Functional Empirical Bayes Methods for Identifying Genes with Different Time-course Expression Profiles

Hong, Fangxin
Li, Hongzhe

Publication Date
2004-11-10

Functional Empirical Bayes Methods for Identifying Genes with Different Time-course Expression Profiles

Fangxin Hong,[1] **and Hongzhe Li**[2,*]

[1] Departments of Statistics, University of California, Davis, California 95616, U.S.A.

[2] Rowe Program in Human Genetics, School of Medicine
University of California, Davis, California 95161, U.S.A.

[*] *email:* hli@ucdavis.edu

Address for correspondence:
Hongzhe Li, Ph.D.
Rowe Program in Human Genetics
University of California Davis School of Medicine
Davis, CA 95616-8500, USA
Tel: (530) 754-9234; Fax: (530) 754-6015
E-mail: hli@ucdavis.edu

Summary

Time course studies of gene expression are essential in biomedical research to understand biological phenomena that evolve in a temporal fashion. Microarray technology makes it possible to study genome-wide temporal differences in gene expression profiles between different experimental conditions/groups. In this paper, we introduce a functional hierarchical model and empirical Bayes approach to model gene expression trajectories over time and to detect temporally differentially expressed (TDE) genes. Monte Carlo EM algorithm is developed for estimating both the gene-specific parameters and the hyperparameters. We use the posterior probability based false discovery rate (FDR) criterion to identify the TDE genes in order to control for the over FDR. We illustrate the methods by using both simulated data sets and a data set from a microarray based gene expression time course study of *C. elegans* developmental processes. Simulation results suggested that the procedure have low false discovery rate but could potentially have high false negative rate when the noise variance is relatively large. Results from both simulations and analysis of *C. elegans* data indicated that the procedure performed better than the two-way ANOVA in identifying TDE genes between the dauer exit process and starved $L1$ worms response to feeding process.

Key Words: Hierarchical model, Empirical Bayes, Gibbs-sampler, B-spline, False discovery rate, gene expression.

1 Introduction

Since many important biological systems or processes are dynamical systems, it is important to study the gene expression patterns over time in the genomic scale. The DNA microarray technologies make it possible to monitor changes in gene expression over time during these biological processes (Spellman *et al.*, 1998; Diaz *et al.*, 2002; Chuang *et al.*, 2002). One important application of such microarray time course (MTC) gene expression experiments is to identify genes that are differentially expressed temporally between two MTC gene expression experiments. We call these genes temporally differentially expressed (TDE) genes. Comparing to gene expression study at one time point, such MTC studies can potentially identify more genes which are differentially expressed (Chuang *et al.*, 2002).

While many statistical methods have been developed in recent years for identifying differentially expressed genes, the focus of these methods is on identifying genes with different mean genes expression level at one single time point (Lonnstedt and Speed, 2002; Efron *et al.*, 2001; Newton *et al.*, 2003). In comparison, methods that are specifically developed for identifying genes with different expression patterns during two MTC experiments are less developed. One common approach is to use the two-way repeated ANOVA treating group, time and their interactions as factors (Park *et al.*, 2003; Wang and Kim, 2003) and to identify the TDE genes by testing the interactions. However, due to the number of replications is usually small, such analysis may not have power to detect truly differentially expressed genes. Xu *et al.* (2002) proposed to model the time course expression with lower order polynomial functions of time for each gene separately, including the interaction terms between times and group indicator. A particular gene is identified to be differentially expressed among groups if any of its coefficients of the interaction terms is significantly different from zero. To control false-positives on the global scale of all genes, they used Bonferroni correction to adjust for multiple comparison. However, simple low order polynomials may not provide the flexibility of modelling more complicated gene expression profiles. In addition, modelling and testing each gene separately makes simultaneous inferences of all genes difficult. Guo *et al.* (2003) formulated MTC data as longitudinal measurements and defined a robust Wald score statistic to detect genes with temporal changes in expression. The methods can account for within-subject correlation of gene expression levels over time. They then used the SAM method (Tusher *et al.*, 2001) on the derived scores to identify those genes with non-constant means over time. Again due the small number replications, the asymptotic variance of the Wald score statistic used in Guo *et al.* (2003) may not be valid. In addition, the methods in Guo *et al.* (2003) was developed for identifying genes with change of expression levels over time, not for identifying TDE genes.

Among the methods for identifying differentially expressed genes, the empirical Bayes methods have gained much popularity due to the fact that they can effectively pool data from different genes (Efron *et al.* 2001; Lonnstedt and Speed, 2002; Newton *et al.*, 2003). In this paper, we propose an empirical Bayes method for identifying the TDE genes between two experimental conditions in the framework of hierarchical models. In our model formulation, we treat the observed time course gene expression data as samples from the true underlying continuous gene expression trajectories and propose to use cubic B-splines (De Boor, 1978) to approximate the true gene expression trajectories. The cubic B-splines provide a flexible curve-fitting methods and have been shown to effectively model various gene expression time course profiles (Luan and Li, 2003; Luan and Li, 2004). Based on the empirical Bayes methods, data from all the genes are combined together into estimate of the posterior probability of differential expression for each individual gene. We propose to use the false discovery rate (FDR) procedure of Story (2003) for identifying the TDE genes while controlling for overall false discovery rates.

The rest of the paper is organized as follows: we first present the hierarchical model and empirical Bayes methods for the problem and the Monte Carlo EM (MCEM) algorithm for estimating the model parameters. We then present simulation results to evaluate the proposed models and the MCEM algorithm. We apply the methods to a *C. elegans* developmental data set (Wang and Kim, 2003) which collect the expression profiles of thousand genes during dauer recovery experiment and starved L_1 worm response to feeding experiment. Finally, we give a brief discussion on the methods and results.

2 Statistical Model and Empirical Bayes Inference

A typical design of MTC studies for comparing the gene expression profiles between two experimental groups can be summarized as in Table 1, in which the two experiment groups are indexed by $i = 1, 2$. Assume that there are K_i ($K_i \geq 1$) replications for the ith group. For each replication, the gene expression levels are measured at T different time points, $t_1, \cdots, t_T$. Assume that there are a total of n genes ($j = 1, ..., n$) on each array slide. Let Y_{jikt} be the log gene expression level for jth gene at time t in kth replication under the ith condition, and let $Y_{jik} = \{Y_{jik1}, \cdots, Y_{jikT}\}$ be the vector of observed gene expression levels over T time points.

2.1 *Functional Hierarchical Model Based on Basis Expansion*

Let $Z_i = 0, 1$ be the indicator for the two experimental groups. We consider data of the form,

$$Y_{jikt} = f_{ji}(t_{jit}) + \sigma_{jit}\epsilon_{jikt} \tag{1}$$

where f_{ji} is the true gene expression function for the jth gene under condition i, σ^2_{jit} is the gene, experiment and time-specific variance, ϵ_{jikt} is the error term with mean 0 and variance 1. Under this model, we assume that the measured log-expression level at time t is the true gene log-expression level plus noise. In practice, since the number of measured time point T is usually small, it is often difficult to estimate the true gene expression function f_{ji} by any nonparametric function. Instead, we assume that the true gene expression trajectory can be modeled by basis expansion, i.e., we assume that

$$f_{ji}(t) = \sum_{l=1}^{p}(\beta_{lj} + Z_i d_{lj})B_l(t_{jit}), \tag{2}$$

where B_l is the basis function, $l = 1, \cdots, p$, β_{lj} is the corresponding coefficient of the lth basis for the jth gene under the condition when $Z_i = 0$, and d_{lj} measures the difference in coefficient at the lth basis function between the two conditions. Note that if $d_{lj} = 0$ for all $l = 1, \cdots, p$ then there is no difference in gene expression trajectories between the two conditions for gene j. If, in addition, we assume that

$$\epsilon_{jik} = \{\epsilon_{ijlt}\}_{t=1}^{T} \sim MVN(corr_{ji}),$$

then

$$\begin{aligned} Y_{jik} &= \{Y_{jikt}\}_{t=1}^{T} | (\mu_{ji}, \Sigma_{ji}) \sim MVN(\mu_{ji}, \Sigma_{ji}) \\ \mu_{ji} &= \{\sum_{l=1}^{p}(\beta_{lj} + Z_i d_{lj})B_l(t_t)\}_{t=1}^{T} \end{aligned} \tag{3}$$

Since the number of replications in a typical microarray time-course gene expression is usually small, it is difficult to estimate the gene expression for each gene separately. Instead we take a hierarchical modeling approach by assuming priors for the difference of the coefficients, $d_j = \{d_{1j}, \cdots, d_{pj}\}$. One possibility it to assume an independent 0-normal mixture prior for each d_{lj} across all the genes, i.e., for each $l = 1, \cdots, p$,

$$\begin{aligned} e_{lj} &\sim Bernoulli(\pi_l) \\ d_{lj} &\sim N(0, e_{lj}\sigma_l^2), \end{aligned} \tag{4}$$

which implies that there exists a Bernoulli random variable e_{lj}, if $e_{lj} = 0$, $d_{lj} = 0$, indicating no difference between the two gene expression trajectories at the lth basis function. On the other hand, if $e_{lj} = 1$, d_{lj} is generated from a normal distribution with mean 0 and variance σ_l^2. We call the model (3 together with prior (4) the independent prior (IP) model. This model includes both the hyperparameters $\theta_h = \{\pi_l, \sigma_l^2, l = 1, ...p, H, q\}$ and gene-specific parameters $\theta_g = \{\beta_{lj}, l = 1, ..., p, j = 1, ..., n\}$.

