Stat Biosci (2010) 2: 95–119
DOI 10.1007/s12561-010-9024-z

Identifying Differentially Expressed Genes for Time-course Microarray Data through Functional Data Analysis

Kun Chen · Jane-Ling Wang

Received: 19 October 2009 / Accepted: 15 October 2010 / Published online: 3 November 2010

Abstract Identification of differentially expressed (DE) genes across two conditions is a common task with microarray. Most existing approaches accomplish this goal by examining each gene separately based on a model and then control the false discovery rate over all genes. We took a different approach that employs a uniform platform to simultaneously depict the dynamics of the gene trajectories for all genes and select differently expressed genes. A new Functional Principal Component (FPC) approach is developed for time-course microarray data to borrow strength across genes. The approach is flexible as the temporal trajectory of the gene expressions is modeled nonparametrically through a set of orthogonal basis functions, and often fewer basis functions are needed to capture the shape of the gene expression trajectory than existing nonparametric methods. These basis functions are estimated from the data reflecting major modes of variation in the data. The correlation structure of the gene expressions over time is also incorporated without any parametric assumptions and estimated from all genes such that the information across other genes can be shared to infer one individual gene. Estimation of the parameters is carried out by an efficient hybrid EM algorithm. The performance of the proposed method across different scenarios was compared favorably in simulation to two-way mixed-effects ANOVA and the EDGE method using B-spline basis function. Application to the real data on *C. elegans* developmental stages also suggested that FPC analysis combined with hybrid EM algorithm provides a computationally fast and efficient method for identifying DE genes based on time-course microarray data.

Research of Jane-Ling Wang is supported by NSF Grant DMS09-06813.

K. Chen (✉)
Dana-Farber Cancer Institute, Boston, USA
e-mail: kun_chen@dfci.harvard.edu

K. Chen · J.-L. Wang
University of California, Davis, USA

Keywords Time-course gene expression · Differentially expressed genes · Functional data analysis · Functional principal component · Hybrid EM · False discovery rate

1 Introduction

One of the fundamental problems in biological studies is to identify genes that are responsive to environmental changes, or responsible for phenotypic differences. This can be carried out by examining gene expression levels of samples in groups exposed to different environments using microarray technology. Time-course experiment design allows the monitor of dynamic process of gene expression. Transient change or random fluctuation can be discerned if the gene expressions are observed frequently over time. Besides, certain biological problems, such as identification of genes active at different development stages, require time-course data. The advantages of the time-course design motivate the wide use of time-course microarray data [8, 17, 22].

A key feature of time-course data is that the repeated measurements of one gene at different time points are likely to be correlated since they are taken from the same subject. Furthermore, the cellular mRNA concentration can serve as a feedback regulatory mechanism, thus influencing future gene expression. This correlation can be called "within-gene correlation." Various statistical methodologies, such as ANOVA test [5], moderated t-test [21], empirical Bayes method [6] and mixture modeling [13] etc., have been developed for the cross-sectional design based on independent samples. These methods are not applicable to time-course microarray study due to the aforementioned "within-gene correlation" present in the longitudinal samples, and also because the time trend of the gene expression profile is not considered in these models.

Several approaches have been proposed to identify differentially expressed genes for time-course microarray data. In general, the goal is accomplished by testing for each gene the null hypothesis of equivalent expression against the alternative hypothesis of differential expression, based on a model depicting the observed time-course data. The multiple test correction method, such as Bonferroni or false discovery rate (FDR), is usually followed to adjust the p-value. A commonly used model, two-way ANOVA model, is employed by treating different time points as multiple levels of time factor [14]. There are two limitations of adapting the ANOVA approach to longitudinal data. First, the timing information of when the measurements are taken is not utilized. Second, it does not take into account the within-gene correlation.

