2_{n_{ijk}}} \right)^{-1} \right).
\]  

We think it was improper for them to reduce (6) to

\[
x_{ijk} \mid \cdot \sim N(\mu + g_i + c_j + r_{ij}, \sigma^2_{n_{ijk}}),
\]

when there is no experimental replication. By our reasoning, no replication is equivalent to putting \( n_{ijk} = 1 \), not \( n_{ijk} = 0 \) as stated in their paper, and \( \sigma^2_e = 0 \), implying that

\[
x_{ijk} = \left( \frac{1}{n_{ijk} + 1} \right) \sum_{l=1}^{n_{ijkl}} y_{ijkl} = y_{ijk}.
\]

**Modeling biological/experimental correlations.** Cho and Lee (2004) assumed independence for both biological and experimental errors, but these assumptions may be relaxed in the HEM. In this section, we expand their HEM to model biological or/and experimental correlation in gene expression data. First, consider correlations among experimental replicates, which may arise when experimental replicates are exposed to some common environmental factors yet collected at different times or locations. In these cases, the gene expression data may show temporal or spatial variation. In bacterial biofilms, for example, microorganisms undergo significant phenotypic changes during the transition from planktonic to biofilm forms, which reflects altered underlying gene expression and regulation as they adapt to form highly organized and structured sessile communities (Schembri et al., 2003; Watnick and Kolter, 2000) with enhanced resistance to antimicrobial treatments and host immune responses. Furthermore, the development of a biofilm may be tuned differently as a response to a nutrient gradient from the bulk media to the inside of the biofilm. Thus, gene expression at different locations in a biofilm may show spatial variation similar to geographic data in such a way that ‘everything is related to everything else, but near things are more related than distant things’ (Tobler, 1970). Biofilms are the dominant mode of expression and regulation as they adapt to form highly organized and structured forms, which reflects altered underlying gene expression and regulation as they adapt to form highly organized and structured sessile communities (Schembri et al., 2003; Watnick and Kolter, 2000) with enhanced resistance to antimicrobial treatments and host immune responses. Furthermore, the development of a biofilm may be tuned differently as a response to a nutrient gradient from the bulk media to the inside of the biofilm. Thus, gene expression at different locations in a biofilm may show spatial variation similar to geographic data in such a way that ‘everything is related to everything else, but near things are more related than distant things’ (Tobler, 1970). Biofilms are the dominant mode of growth for bacteria in the environment, and are of great concern in medical, environmental and industry settings. Therefore, considering the spatial correlation in a model would help better understanding the formation of biofilms as well as their altered biological activities.

Consider a balanced experimental design with observations on \( i = 1, \ldots, G \) genes under \( j = 1, \ldots, C \) conditions with \( k = 1, \ldots, m \) biological samples each measured in \( l = 1, 2, \ldots, n \) chips. Suppose that the chips represent \( n \) experimental replicates (e.g. layers of a biofilm). Let \( R_i \) be a \( n \times n \) spatial variance–covariance matrix such that the residuals \( e \sim N(0, I_{nm} \otimes R_i) \), where \( I_{nm} \) is an \( m \times m \) identity matrix (\( m^* = G \times C \times m \) and \( \otimes \) indicates the Kronecker product operation. Thus, in the posterior formulation as shown in (2), \( p(y \mid x, \sigma^2_e) \) is then replaced with \( p(y \mid x, R_e) \) and the latter is given as

\[
p(y \mid x, R_e) \propto |I_{nm} \otimes R_e|^{-1/2} \exp((y-x)'(I_{nm} \otimes R_e)^{-1}(y-x)).
\]