Alternatively, we can assume

$$d_j = \{d_{1j}, d_{2j}, \cdots, d_{pj}\}' \sim (1-\pi)\delta(0) + \pi MVN(0, A), \tag{5}$$

treating the difference vector $\{d_{1j}, \cdots, d_{pj}\}$ as a p-dimensional random vector which follows a mixture distribution of vector 0 and multivariate normal distribution with mean 0 and variance-covariance matrix A. This is equivalent to assuming that for gene j, there is a Bernoulli random variable e_j with probability of π being 1, and if $e_j = 0$, $d_j = 0$, otherwise d_j follows $MVN(0, A)$, i.e.,

$$d_j | e_j \sim \begin{cases} 0 & \text{if } e_j = 0 \\ MVN(0, A) & \text{if } e_j = 1 \end{cases}$$

We call the model (1) together with prior distribution (5) the joint-prior (JP) model. Comparing to the IP model, the JP model allows potential dependency in the prior distribution of changes across different basis functions.

To finish the specification of the hierarchical model, we assume an inverse Wishart prior for the variance-covariance matrix of the error terms, i.e.,

$$\Sigma_{ji}^{-1} \sim W(B_i, q_i).$$

With additional assumptions, the distribution of gene-specific covariance matrix Σ_{ji} can be simplified. For example, we can assume that the gene specific error covariance to be diagonal as

$$\Sigma_{ji} = diag\{\sigma^2_{ji1}, \cdots, \sigma^2_{jit}, \cdots, \sigma^2_{jiT}\}$$

and each σ^2_{jit} follows an inverse gamma distribution $IG(a_t, b_t)$. We can further assume that the variance at all time points are the same for a given gene but different from gene to gene, i.e., $\sigma^2_{ji1} = \cdots = \sigma^2_{jiT} = \sigma^2_{ji}$, and assume $\sigma^2_{ji} \sim IG(a_i, b_i)$. The simplest model is to assume that error variances are the same for all the genes at all the time points. The specification of the covariance depends on the design of the experiments and also the availability of replications. For example, if the number of replications is small or zero, it is difficult to estimate gene-specific variance, in which case we may want to assume a constant variance across all the genes.

2.2 *Parameter Estimation and Inference*

We propose to estimate both the gene-specific parameters θ_g and the hyper-parameters θ_h in the prior distributions by maximizing the marginal likelihood of the observed data. However, directly maximizing such marginal likelihood is difficult, we therefore employ the Monte Carlo EM algorithm for obtaining these parameter estimates, where Gibbs sampling is used in the E-step to approximate the required expectations (Carlin and Louis, 1996). Details of the MCEM algorithm and the conditional distributions in the Gibbs steps are given in the Appendix.

After obtaining the parameter estimates, we can calculate the posterior probability or log posterior odds for gene j to have different gene expression trajectories, i.e., to be a TDE gene,

$$P_j = 1 - Pr(e_j = 0|Y),$$

and

$$B_j = \log \frac{1 - Pr(e_j = 0|Y)}{Pr(e_j = 0|Y)},$$

for both the IP and the JP models, where in the IP model, $e_j = \{e_{1j}, \cdots, e_{pj}\}$. Similarly, for the IP model, we can calculate the posterior probability and posterior log odds of having different coefficients for each basis function in the mean model as

$$P_{lj} = Pr(e_{lj} = 1|Y),$$

$$B_{lj} = \log \frac{Pr(e_{lj} = 1|Y)}{Pr(e_{lj} = 0|Y)},$$

for $l = 1, \cdots, p$. These basis-specific posterior values can be used to classify gene expression patterns for those TDE genes. Furthermore, the estimated posterior mean of the random effect d_{lj}, which is denoted as M_{lj}, can be used to assess the magnitude of the expression profile change.

2.3 *False discovery rate*

Based on the estimated marginal posterior probability $P_j, j = 1, \cdots, n$, we can select a list of genes by ranking from smallest to largest by P_j and cutting the list at some point chosen beforehand. Alternatively, the cutoff probability κ can be selected by controlling the overall FDR. The FDR has emerged in the context of multiple hypothesis testing as a practical object to be controlled (Benjamini and Hochberg, 1995). The focus of most recent work on FDR has been based on the p-values, not the posterior probabilities. However, Newton *et al.* (2003) noticed that the posterior probability, which cause the gene to be selected or not, also measures the probability of type I error if a given gene is selected. Following the notation used in Newton *et al.* (2003), let

$$J(\kappa) = \{j \in \{1, 2, ..., m\} : P_j \geq \kappa\}$$

denote the list of genes identified. Conditionally upon the data, and in the context of the model, the expected number of type I errors (i.e., false discoveries) is

$$E[\sharp FD|data] = \sum_{j \in J(\kappa)} (1 - P_j).$$

Typically we can find κ as large as possible so that $J(\kappa)$ is not empty and also

$$E[\sharp FD|data]/|J(\kappa)| \leq \alpha$$

for some target error rate α, where $|J(\kappa)|$ is the size of the list. The left-hand side is similar to the positive FDR (Story, 2003), except that the expectation is conditional on data. By plugging in the estimated posterior probability, the above equation can be approximated by

$$\sum_{j \in J(\kappa)} (1 - P_j)/|J(\kappa)| \leq \alpha.$$

The cutoff point κ is found conditionally on data, either to make a list of some fixed size or to approximately achieve the target error rate α. However, as noted by Newton *et al.* (2003), the control of the false discovery rate at α is only approximate because it rests on the fitted probability model being an accurate approximation of the data-generating mechanism.

3 Simulation Studies

In this section, we present results based on simulated data to evaluate the MCEM algorithm and the results on identifying TDE genes. We also investigate by simulations how error variance affects the identification results from the proposed methods and from the ANOVA method.

3.1 *Simulations based on B-spline models and parameter estimates*

We simulated several data sets based on the proposed IP model. We assume that the true gene expression trajectories can be modeled by cubic B-splines with six basis functions. We generated gene expression data at $T = 19$ equal-spaced time points for 500 genes. The coefficients of the baseline gene expression level β_{lj} are grouped into 10 groups and are randomly selected from a uniform distribution $U(-2, 2)$. For each condition, two replications were simulated, i.e., $K_1 = K_2 = 2$. In order to evaluate the proposed MCEM algorithm for parameter estimates, for a given set of true parameters, we simulated fifty replications based on the true IP model. The error variance is fixed at $\sigma_e^2 = 0.03$ across all the genes and all the time points.

For the first model, we assume that the true values for the prior probabilities and basis-wise variances are

$$\begin{aligned}\text{Model 1:} \qquad \pi &= (\pi_1, ..., \pi_6) = (0, 0, 0, 0, 0.1, 0.1),\\ \sigma^2 &= (\sigma_1^2, ..., \sigma_6^2) = (0, 0, 0, 0, 2, 2), \end{aligned} \tag{6}$$

This model assumes that the gene expression differences are only on the last two basis functions. Figure 1 shows the gene expression profiles for ten simulated genes, which are quite typical profiles we normally observe.

The parameter estimates from the MCEM algorithm resulted in $\pi_1 = \cdots = \pi_4 = 0$ for the first four basis function, indicating no changes in the gene expression profiles on the first

4 basis functions. Figure 2 (a) shows the scatter plot of the MCEM estimates of π_5 and π_6 versus the estimates based on the true values of e_{5j} and e_{6j} for $j = 1, \cdots, 500$ for a total of 50 replications. Similarly, Figure 2 (b) shows the scatter plot of the MCEM estimates of σ_5^2 and σ_6^2 versus the estimates based on the true values of d_{5j} and d_{6j} for $j = 1, \cdots, 500$. Both plots indicate that the MCEM algorithm estimates the hyperparameters reasonably well. In general, we observe that the EM estimates of the prior probabilities are smaller than the estimated based on the realized values of the variables e_{5j} and e_{6j}. On the other hand, the EM estimates of the variances in the hyperparameters are usually larger than those estimated by the realized values of d_{5j} and d_{6j}. This is not surprising since even when e_{5j} or e_{6j} is 1, the realized values of d_{5j} or d_{6j} can still be small.