The first problem can be solved by delineating the temporal expression profile as a function of time. The temporal expression profile is modeled in [23] as a function of lower-order polynomials in time. Rather than the lower-order polynomials, some use a more flexible representation of time-course gene expression profiles by B-splines [3, 7, 19] or functional principal components (FPCs) [12], techniques often used in the statistical literature of functional data analysis (FDA) [15, 16, 24]. Compared to the ANOVA model, use of the basis functions reduces the number of parameters needed for describing an expression profile, thus saving the degree of freedom and increase the efficiency. The other benefit to use basis representations, besides model efficiency, is that this method can be applied to missing data without imputation.

The second issue can be resolved practically by incorporating random effects explaining the temporal correlation like a two-way mixed-effects ANOVA model [22]. Analogously, the random effects can also be added into the basis function-based model. The model with a random intercept in [19] assumes the expression levels sampled at every pair of time points have the same correlation no matter how far or close these time points are. A more flexible correlation structure is allowed by modeling individual profiles as B-spline functions [7] or FPCs [12] with random coefficients. Some other existing methods model the covariance structure of the expression process without utilizing the time trend. For example, the temporal profiles are treated as multivariate normal vector in [20] with a hierarchical distribution imposed on the covariance matrix of the multivariate normal vector. A Hidden Markov model [25] is proposed by considering the observed profile as being influenced by an underlying Markov process, where the within-gene correlation is implied in the presence of first-order dependence structure of the underlying process. B-spline and FPC models require least assumptions on the covariance structure. However, for B-spline model, many basis functions might be required to capture the variation of the temporal expression, and the selection of proper knots number and their locations is critical, which increases the computational burden.

In these methods, the test statistics are constructed either for each gene separately [2, 12, 14, 19, 23], or part of the statistics involves information from all genes, for example, shrinkage variance estimates [7, 20] and maximum a posteriori (MAP) estimates [25], which are in the same spirit as moderated t-tests [21] and empirical Bayes methods [6] for independent microarray samples. In the former case, a separate model is fitted for each gene and often a large number of parameters are involved for a relatively modest sample size, leaving few degrees of freedom for the inference procedure. Also, gene-specific variance estimates are not precise due to small number of replications in usual microarray studies. For instance, in [12], FPC model is fitted for one gene at a time to estimate the gene-specific group means and the covariance structure nonparametrically, and to construct the statistics subsequently. The small sample size in both groups, 6 cases and 12 controls in [12] example, may result in an overfitting of the model. The test statistics with shrinkage alleviate these problems to some extent.

In this work, we propose a unified approach to model all gene profiles using the techniques in functional principal component analysis (FPCA), which allows flexible representation of the gene expression trajectories and the temporal covariance structure. In our model, each gene can be considered as a random realization from the gene pool, in contrast to [12, 19], where each gene is considered as an isolated case. Thus, fewer parameters per gene are needed and the information across genes can be borrowed to enhance the variance estimation. The proposed model reflects the commonality of all gene expressions, yet allows for subject-specific variation of individual genes. We aimed to measure the magnitude of differential expression based on the gene specific variation. In [19] and [12], bootstrap method is used to obtain the p-value for each gene. This is a computationally expensive strategy, especially in [12], as each bootstrap sample requires a repeat of the FPCA procedure. An appealing approach is proposed in [7], where the posterior probability of differential expression for each individual gene is estimated and then used to select significant genes. We adopt such an approach to impose a mixture distribution on the gene specific variation and using an indicator to reflect whether the gene is differentially expressed.

The proposed method is compared with existing methods in the literature via both simulated and real data. The real data set is taken from the aforementioned study [22], where 2430 genes are candidates, showing significant change of expression during the dauer exit compared to a reference group. Because some of the genes may change their expression solely by food induction, a known mechanism causing dauer exit, in order to identify the true genes responsive to dauer exit, we have to compare the gene expression profiles during this course to those during the reintroduction of food for starved worms in L1 stage, a stage before entering the dauer stage. In the following, we will use the terms dauer and L1 to denote the corresponding two groups.