If we assume that the prior distribution of \( R_e \) to be an inverted Wishart distribution with scale matrix \( V_e \) with a degree of freedom parameter \( \kappa_e \) [denoted as \( \text{Wishart}^{-1}(\kappa_e, V_e) \)], it can be shown that posterior samples of \( R_e \) are sampled from the following:

\[
R_e \mid \cdot \sim \text{Wishart}^{-1}(m^* + \kappa_e, S_e + V_e),
\]

where \( S_e \) is a \( n \times n \) matrix with the (\( i, l \)) diagonal element being \( \sum_j \sum_r (y_{ijkl} - x_{ijkl})^2 \) and the (\( i, l \)) off-diagonal element being \( \sum_j \sum_r (y_{ijkl} - x_{ijkl})(y_{iljk} - x_{iljk}) \). We assumed inverted Wishart distributions instead of inverted multivariate gamma distributions as the prior distributions for the variance–covariance matrix for the sake of convenience in programming. Thus, each element in \( x \) is sampled from the following fully conditional distribution

\[
x_{ijk} \mid \cdot \sim N \left( \frac{\sigma^2_e}{\sigma^2_e + \sigma^2_{n_{ijkl}}} \left( \frac{n_{ijkl}}{\sigma^2_e} + \frac{1}{\sigma^2_{n_{ijkl}}} \right)^{-1} \right),
\]

where \( \mu^* = \mu + g_i + c_j + r_{ij} \) or \( \gamma^* \) is the (\( \ell, \ell' \)) element of matrix \( R^{-1} \), \( tr \) represents the trace matrix operation. Under the assumption of common experimental error variance (i.e. \( \gamma^0 = 0 \)), (8) is reduced to (6).

To model correlated biological errors, we similarly define a \( m \times m \) covariance matrix \( \Delta_m \) such that \( b_{ijk} \sim N(0, I_m \otimes \Delta_m) \) where \( I_m \) is an \( m \times C \times C \) identity matrix (\( C^* = G \times C \)). Biological covariance matters in analyzing gene expression data when samples are collected from a single experimental system. Then, \( p(x \mid \mu, g, c, r, \Delta) \) in (2) is replaced with the following

\[
p(x \mid \mu, g, c, r, \Delta) \propto |I_{C^*} \otimes \Delta_m|^{-1/2} \exp((x - \mu - g - c - r)'(I_{C^*} \otimes \Delta_m)^{-1}(x - \mu - g - c - r))
\]

Again, we assume the prior distribution for \( \Delta_m \) to be an inverted Wishart distribution with scale matrix \( V_m \) and degree of freedom parameter \( \kappa_m \), leading to the following fully conditional distribution of \( \Delta_m \):

\[
\Delta_m \mid \cdot \sim \text{Wishart}^{-1}(C^* + \kappa_m, S_m + V_m)
\]

where \( S_m \) is a \( m \times m \) matrix with the diagonal element \( \sum_j \sum_r (x_{ijkl} - x_{ijkl})^2 \) and the off-diagonal element \( \sum_j \sum_r (x_{ijkl} - x_{ijkl})(x_{iljk} - x_{iljk}) \). The fully conditional distributions for elements in \( g, c \) and \( r \) are derived as below:

\[
\mu \mid \cdot \sim N \left( \sum_j \sum_r \delta(\xi_{ijkl} + \Delta r) \sum_j \sum_r \delta(\xi_{iljk} + \Delta r), (\Lambda^*)^{-1} \right)
\]
 Briefly, the grand mean was given as a constant. Gene effects and interaction effects from 5 biological samples with 3 experimental replicates of each and expression data of 1000 genes, which were measured using Bayesian hierarchical error model for analysis of gene expression data

to show how our expanded model works, we simulated the decomposition of mean squared errors (MSE) of each parameter (say $g_i$) such that

$$ \text{MSE} = \text{var}(\hat{g}_i) + (\text{bias}(\hat{g}_i))^2. $$

where $\delta^{kk'}$ is the $(k,k')$ element in the inverse matrix $\Delta^{-1}$, $x_{ijk}(\mu) = x_{ijk} - g_i - c_j - r_{ij}$ and

$$ \Lambda^* = \sum_j \sum_k \sum_k \left( \delta^{kk} + \sum_{k'=1}^k \delta^{kk'} \right), $$

$$ g_j \cdot \sim N \left( \sum_k \sum_k \frac{\delta^{kk} x_{ijk}(\mu) + \sum_{k'=1}^k \delta^{kk'} (x_{ijk} + x_{ijk}(\mu))}{\sigma^2 + \Lambda^*}, \left( \sigma^2 + \Lambda^* \right)^{-1} \right). $$

where $x_{ijk}(\mu) = x_{ijk} - \mu - c_j - r_{ij}$ and $\Lambda_j^* = \sum_k \sum_k (\delta^{kk} + \sum_{k'=1}^k \delta^{kk'}).$