To further demonstrate the proposed model and methods, we provide some more detailed analysis for one simulated data set. Figure 3 (a1) shows the MCEM estimates of the Spline coefficients $\beta_{lj}, l = 1, \cdots, p, j = 1 \cdots, n$, versus the true values, indicating that the algorithm estimates the B-spline coefficients and therefore the gene expression trajectories very well. Figure 3 (b1) plots the estimated posterior mean of d_{lj} versus the true realized values of d_{lj}, for $l = 5, 6$. We observed that for large d_{lj}, the posterior means of the random effects are very close to the true realized values. However, for small values of d_{lj}, the posterior means of the random effects shrink towards zero, which implies that for some genes with small true d_{jl}, the posterior probabilities can be small. Figure 3 (c1) shows the log posterior odds B_j versus the maximum of the posterior mean of d_{lj}, where for genes with large posterior odds, a ceiling of 35 was applied. This plot shows that the genes with large posterior odds tend to have large absolute value of posterior mean of the difference d_{lj} and therefore large difference in gene expression trajectories.

For the second model, we assume

$$\begin{aligned} \text{Model 2:} \qquad & \pi = \{0.05, 0.1, 0.1, 0.2, 0.1, 0.1\}, \\ & \sigma^2 = \{1, 2, 2, 3, 2, 2\}. \end{aligned} \tag{7}$$

Under this model, there is a probability of difference on each of the six basis functions used to characterize the expression profiles. The third model assumes that half of the genes with the expression change occurring only at the first two bases, and the other half only at the last two bases, with the following prior probabilities and basis-specific variance,

$$\begin{aligned} \text{Model 3:} \qquad & \pi = \{0.1, 0.1, 0, 0, 0, 0\}, \sigma^2 = \{2, 2, 0, 0, 0, 0\}, \\ & \pi = \{0, 0, 0, 0, 0.1, 0.1, \}, \sigma^2 = \{0, 0, 0, 0, 2, 2\}. \end{aligned}$$

For both models, we simulated a total of 800 genes for each replication. The results indicated that the MCEM algorithm can estimated both the gene-specific parameters and the hyperparameters very well. For the third model, the estimated prior probability for basis 1,2,5,6 are close to 0.5. As examples, the top panel of Figure 3 shows results for one data set simulated under the Model 2, and the bottom panel of Figure 3 shows results for one data set simulated under the Model 3. We observe that the estimated gene-specific B-spline coefficients are close to the true values and the posterior mean of gene specific random effects are close to the observed values. In addition, we also observed the V-shaped plot of the log posterior odds B_j versus the maximum of the posterior mean of d_{lj}. These results indicate that the proposed MCEM algorithm can indeed estimate the parameters well and that large estimated posterior probabilities indeed imply that the genes have large difference in expression profiles across times.

3.2 *Effect of error variance and number of replications on identifying TDE genes*

The simulation results presented in previous section indicate that the MCEM algorithm can indeed be applied to estimate the model parameters when the error variance σ_e^2 is small. We now examine how error variance affects the results on identifying TDE genes. We simulated 10 identical data based on the true parameters as in Model 1 in previous section with the same realized values of e_{lj} and d_{lj}, but we simulated the error terms from 10 different true error distributions with variance σ_e^2 ranging from 0.01 to 0.10. The estimated prior probability of π_5 and π_6 based on realized e_{5j} and e_{6j} are 0.101 and 0.104, which are close to the true value of 0.1. Similarly, the estimated prior variance σ_5^2 and σ_6^2 based on realized d_{5j} and d_{6j} are 2.05 and 2.1, which are also close to the true value of 2.0. Figure 4 (a) and (b) show the estimated p_5 (p_6) and σ_5^2 (σ_6^2) as the error variance increases from 0.01 to 0.10. We observe that as the error variance increases, the EM estimate of prior probability decreases, but the EM estimate of the variance of the differences increases. The results are not surprising since large noise tend to mask the true difference in gene expressions, which results in smaller estimate of π_5 and π_6.

The effects of error variance on the parameter estimates also directly translate into the results on identifying TDE genes. For 97 true TDE genes we simulated, Table 2 presents the identification results for the number of replications of $K_1 = K_2 = 2$ and $K_1 = K_2 = 10$ and for different error variance using a FDR of 0.01. As the error variance increases, the method identifies a smaller number of the true TDE genes. However, we observe that increasing the number of replications helps to identify more true TDE genes. As a comparison, we also report in this Table the identification results based on two-way ANOVA. We observed that the proposed methods indeed identified more true TDE genes than the ANOVA method, especially when the number of replications is small. In addition, both methods resulted in 0 false identification. This is also expected, since we set FDR=1% and there are only 500 genes.

3.3 *Simulating data from other models*

We also simulated a data set of $n = 500$ genes with $T = 19$ and $K_1 = K_2 = 2$ based on the following model,

$$y_{jikt} = \sum_{l=1}^{p} \beta_{lj} B_l(t) + Z_i d_{jt} + \epsilon_{jikt}$$

where $p = 6$, $B_l(t)$ is the B-spline basis function, β_{lj} is the same as in previous simulations and $\mathrm{var}(\epsilon_{jikt}) = 0.1$. We simulated 100 genes with $d_{jt} \neq 0$, including

- 50 genes with Constant difference: $d_{jt} = d_j$, $d_j \sim N(0, 3)$
- 25 genes with Linear difference: $d_{jt} = \alpha_{j0} + \alpha_{j1} t$,
  $\alpha_{j0}, \alpha_{j1} \sim Unif[-0.2, 0.2]$
- 25 genes with a step difference: $d_{jt} = \alpha_{j0} I(t < T_j) + \alpha_{j1} I(t \geq T_j)$
  $\alpha_{j0}, \alpha_{j1} \sim N(0, 2)$ and $T_j \sim Unif[2, 18]$.

For this data set, for FDR=0.01, the IP model with six B-spline basis functions identified 82 true TDE genes with no false positive, the JP model with six B-spline basis functions identified 84 true TDE genes with no false positive. This simulated data set indicates that the proposed models work well in identifying genes with linear difference. Note that in such case, we would expect that the ANOVA perform equal well. Indeed, the ANOVA model identified 82 true TDE genes with one false positive.

4 Application to *C. elegans* Developmental data

4.1 *Description of the data set*

When conditions are not favorable, such as low amount of food, high population density and temperature, first larval stage ($L1$) *C. elegans* worm can develop into dauer larvae to maximize survival. In this facultative stage, worms become developmentally arrested, non-feeding, stress-resistant and long-lived. These dauers continue development when conditions become favorable, particularly under the condition of a high food to pheromone ratio and low temperature. On the other hand, the $L1$ worms will also arrest in the absence of food. However, the arrested $L1$ larvae do not display dauer like properties but will continue development with the addition of food. Wang and Kim (2003) reported a cDNA microarray time course gene expression study on the dauer recovery and $L1$ starvation responses to feeding, in which they measured gene expression levels of about 17,088 genes at $0,1,\cdots$ and 12 hrs after feeding. For each time point, there are three or four replicates. The biological question is to identify genes that are related to common feeding program, i.e. those genes with similar expression patterns over time and dauer-recovery specific genes, which are the genes with different expression profiles over time.

Wang and Kim (2003) first identified 2430 genes which change their expression during the dauer exit time course by using a standard one-way ANOVA. They further analyzed those 2430 genes with two-way mixed-effects ANOVA and identified 1984 genes to be differentially expressed between dauer exit and $L1$ starvation time course (p-value< 0.05 from ANOVA).