This paper is organized as follows: in Sect. 2 we describe the development of FPCA model for detection of differentially expressed genes, in Sect. 3 we apply the proposed approach to the gene expression data collected during the dauer exit process of *C. elegans* [22], in Sect. 4 we demonstrate the performance of the proposed approach by simulation studies, and in Sect. 5 we discuss possible extensions of the methodology.

2 Model Development

We first elaborate on FPCA approach for time-course microarray data. Our method consists of two steps. In the first step, the individual expression trajectory over time is reduced to a finite number of FPC scores, which are then used in the second step to construct the criterion to identify differentially expressed genes. This approach accommodates replications through a multi-level mixed-effects model and is broadly applicable to detect differentially expressed genes.

2.1 FPCA Model for Grouped Microarray Data

We adopt the point of view in functional data analysis to view time-course expression levels of each gene as one realization of a random curve out of the gene pool, consisting of genes in the whole genome. The true expression profile of each gene is determined at the level of gene, but can be regarded as a random outcome at the level of the gene population. More specifically, a single expression profile $X(t)$, assumed to be a smooth function in the time interval $[a, b]$, can be modeled as the sum of the population mean expression profile $\mu(t)$ and a random deviation from $\mu(t)$. The random deviation is further expanded through a set of orthogonal basis functions $\{\phi_l(t)\}$ yielding

$$X(t) = \mu(t) + \sum_{l=1}^{\infty} b_l \phi_l(t), \quad a \le t \le b. \tag{1}$$

The randomness of the process $X(t)$ is now represented by the sequence of random variables b_l. In FPCA model, the basis functions $\phi_l(t)$ are chosen to be eigenfunctions, satisfying

$$\int_a^b G_X(s,t)\phi_l(s)\,ds = \lambda_l \phi_l(t).$$

Here $G_X(s,t)$ is the covariance function of $X(t)$, defined by

$$G_X(s,t) = \mathrm{Cov}\big(X(s), X(t)\big) = E\big(X(s) - \mu(s)\big)\big(X(t) - \mu(t)\big).$$

Such a decomposition in (1) is known as the Karhunen–Loève expansion [1] for functional data.

The eigenfunctions, $\phi_l(t)$, analogous to the eigenvector in multivariate analysis, maximize the variability of $\int_a^b \phi_l(t)X(t)\,dt$ subject to $\int_a^b \phi_l^2(t)\,dt = 1$. Besides this interpretation, the eigenfunctions reflect the direction of major shape deviation from the mean function. The random coefficient, b_l, associated with the corresponding eigenfunction $\phi_l(t)$, explains how much a gene deviates from the mean function in the direction of that eigenfunction.

The aforementioned principal component approach is suitable for situations when all the gene expression profiles have the same mean structure. When expression profiles are sampled from different groups, such as a control and a treatment group, we need to make some adjustment to allow for different mean structures for different groups before we pool all the gene expression profiles. To see this, consider two populations with respective mean functions, $\mu_0(t)$ and $\mu_1(t)$, covariance functions, G_0 and G_1, and mixed together with proportion p and $1-p$. The covariance function of the mixed (pooled) population becomes

$$G_X(s,t) = pG_0(s,t) + (1-p)G_1(s,t) + p(1-p)\big(\mu_0(s) - \mu_1(s)\big)\big(\mu_0(t) - \mu_1(t)\big),$$

which not only contains the mixture of two covariance functions but also a third term attributed to the difference between the two group means. To avoid the distortion of covariance function due to different population means, we first subtract the corresponding group mean function from each gene profile before we perform further analysis. This operation can be performed by estimating the mean expression profiles separately through the nonparametric approach in [24], among others. A nice consequence is that the covariance G_X of the mixed population becomes

$$G_X(s,t) = pG_0(s,t) + (1-p)G_1(s,t), \tag{2}$$

which collapses to $G_X = G_0 = G_1$ when $G_0 = G_1$. In more general cases, the covariance structure from different groups, G_0 and G_1, may not be the same. They can be estimated separately through nonparametric approach for each group and plugged in back to (2) to get the pooled estimate for G_X.