Fig. 2. 3D patterns of estimated (co)variance components among (a) five biological sample and (b) three experimental replicates.

$$ \text{var}(\hat{g}_i) = \frac{1}{T} \sum_{t=1}^T (\hat{g}_i - g_i)^2 $$

represented sampling variation and

$$(\text{bias}(\hat{g}_i))^2 = \left( \frac{1}{T} \sum_{t=1}^T (\hat{g}_i - g_i)^2 \right) $$

represented squared bias, and $\hat{g}_i$ was an estimate of $g_i$ at saved iteration $t=1, 2, \ldots, T$.

Our extended model facilitates estimating patterns of biological or/and experimental (co)variance or correlations (Fig. 2) whereas such information could not be inferred using the previous model (Cho and Lee, 2004). As mentioned before, biological or experimental correlations could be of considerable interest and necessary
to help us better understand the nature of correlated gene expression. For example, suppose the pattern of experimental (co)variance in Figure 2b represents the environmental correlations in a biofilm. We might conclude that nutrient differences result in higher correlations between the inner and bottom layers than between the top and the other two layers.

Further, we investigated how different model specifications about biological/experimental covariance (i.e. zero versus non-zero) would affect the posterior estimates of parameters of interest. Ignoring experimental covariance in the model did not lead to significant differences in parameter estimation (data not shown). However biological covariance components are important for estimating gene and interaction effects. If the biological covariance components are ignored then gene effects and interaction effects will be estimated with large biases. However, grand mean, condition effects and the common residual variance appear to be more robust, and gave similar results under either model specification about biological (co)variance. As shown in Table 1, considering non-zero biological (co)variance in the model considerably decreased estimation biases of gene and interaction effects. Squared biases, and consequently MSE, of these parameters were smaller in the analysis considering biological covariance components than by ignoring them. Given the variation that occurs during the collection of gene expression data, posterior estimation of gene and interaction effects was better, in terms of MSE, sampling variation and squared biases, when more samples were used (Table 1). This result would challenge the present practice where a very limited number (say three) of samples are used in most gene expression DNA array experiments.

Finally, we show how model specifications about biological covariance would affect the estimation of gene expression differences, \( (\mu + g_i + c_j + r_{ij}) - (\mu + g_i + c_j) + (r_{ij} - r_{ij}') \) between various conditions (Fig. 3), which is of importance in identifying differential gene expression. Clearly, the estimated gene expression differences were in better agreement with true differences when considering non-zero biological covariance as compared to ignoring them in the model. In the former case, the \( R^2 \) value for the linear regression of the estimated and true expression differences was higher and the regression slope is closer to 1.0. The low slope in Figure 3b suggests an obvious systematic bias, which resulted from ignoring non-zero biological covariance. On average, the higher the biological correlations were, the greater the difference. In 10 independent simulations starting with different datasets, we found that 6.5–38.7\% of the top 100 differentially expressed genes could vary when one assumes zero or non-biological covariance, respectively.

In conclusion, we extended the model of Cho and Lee (2004) to allow the analysis of gene expression data with biological/environmental correlations, which not only facilitate estimating biological/expression correlation patterns, but it also improved estimation of model parameters and consequently identifying differentially expressed genes in DNA array experiments. A FORTRAN 90 program was developed to implement our extended method and it is freely downloadable at http://www.ibest.uidaho.edu/tools/banal_array/
REFERENCES