4.2 *Results of analysis*

We first applied the JP model with gene- and time- specific error variance to fit the data in order to identify genes with different expression patterns between the dauer recovery experiment and starved $L1$ response to feeding experiment. We treated dauer recovery as condition

1 ($i = 1, Z_i = 0$) and starved $L1$ response to feeding as condition 2 ($i = 2, Z_i = 1$). Due to the small number of time points measured, we used B-splines with four basis functions to approximate the gene expression trajectories. The MCEM algorithm converged after 50 steps. The estimated parameters in the prior distributions are $\pi = 0.42$, $a = 2.64$ and $b = 2.28$ in the inverse gamma prior for the variance, and

$$A = \begin{pmatrix} 3.34 & 0.93 & 0.96 & 0.75 \\ & 5.36 & 0.36 & 1.26 \\ & & 1.83 & 0.49 \\ & & & 1.21 \end{pmatrix}$$

These parameter estimates imply that prior probability of being a TDE gene is 0.42. Our method identified 1011 and 1049 genes for FDR=0.01 and 0.05 respectively (see Table 3) . As examples, Figure 5 shows the observed mean expression profiles and the fitted smooth curves for nine TDE genes we identified. Clearly, the smoothed curves fit the data well. The four genes in the first column of Figure 5 show different expression at the end of the time course. On the other hand, the four genes in the second column of Figure 5 have different expression at the start of the time course. The genes in the last columns have different expression profiles throughout the whole time course. Figure 6 shows the observed and fitted gene expression data for nine genes with the smallest posterior probabilities of being TDE. Clearly, the gene expression profiles of these 9 genes are very similar. We also see that the fitted smooth gene expression profiles agree with the observed data very well for the 12 genes in Figure 5 and the 9 genes in Figure 6.

We also performed analysis by assuming the JP model but with constant variance for the error terms. The estimated prior probability of π is 0.38 and $\sigma_e^2 = 0.26$. For a FDR=0.01 or 0.05, the genes identified by this model are essentially the same as those identified by the model assuming time- and gene-specific variance (see Table 3).

Wang and Kim (2003) applied a two-way ANOVA with replication-specific random effect to each gene and obtained the p-value of having different expression profiles by testing the time by group interaction for each gene. They identified 1681 and 1949 genes for FDR=0.01 and 0.05. These numbers are very different the numbers of TDE genes identified by our methods. There are a large number of genes which were identified as TDE genes by ANOVA but not by our proposed empirical Bayes method. The 800 - 900 genes identified by both methods indeed show clear differential expression patterns over times. However, for the genes that were identified by ANOVA but not our methods, we examined their expression profiles under the two conditions and observed apparent similar expression profiles between the two time course after smoothing the data. Figure **??** present the observed and fitted data for 40 randomly selected genes that were identified as TDE genes by ANOVA only. These genes do not seem to have differential expression patterns over time. Plots for additional genes identified by ANOVA only can be found in our website, http://dna.ucdavis.edu/∼hli/Ebspline.html. One possible reason that ANOVA identified so many genes is that the F-test for interaction is very sensitive to the normality assumption and outliers.

We also performed analysis using B splines with five basis functions and obtained almost identical results on genes identified. However, based on the deviance information criteria (DIC) (Spiegelhalter *et al.*, 1998), the two models give very similar fit, with score of 3.19×10^5 and 3.10×10^5 for the model with 4 basis functions and 5 basis functions, respectively. Note that the B-spline function with 4 basis function is equivalent to the third-order polynomial function.

5 Discussion and Conclusions

We have proposed a hierarchical model and an empirical Bayes method to identify genes with temporal differential expression patterns based on microarray time course gene expression data. The method utilizes information from all the genes to estimate the posterior probability and posterior log-odds of being TDE for each gene. These posterior probabilities are then used for identifying the TDE genes in the framework of FDR. Simulations and application to real data set indicated that the methods are able to identify the TDE genes with low FDRs. The proposed model can be applied to both replicated or non-replicated time-course gene expression comparisons. However, our simulation also indicated that when the error variance is large and the number of replications is small, real TDE genes can be missed.

Our proposed methods assume that the true gene expression trajectories over time are continuous and smooth and enough samples are taken to approximate these trajectories. For simplicity, we used B-splines with pre-specified number and location of knots to approximate and characterize these gene expression trajectories. However, it should be emphasized that other curve-fitting procedure such as smooth splines and other basis functions can be applied in the same modeling framework. Using expansions in basis functions have been commonly used in both functional and longitudinal data analysis (Ramsay and Silverman, 1997). Our model can be interpreted as searching for TDE genes in the transformed functional spaces formed by the basis functions, so it serves as a way of dimension reduction. Another advantage of the proposed method is that it does not require the same sampling times between the two experimental conditions. We did not extensively study the issue of knots selection for B-splines, except that we applied the DIC to select between two different models in our analysis of the *C. elegans* data set. The location of knots may also be driven by our prior knowledge of the biological process. For typical microarray time course gene expression data, we found that B-splines with small number of evenly spaced knots fit the observed data quite well (Luan and Li, 2003; Luan and Li, 2004). Of course one should always check how well the curves fit the data. While it is an advantage to estimate the true gene expression trajectories by B-splines, one limitation of the proposed methods is that when the sampled time points are small, the true gene expression trajectories can not be modeled very well. In this case, we can develop a vector based empirical Bayes approach in a similar modeling framework. We are currently pursuing this approach.

In our proposed hierarchical model, we assume the prior probability of π to be unknown parameter and estimate them using the data. If the number of replications is small, there might be numerical instability in estimating these parameters. Alternatively, we can fix these prior probabilities to some reasonable values. Of course the posterior probabilities cannot be used with FDR to select the TDE genes. However, we can still rank the genes based on these posterior probabilities. For example, if we fix the prior probability at $\pi = (0, 0.5, 0.1, 0.5)$ in our analysis of the *C. elegans* data set, the top 1900 genes identified are the same as using the estimated prior probabilities. We also assume that the difference in coefficients of B-spline basis functions follows a normal distribution across the genes. If some genes have very extreme difference in expression profiles between the two conditions, distributions with long tails such as Cauchy or Laplace distribution may provide more suitable model for the data. How different distributional assumption affects the results on genes identified deserves future investigation.

In summary, we have introduced a B-spline based hierarchical model for identifying genes with different expression patterns over time. The methods have been successfully applied to analysis of the *C. elegans* developmental data set in identifying both dauer recovery specific genes and genes that the related to common feeding programs in *C. elegans*. As more and more microarray time course gene expression data are being generated in the such areas as cancer research (Chuang *et al.*, 2002) and neurobiology (Diaz *et al.*, 2002), we expect to see more applications of the proposed methods in identifying genes with different expression patterns over time.

Acknowledgments

This research is supported by NIH grant ES09911. We thank Dr. Stuart Kim for making the *C.elegans* data available and insightful comments.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of Royal Statistical Society* Ser B 57:289-300.

Chuang YY, Chen Y, Gadisetti VR, et al (2002): Gene Expression after Treatment with Hydrogen Peroxide, Menadione, or t-Butyl Hydroperoxide in Breast Cancer Cells. *Cancer Research*, 62:6246-6254.

Carlin PC and Louis TA (1996): *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall/CRC, New York.

De Boor C (1978): *A Practical Guide to Splines*, Springer-Verlag.

Diaz E, Ge Y, Yang YH, Loh KC, Okazaki Y, Hayashizaki Y, Serafini, TA, Speed TP, Ngai J, and Scheiffele P. (2002): Molecular analysis of gene expression in the developing pontocerebellar system. *Neuron*, 36, 417-434.

Efron B, Tibshirani R, Story JD, and Tushe V (2001): Empirical Bayes Analysis of Microarray Experiment *Journal of the American Statistical Association*, 96:1151-1160.

Guo X, Qi H, Verfaillie CM and Pan W (2003): Statistical significance analysis of longitudinal gene expression data *Bioinformatics*, 19:1628-1635.

Lönnstedt I, Speed T (2002): Replicated microarray data. *Statistica Sinica*, 12:,31-46.

Luan Y, Li H (2003): Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19:474-482.

Luan Y, Li H (2004): Model-based methods for identifying periodically regulated genes based on the time course microarray gene expression data. *Bioinformatics*, accepted.

Newton MA, Noueiry A, Sarkar D, Ahlquist P (2003): Detecting differentially gene expression with a semiparametric hierarchical mixture method. *UW-Madison Biostatistics TR 1074*.

Park T, Yi SU, Lee S, Lee SY, Yoo D, Ahn J, and Lee YS (2003): Statistical tests for identifying differentially expressed genes in time-course microarray experiments *Bioinformatics*, 19:694-703.

Ramsay JO and Silverman BW (1997): *Functional Data Analysis*, Springer-Verlag, New York.

Spellman P, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B (1998): Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273-3297.

Storey (2003): The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, in press.

Tusher VG, Tibshirani R, Chu G (2001): Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of National Academy Sciences USA*, 98: 5116-5121.

Wang J and Stuart K (2003): Global analysis of dauer gene expression in Caenorhabditis elegans. *Development* 130: 1621-1634.