In real experiments, the gene expression profiles are typically measured on different experimental subjects (or replicates). We denote the expression profile curve of gene i, $i = 1, \dots, n$, from experimental subject (or replicate) $j = 1, \dots, J$, by $X_{ij}(t)$. These J subjects can be from several groups, but for simplicity of presentation we assume that there are only two groups, a control and a treatment group, and note that our approach can be extended to multiple groups as well. We use the notation z_j to indicate the group membership of the jth experimental subject,

$$z_j = \begin{cases} 0, & \text{if the } j\text{th subject is from a control group,} \\ 1, & \text{if the } j\text{th subject is from a treatment group.} \end{cases}$$

Let J_0 and J_1 be the number of subjects in a control and a treatment group, respectively. For the *C. elegans* data in [22], $J_0 = J_1 = 4$ for both groups. The individual gene expression profile can thus be represented as:

$$X_{ij}(t) = \mu_{z_j}(t) + \sum_{l=1}^{\infty} b_{ijl}\phi_l(t), \tag{3}$$

where $\mu_{z_j}(t)$ is the population mean profile for a control group if $z_j = 0$ or a treatment group if $z_j = 1$, and b_{ijl} are the principal scores which are uncorrelated random variables representing the between-subject variations. Typically, only a few scores suffice to summarize the information in the gene trajectories and we will modeled these scores in the next subsection.

2.2 Decomposition of the Random Coefficients

Since the purpose of the study is to identify the differentially expressed genes and differences between individual trajectory are carried in the principal component scores, we set up another model for b_{ijl}. The random coefficients b_{ijl} in (3) can be further decomposed into a gene effect u_{il}, accommodating gene-specific shape, an effect w_{il} that particularly accounts for the effect of differential expression, a replicate effect v_{jl} for the whole array, and the gene-specific replicate effect e_{ijl}. Thus our final model of trajectory function $X_{ij}(t)$ becomes

$$X_{ij}(t) = \mu_{z_j}(t) + \sum_{l=1}^{\infty}(u_{il} + z_j w_{il} + v_{jl} + e_{ijl})\phi_l(t). \tag{4}$$

In model (4), z_j indicates the group of jth experimental subject and $\mu_{z_j}(t)$ the corresponding group mean function as described previously. The random effects above all have zero means with variance $\mathrm{var}(u_{il}) = \sigma^2_{u,l}$, $\mathrm{var}(w_{il}) = d_i\sigma^2_{w,l}$, $d_i \sim \text{Bernoulli}(\pi)$, $\mathrm{var}(v_{jl}) = \sigma^2_{v,l}$ and $\mathrm{var}(e_{ijl}) = \sigma^2_{e,l}$, $l = 1, 2, \ldots$. The random term v_{jl} can be dropped if the array replicate effect is very weak and not needed, as in the *C. elegans* data in Sect. 3. When $d_i = 0$, random variable w_{il} has a degenerate distribution such that the ith gene will not have differential effect. All the random variables considered in the model are independent across the subscripts.

Although the number of basis in (4) is infinite in theory, in reality only a few bases are needed. Standard protocol is to choose the first L eigenfunctions that explain a sizable fraction of total variation in the data. This selection criterion based on the fraction of total variation explained (FVE) is rather subjective and may be assisted by a scree plot. Alternatively, AIC and BIC criteria are often used to select L. A good strategy is to assess the results from AIC, BIC and FVE, and then subjectively decide the number of basis functions. There is a risk in using a small number of basis functions to model potentially highly variable expression as some highly variable DE genes may be left out. One remedy is to increase the number of bases until the size of the identified set does not seem to change substantially. The approach based on the FPCs has the advantage that it lists all components of variations in descending order along with the corresponding eigenfunctions, so the final decision could be data-adaptive.