Xu XL, Olson JM and Zhao LP (2002): A regression-based method to identify differentially expressed genes in microarry time course studies and its application in an inducible Huntington's disease transgenic model *Human Molecular Genetics*, 11:1977-1985.

Appendix

Parameter Estimation Using Monte Carlo EM Algorithm

We give some details of the MCEM algorithm for the IP model. Details for the JP model are similar and are omitted. The IP model includes both the hyperparameters $\theta_h = \{\pi_l, \sigma_l^2, l = 1, ...p, B, q\}$ and gene-specific parameters $\theta_g = \{\beta_{lj}, l = 1, ..., p, j = 1, ..., n\}$. Let $\tau_j = \Sigma_j^{-1}$ and denote $\theta = (\theta_h, \theta_g)$. We make the following conditional independence assumptions: $d_{lj}|e_{lj}$ are independent for all $l = 1, ..., p$, $Y_{jik}|(d_{lj}, l = 1, ..., p, \Sigma_j)$ are independent for all replications $k = 1, ..., K_i$ and two conditions $i = 1, 2$, and the n genes are independent. Under these assumptions, we have

$$\begin{aligned} f(e_j) &= \prod_{l=1}^{p} (\pi_l)^{e_{lj}} (1 - \pi_l)^{1-e_{lj}}, \\ f(d_j|e_j) &= \prod_{l=1}^{p} \{\phi(d_{lj} - 0; \sigma_l^2)\}^{e_{lj}} \{\delta(d_{lj})\}^{1-e_{lj}}, \\ f(\tau_j) &= \frac{|\tau_j|^{(q-T-1)/2} \exp\{-tr(\tau_j B^{-1})/2\}}{2^{Tq/2} \pi^{T(T-1)/4} |B|^{q/2} \prod_{i=0}^{T+1} \Gamma\{1/2(q-i)\}}, \end{aligned}$$

$$\begin{aligned} & f(Y_{jik}, i = 1, 2, k = 1, ..., K_i|\tau_j, \mu_{ji}, i = 1, 2) \\ = \; & f(Y_{j1k}, k = 1, ..., K_1|\tau_j, \mu_{j1}) f(Y_{j2k}, k = 1, ..., K_2|\tau_j, \mu_{j2}) \\ = \; & \prod_{k=1}^{K_1} \Phi(Y_{j1k} - \mu_{j1}, {\tau_j}^{-1}) \prod_{k=1}^{K_2} \Phi(Y_{j2k} - \mu_{j2}, {\tau_j}^{-1}), \end{aligned}$$

where $\phi(\centerdot - \mu, \sigma^2)$ and $\Phi(\centerdot - \underline{\mu}, \Sigma)$ are normal and multivariate normal density functions. Let $Y_j = \{Y_{jik}, i = 1, 2, k = 1, ..., K_i, \tau_j, e_j, d_j\}$ denote the complete data for gene j and let $Y = \{Y_1, \cdots, Y_n\}$ be the complete data for all the n genes. Then the complete data likelihood function for the jth gene is

$$L_j = \{\prod_{l=1}^{p} f(e_j)\}\{\prod_{l=1}^{p} f(d_j|e_j)\} f(Y_{j1k}, k = 1, ..., K_1|\tau_j, \mu_{j1}) f(Y_{j2k}, k = 1, ..., K_2|\tau_j, \mu_{j2}),$$

and the overall complete data likelihood function is $L = \prod_{j=1}^{n} L_j$. Let $l_j = \log L_j$ and $l = \sum_{j=1}^{n} l_j$ be the corresponding log likelihood functions.

It is easy to see that the EM equations for updating the parameters in θ are given as the following,

$$\pi_l^{(new)} = \frac{\sum_{j=1}^n E(e_{lj}|Y)}{n},$$

$$\sigma_l^{2(new)} = \frac{\sum_{j=1}^n E(e_{lj}d_{lj}^2|Y)}{\sum_{j=1}^n E(e_{lj}|Y)},$$

$$\beta_l^{(new)} = \frac{1}{K_1+K_2}\left[B\{E(\tau_j|Y)\}B^{'}\right]^{-1}\left[B\{E(\tau_j|Y)\}(\sum_{i=1}^{2}\sum_{k=1}^{K_i} Y_{jik}) - K_2 B\{E(\tau_j B'(ed)_j|Y)\}\right],$$

$$q^{(new)} = argmax\left[\frac{q-T-1}{2}\sum_{j=1}^{n} E(\log\tau_j|Y) - \frac{Tqn}{2}(1+\log 2) - n\sum_{i=1}^{T}\log\Gamma\{\frac{1}{2}(q+1-i)\}\right],$$

$$H^{(new)} = \frac{1}{q^{(new)}n}\sum_{j=1}^{n} E(\tau_j|Y),$$

where $B = (B_1, ..., B_p)'$ and $B_l = \{B_l(1), ..., B_l(T)\}'$ is the matrix and vector of curve-modelling basis functions evaluated at the observed data points, and $(ed)_j = (e_{1j}d_{1j}, ..., e_{pj}d_{pj})$. For updating parameters H and q, we fist update q by one-dimensional minimization, then update H based on the new q.

In order to update these parameters in the M-step, we need the following expectations, $E(e_j|Y)$, $E(e_j d_j^2|Y)$, $E(\tau_j|Y)$ and $E(\tau_j X'(ed)_j|Y)$ in the E-step. However, direct calculations of these expectations are not feasible due high-dimensional integrations. Here we propose to approximate these expectations by using the Gibbs sampling, which involves sampling conditional distribution sequentially (see Carlin and Louis (1996) for more information on using Gibbs sampling for empirical Bayes analysis). In particular, for the nth Gibbs step, we perform the following three sequential sampling steps,

1. For each gene j, draw $\tau_j^{(n)}$ from the condition distribution of $\tau|\{Y_j, e_j^{(n-1)}, d_j^{(n-1)}\}$, which can be shown to follow a Wishart distribution, $W(H_j^{(n)}, q_j^{(n)})$ with

$$\begin{aligned} H_j^{(n)} &= H^{-1} + \sum_{i=1}^{2}\sum_{k=1}^{K_i}(Y_{jik}-\mu_{ji})(Y_{jik}-\mu_{ji})', \\ q_j^{(n)} &= q + K_1 + K_2, \end{aligned}$$

where $\mu_{j1} = B^{'}\beta_j^{old}$, $\mu_{j2} = B^{'}(\beta_j^{old}+de_j^{(n-1)})$ and $de_j^{(n-1)} = (d_{1j}^{(n-1)}e_{1j}^{(n-1)}, ..., d_{pj}^{(n-1)}e_{pj}^{(n-1)})'$.

2. For each gene j, draw $e_l^{(n)}$, from conditional distribution of $e_l|\{Y_j, \tau_j^{(n)}, d_j^{(n-1)}\}$, which is a discrete distribution with 2^p possible outcomes.

3. For each gene j, draw $d_j^{(n)}$ from the conditional distribution of $d_j|\{Y_j, \tau_j^{(n)}, e_j^{(n)}\}$, which is a multivariate normal distribution $MVN(\nu_j^{(n)}, \Sigma_j^{(n)})$ with parameters

$$\begin{aligned} \nu_j^{(n)} &= \Sigma_j^{(n)} Q_j (\sum_{k=1}^{K_2} Y_{j2k} - K_2 B' \beta_j^{old}), \\ \Sigma_j^{(n)} &= (A_j + K_2 Q_j \tau_j^{old} Q_j')^{-1}, \end{aligned} \tag{8}$$

where $A_j = diag(\frac{e_{1j}^{(n)}}{\sigma^2{}_1^{old}}, ..., \frac{e_{pj}^{(n)}}{\sigma^2{}_p^{old}})$, $Q_j = (Q_{1j}, ..., Q_{Tj})$, and $Q_{tj} = \{B_1(t)e_{pj}^{(n)}, ..., B_p(t)e_{pj}^{(n)}\}'$.