2.3 Model for the Observed Time-course Data

Equation (4) describes the true gene expression profiles. However, in reality, the true gene expressions may not be observed directly or completely, instead they may be observed at finite time points and further disrupted by the noisy signal ϵ with mean 0 and variance σ_ϵ^2. Denote by y_{ijk} the observed mRNA expression of the ith gene from the jth experimental subject (or replicates) at the kth time point, t_{ijk}; the model for the observation can be written as

$$y_{ijk} = X_{ij}(t_{ijk}) + \epsilon_{ijk}, \quad \text{for all } i = 1, \dots, n,\ j = 1, \dots, J,\ k = 1, \dots K_j, \tag{5}$$

where n is the number of genes in the microarray study, J is the number of experimental subjects or replicates, and K_j is the number of sampling time points and may be the same for n genes in the jth subject.

The parameters to be estimated in the proposed model (see (4) and (5)) include population mean functions of control and treatment groups, covariance matrix, variance components for the random terms ($\sigma_{u,l}^2$, $\sigma_{w,l}^2$, $\sigma_{v,l}^2$, $\sigma_{e,l}^2$, $l = 1, \dots, L$), proportion of differentially expressed genes (π) and an error term (σ_ϵ^2). The population mean functions and the covariance matrix are estimated nonparametrically via smoothing methods, so only the remaining parameters are being estimated and there are $4L + 2$ of them, where L is the number of bases selected. In our example of the nematode data, $L = 2$, so a total of 10 parameters are used for all 2430 genes.

Details for the estimation of the population mean function $\mu(t)$, the eigenfunctions $\phi_l(t)$ and eigenvalues λ_l of $G_X(s,t)$, and the FPC scores b_{ijl} can be found in Appendix A.1 and [24]. To obtain the estimates for variance components $\sigma_{u,l}^2$, $\sigma_{w,l}^2$, $\sigma_{v,l}^2$ and $\sigma_{e,l}^2$, a hybrid EM algorithm was developed, which uses Least Squares Method to efficiently get the estimates $\hat{\sigma}_{u,l}^2$, $\widehat{\pi\sigma}_{w,l}^2$, $\hat{\sigma}_{v,l}^2$ and $\hat{\sigma}_{e,l}^2$ and EM algorithm to disentangle the π and $\sigma_{w,l}^2$. We impose the normal distribution on FPC scores for EM to apply, as the normal distribution fits the scores with satisfactory result, observed from the quantile plot of FPC scores versus the theoretical normal quantiles (not shown here). Details of the algorithm can be seen in Appendices A.2 and A.3.

2.4 The Connection Between FPCA Model and Two-way Mixed ANOVA Model

The two-way mixed ANOVA model for the ith gene at time t_k in jth replication can be written as

$$y_{ijk} = \mu_i + \alpha_{iz_j} + \beta_{ik} + \gamma_{iz_jk} + r_{ij} + e_{ijk}, \tag{6}$$

with the constraints $\alpha_{i0} = 0$, $\beta_{i1} = 0$ and $\gamma_{i0k} = \gamma_{i11} = 0$ for all i, k. Here μ is the expected baseline expression level in control group, α_{z_j} and β_k are the main effects of group and time on the gene expression level respectively, $\gamma_{z_j k}$ represents the interaction effect between group and time, and e_{ijk} is the independent measurement error with mean 0 and gene-dependent variance σ_i^2. There is only one random term, $r_{ij} \sim N(0, \tau_i^2)$, the replicate effect of individual gene, explaining the correlation between repeated measurements from one gene. If subjects in control and treatment groups are matched, the replicate effect r_{ij} can be modified to reflect the fact that the matched subjects share one random effect.

We combine some terms in model (6) and let $\mu_i + \beta_{ik} = \mu'_{ik}$ and $\alpha_{iz_j} + \gamma_{iz_jk} = \alpha'_{iz_jk}$. Then the new term μ'_{ik} has the interpretation of average temporal expression profile of ith gene at time t_k in the control group, and it is equivalent to $\mu_0(t_k) + \sum_{l=1}^{\infty} u_{il}\phi_l(t_k)$ in model (4). Similarly, α'_{iz_jk} can be interpreted as the difference of the mean expression profile between treatment and control groups at time t_k, and is equivalent to $\mu_1(t_k) - \mu_0(t_k) + \sum_{l=1}^{\infty} w_{il}\phi_l(t_k)$ in model (4). The replicate effect r_{ij} in model (6) is comparable to $\sum_{l=1}^{\infty}(v_{jl} + e_{ijl})\phi_l(t)$ in model (4). However, r_{ij} is constant over time in a two-way mixed ANOVA model, while FPCA model allows for a time-varying replicate effect.