Repeat the Gibbs-sampling N times, we can approximate the conditional expectation of the function $g(e_j, d_j, \tau_j)$ by

$$E(g(e_j, d_j, \tau_j)|Y_j) \approx \frac{1}{N} \sum_{n=1}^{N} g(e_j^{(n)}, d_j^{(n)}, \tau_j^{(n)}).$$

Calculation of the posterior probabilities and likelihood function

After obtaining the parameter estimation, for each gene we can calculate the posterior probability of being TDE gene, P_j by the Gibbs samples at the last EM step. For the model with a constant error variance σ_e^2, we can calculate this probability analytically as

$$\begin{aligned} Pr(e_j|Y_j) &\propto Pr(e_j, Y_j) \\ &= \int Pr_j(e_j, d_j, Y_j) \mathrm{d}d_j \\ &= C_j(e_j, l = 1, .., p) \int \Phi(d_j - \nu_j, \Sigma_j) \mathrm{d}d_j \\ &\propto \prod_{l=1}^{p} \left[(\hat{\pi}_l)^{e_{lj}} (1 - \hat{\pi}_l)^{1-e_{lj}} \{ \frac{1}{\sqrt{2\pi \hat{\sigma_l^2}}} \}^{e_{lj}} \right] \exp(\frac{1}{2} \nu_j^{*'} \Sigma_j^{*-1} \nu_j^*) (2\pi)^{p*/2} (|\Sigma_j^*|), \end{aligned}$$

where ν_j^* and Σ_j^* are the sub-vector and sub-matrix of ν_j and Σ_j which corresponds to non-zero element of e_j and ν_j and Σ_j are given in equation (8) evaluated at the converged parameter values. Finally,

$$Pr_j(e_{1j} = 0, ..., e_{pj} = 0|Y) \;=\; \frac{C_j(0, ..., 0)}{\sum_{e_{1j}, ..., e_{pj}} C_j(e_j, l = 1, .., p)}.$$

In addition, we can also obtain the posterior mean of the random effect d_{lj} for each gene based on the Gibbs samples obtained at the last EM algorithm.

The observed data likelihood can be computed as:

$$\begin{aligned} f(Y_{jik}, i = 1, 2, k = 1, ..., K_i, j = 1, ..., n) &= \prod_{j=1}^{n} \sum_{e_{1j},...,e_{pj}} f(Y_{jik}, e_{lj}, , i = 1, 2, k = 1, ..., K_i) \\ &= \prod_{j=1}^{n} \sum_{e_{1j},...,e_{pj}} C_j(e_j, l = 1, .., p), \end{aligned}$$

where $C_j(e_j, l = 1, .., p)$ are obtained when calculating posterior probability above. The likelihood value can then be applied in the model selection step using AIC or BIC.

Table 1: Array lay-out for studies comparing two different experimental conditions, where Y_{jikt} is the log of the expression level for the jth gene at time t in the kth replication under condition i.

condition i	————————			***********************		
replication k	——	....	——	——	...	——
time point t	$t_1, \cdots, t_T$	...	$t_1, \cdots, t_T$	$t_1, \cdots, t_T$	...	$t_1, \cdots, t_T$
array	Array ikt					
gene j	gene expression level Y_{jikt}					

Table 2: Identification results for 97 true TDE genes using for a FDR of 0.01 for $K = 2$ and 10 replications. Ebayes: proposed empirical Bayes methods; Anova: two-way ANOVA. The entry is the number of falsely identified genes/number of TDE genes identified.

		Error variance (σ_e^2)				
K	Method	0.01	0.05	0.1	0.5	1.0
2	EBayes	0/86	0/67	0/40	0/4	0/2
	Anova	0/76	0/33	0/15	0/0	0/0
10	Ebayes	0/93	0/87	0/81	0/49	0/33
	Anova	0/91	0/80	0/73	0/33	0/12

Table 3: Number of TDE *C. elegans* genes identified by the Ebayes methods and the Anova method by Wang and Kim (2003).

	# of genes identified			
FDR	EBayes	Anova(Kim)	Overlap	Overlap(top)
0.01	1011 (918)	1681	850 (824)	636 (608)
0.05	1049 (957)	1949	948 (904)	669 (640)

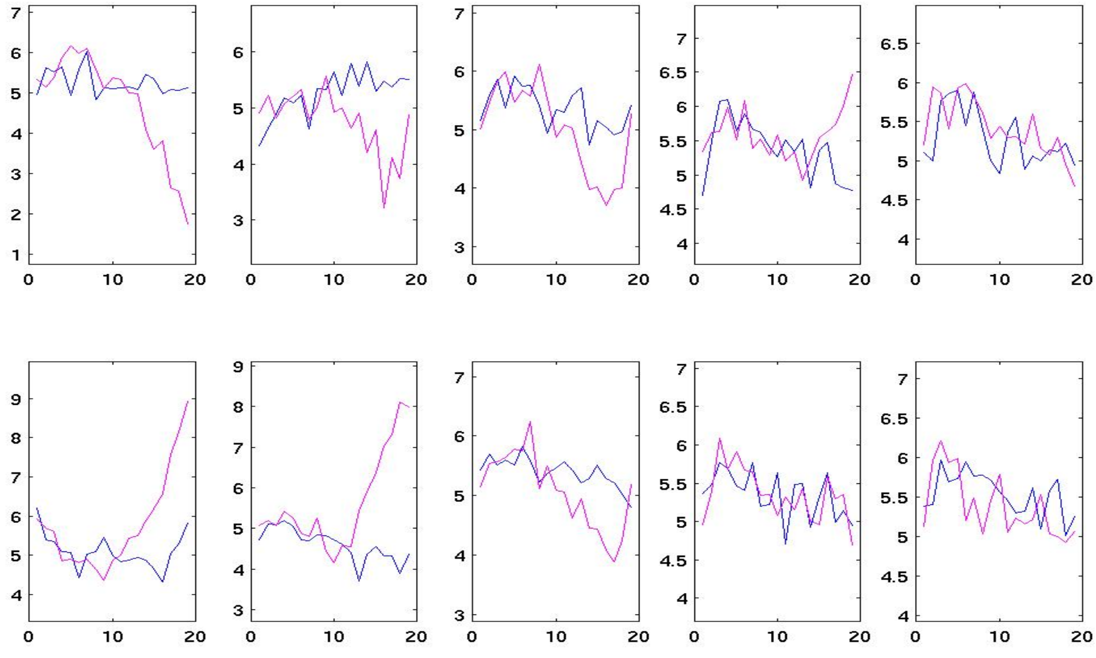

Figure 1: Examples of simulated gene expression profiles

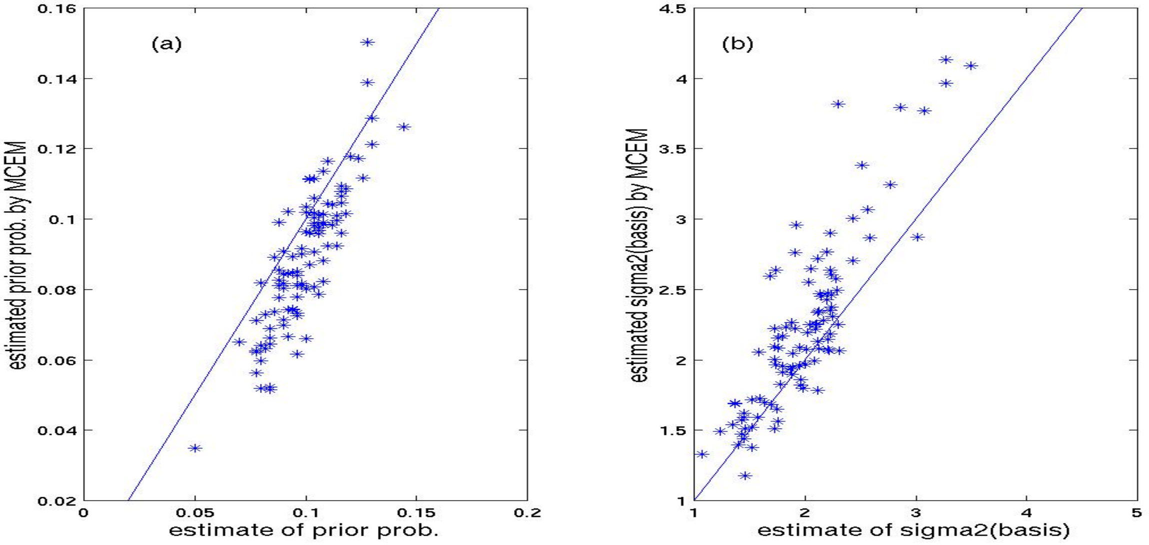

Figure 2: Simulation results based on 50 replications for Model 1. (a) Estimated π_5 and π_6 by the EM algorithm versus the estimates based on simulated values of e_{5j} and e_{6j}. (b) Estimated σ_5^2 and σ_6^2 by the EM algorithm versus the estimates based on simulated values of d_{5j} and d_{6j}.

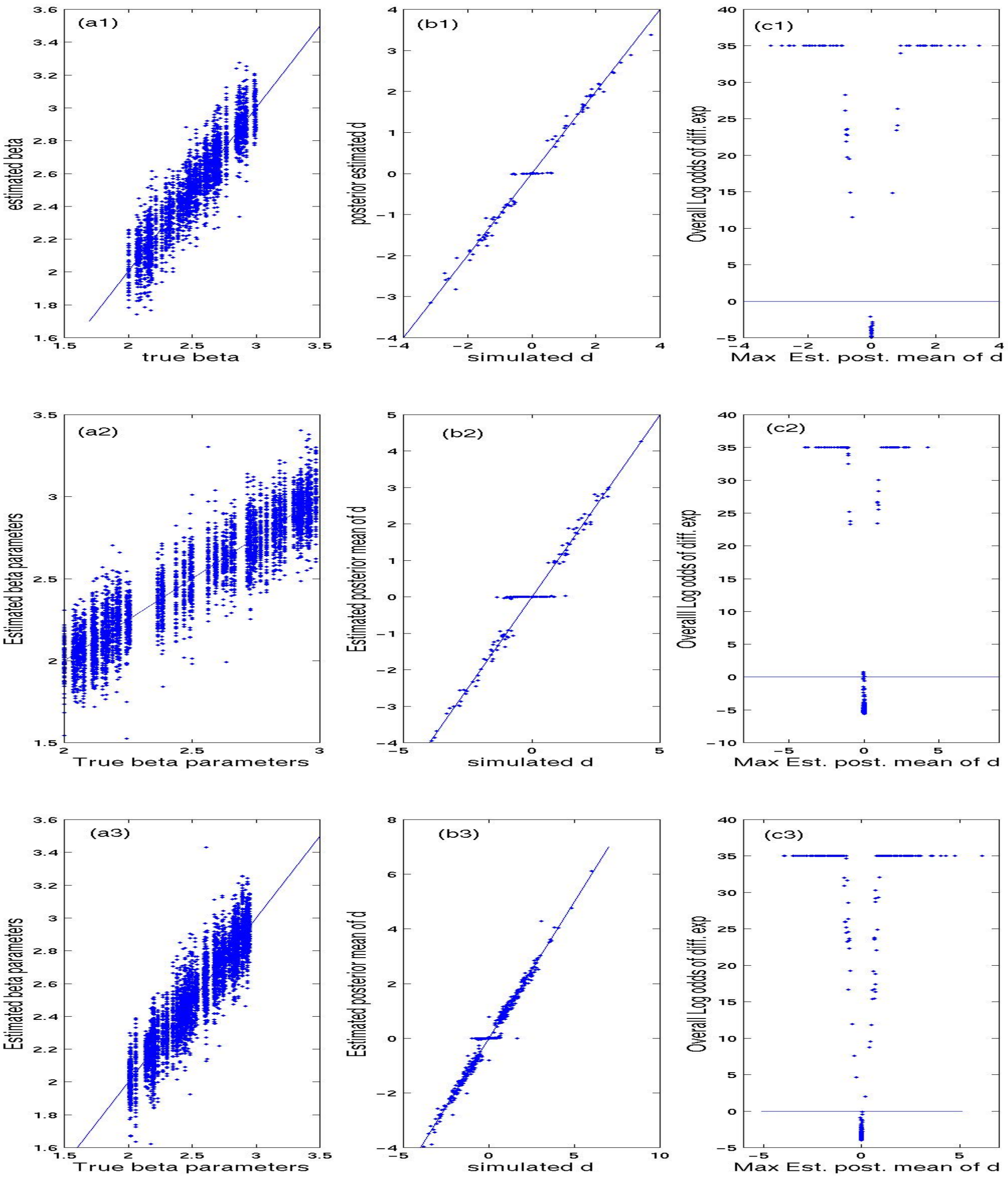

Figure 3: Demonstration of results for one simulated data set for Model 1 (top panel), Model 2 (middel panel) and Model 3 (bottom panel). (a1), (a2), (a3): Estimated gene specific β_{lj} versus the true values; (b1), (b2), (b3): Posterior mean of d_j versus the simulated values; (c1), (c2), (c3): log posterior odds versus the maximum of the posterior means of d_{lj} with a ceiling of log odds ratio of 35 applied.

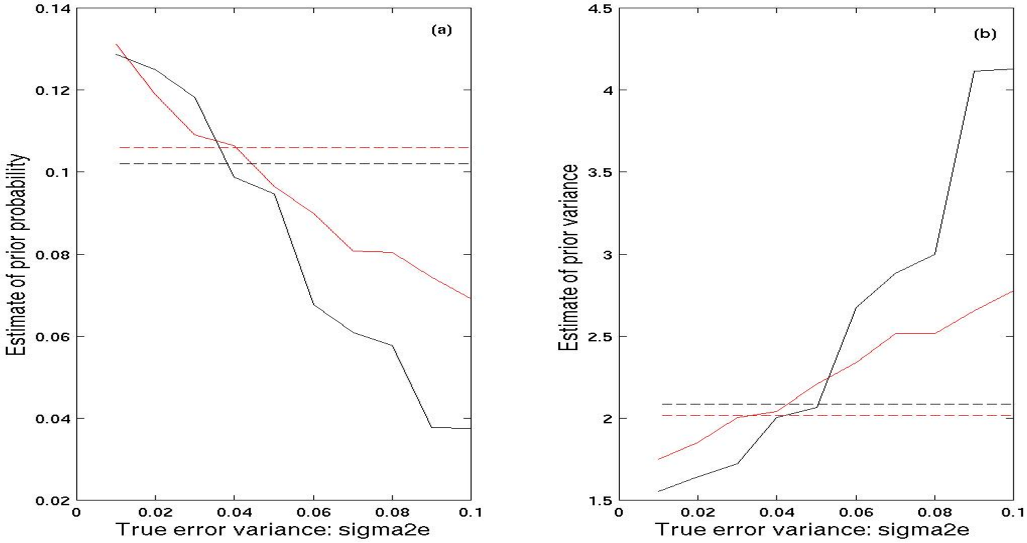

Figure 4: EM estimates of π_5 and π_6 (plot (a)) and σ_5^2 and σ_6^2 (plot (b)) for increasing error variance σ_e^2 (x-axis). Solid lines represent the estimated parameters by the MCEM algorithm; dotted horizontal lines represented estimates based on simulated values.