2.5 Selection of Differentially Expressed Genes

Three methods can be used to identify differentially expressed genes based on the estimate $\hat{d}_i$. First, we can select the genes with $\hat{d}_i > 0.5$ based on the literal meaning of posterior probability. Alternatively, a cut-off value κ instead of 0.5 is used to control the false discovery rate (FDR). Earlier work about controlling the FDR instead of family-wise error can be seen in the landmark papers [4] and [18]. The latter defined the positive false discovery rate (pFDR) and described examples in which the pFDR can be written in the from of posterior error rate [18]. In this sense, the posterior error rate could be used to estimate the pFDR given a cut-off critical value

$$\mathrm{p\hat{F}DR}(\kappa) = \frac{\sum_{i:\, \hat{d}_i > \kappa}(1 - \hat{d}_i)}{\text{number of } \{i : \hat{d}_i > \kappa\}},$$

where $\hat{d}_i$ is previously defined posterior probability that the positive finding i is in reality negative. This data-based FDR control procedure is proposed in [13]. Finally, when the scientists have in mind the number of candidate genes for downstream work, we could just select top N genes based on the $\hat{d}_i$.

3 Application to *C. elegans* Data

We applied our method to the *C. elegans* data set [22], in which there are $n = 2430$ genes of interest across two conditions, L1 starvation and dauer exit. Under L1 starvation condition, the gene expression levels are supposed to be at the control level. Genes that show differential expressions during the dauer exit are of interest. Gene