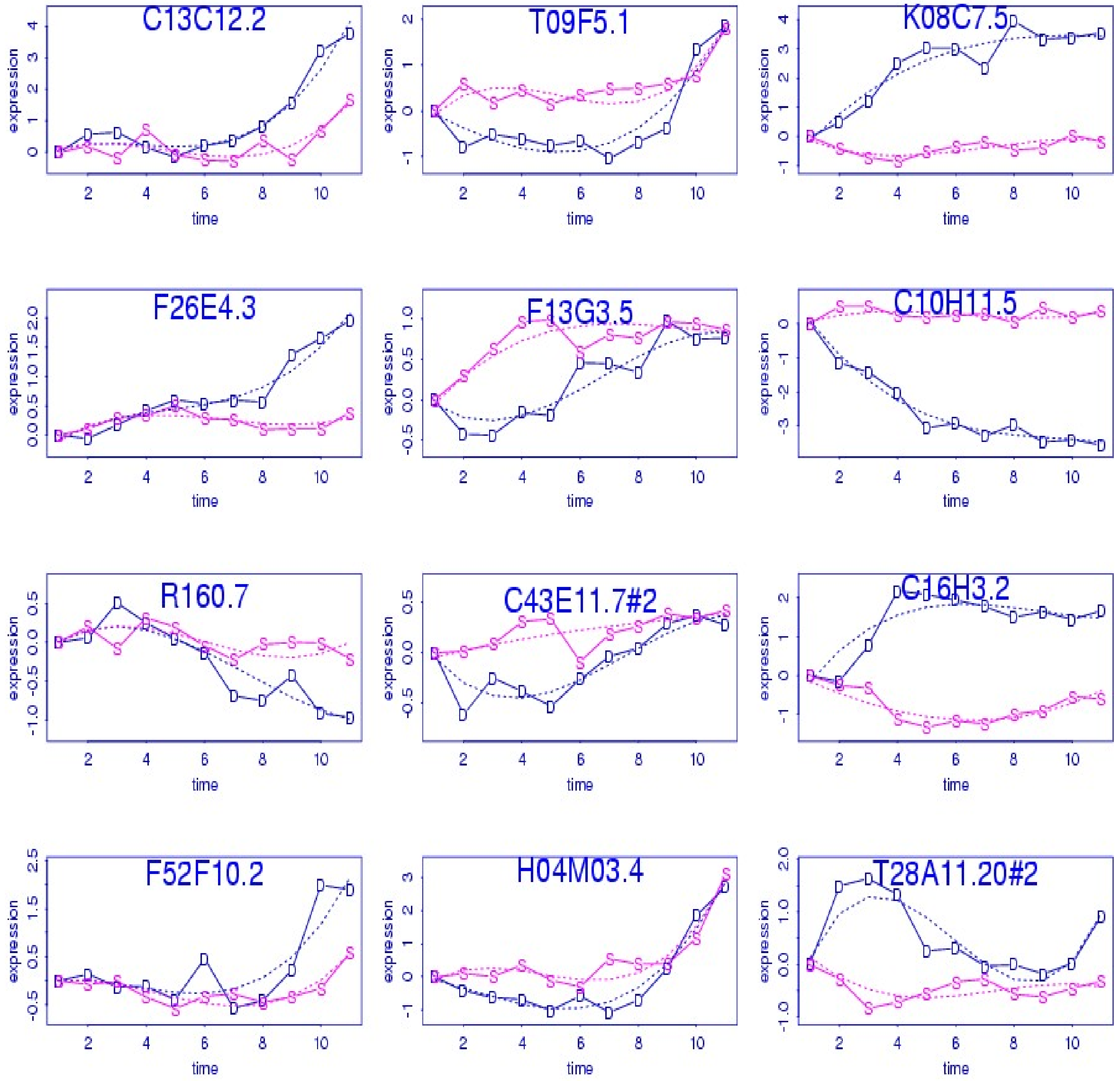

Figure 5: Observed (solid) and fitted (dashed) expression profile of the selected genes identified to be differentially expressed at late time (left column), at early time (middle column) and over all time points (right column). The KimLab IDs for these genes are: 1st column, *C13C12.2, F26E4.3, R160.7, F52F10.2*; 2nd column, *T09F5.1, F13G3.5, C43E11.7*#2, *H04M03.4*; 3rd column: *K08C7.5, C10H11.5, C16H3.2, T28A11.20*#2.

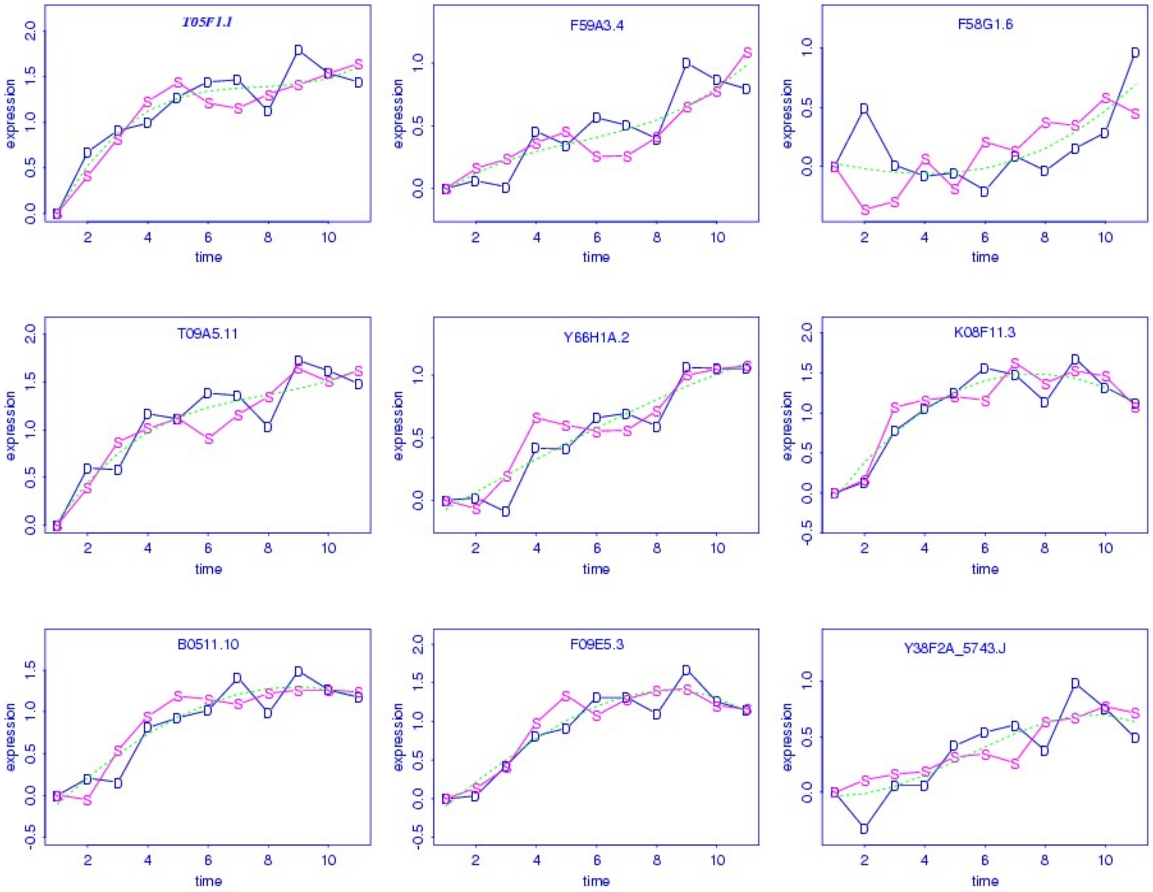

Figure 6: Observed (solid) and fitted (dashed) expression profile of nine genes with the smallest posterior probability of being TDE. The KimLab IDs for these genes are: Top three: *"F22F7.2", "T05F1.1", "F59A3.4"*, middle three: *"F58G1.6", "T09A5.11", "Y66H1A.2"*, bottom three: *"K08F11.3", "B0511.10", "F09E5.3"*

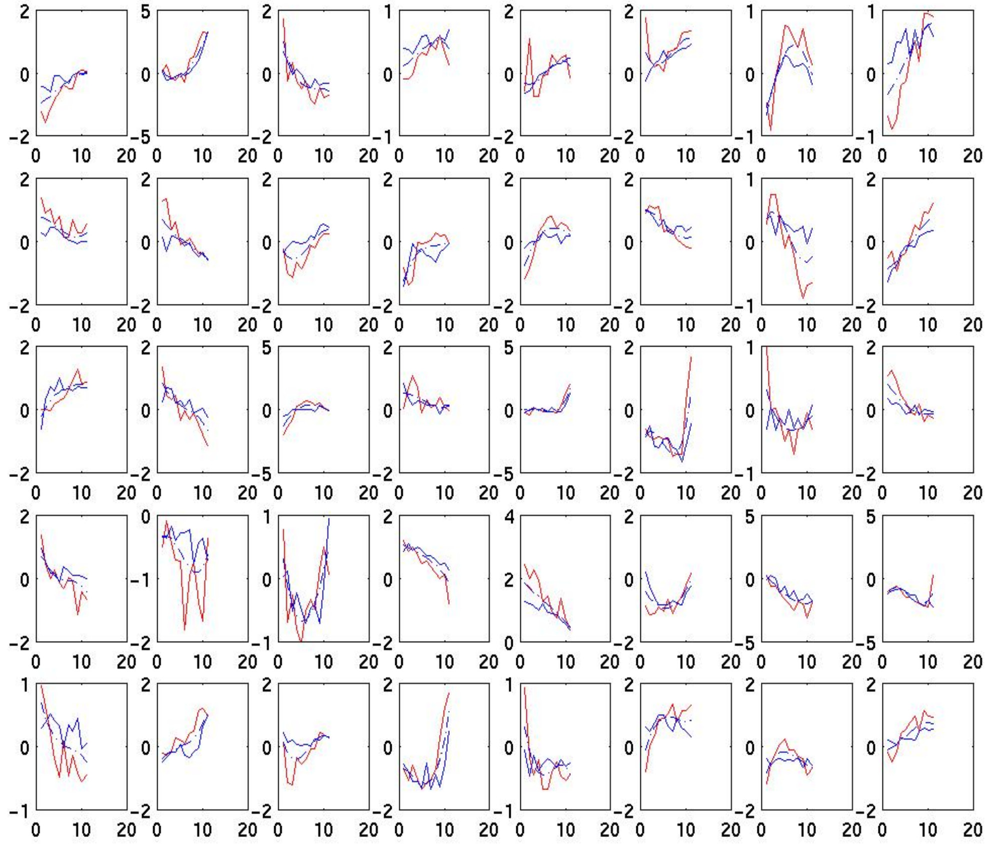

Figure 7: Average gene expression profiles and fitted smooth curves for forty randomly selected genes that were selected by ANOVA but not by our proposed empirical Bayes methods.