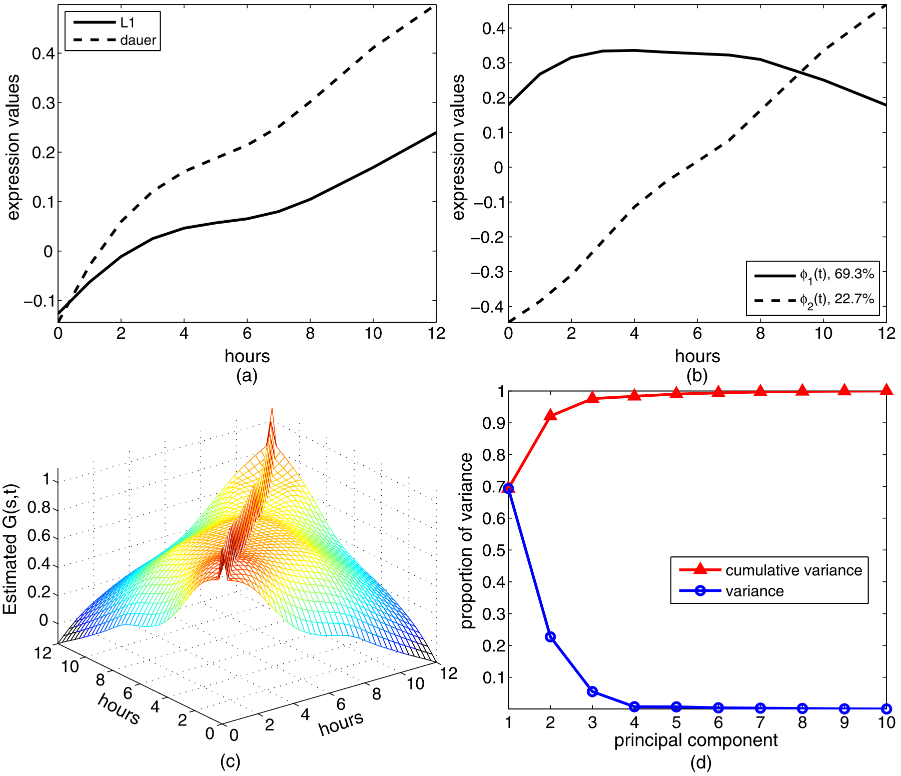


Fig. 1 (**a**) Mean functions of L1 and dauer groups. The bandwidths for the two mean functions are $h_1 = 2$ and $h_2 = 2$ respectively chosen by generalized cross-validation (GCV). (**b**) First two eigenfunctions extracted from the pooled covariance functions of two groups. (**c**) Estimated covariance function $\hat{G}(s,t)$ by pooling all genes in both L1 and dauer groups. The bandwidths for the covariance function are (1, 1) chosen by GCV. (**d**) Scree plot of the variance component percentage sizes

expression levels were measured at hours 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12 with 4 replications under each condition.

The FPCA method was applied to the combined data from both groups after the respective mean function is subtracted from each group. The mean functions for the two groups and the first two PCs are shown in Fig. 1, (a) and (b). Here, aided by the scree plot (Fig. 1(d)), we chose the first L eigenfunctions that cover at least 90% of the total variation in gene expression values. Two bases were extracted from $\hat{G}_X(s,t)$ that together cover about 92% of the variation in the data (Fig. 1(b)). It is interesting to observe from Fig. 1(b) that the first eigenfunction, which accounts for 69.3% of the total variation, is almost constant in the center portion of the time interval (3 to 8 hours) but shows a decline at both ends of the time interval. Thus, the highest variation of the gene trajectories occurs in the mid time zone with a vertical shift. While it is tempting to declare a constant vertical shift for the entire time duration, such a conclusion is not supported by the fact that all time-course data have been shifted to start from the origin. Moreover, there is a second PC which is linear in time and passes through the origin. This second PC explains about 22.7% of the variations,

Table 1 Estimates for variance components assuming the data with or without replicate effects. The bootstrap estimates (see Appendix A.4) for the standard deviations are shown in the parentheses

	with replicate effects		without replicate effects	
	$l=1$	$l=2$	$l=1$	$l=2$
$\sigma^2_{u,l}$	4.4572 (0.2339)	1.0344 (0.0395)	4.4572 (0.2600)	1.0344 (0.0395)
$\sigma^2_{w,l}$	2.9550 (0.4544)	2.5994 (0.1016)	2.9543 (0.4268)	2.5987 (0.1063)
$\sigma^2_{v,l}$	0.0113 (0.0282)	0.0027 (0.0086)	–	–
$\sigma^2_{e,l}$	2.1830 (0.1381)	0 (0.0168)	2.1944 (0.1309)	0 (0.0178)
π	0.8161 (0.0172)		0.8163 (0.0194)	

meaning that the slope of the linear random effect over time is another major source of variation for this time-course gene expression data. Another flexible random-effects model is studied in [7] along the directions of B-spline basis functions, but it needs more basis functions than the FPCA method to explain the same amount of variation and there is no clear biological meaning for these basis functions.

Figure 1(c) displays the estimated covariance surface for the observed expression profile over time, obtained by procedures in [24]. Clearly, the further apart the two time points are, the less correlated the measurements at these two points are. Although decaying over time, the correlation maintains positive even between points moderately far apart.

For this data set, we performed two separate analyses, one with and one without the replicate effects v_{jl} in the model. The rationale for the latter to exclude the replicate effects is because the replicated experiments were not conducted longitudinally, so the conventional "replicate" effect may not be applicable here. The result in Table 1 confirms that the replicate effects can be neglected since the within-replicate variation is larger than the between-replicate variation. In contrast, the models employed in [22] and [19] both contain only one constant random effect that is additive and assigned to the replicate effect. Our analysis suggests that such models may not be suitable for the *C. elegans* data.

The hybrid EM algorithm described in Appendices A.2 and A.3 is used to estimate the variance components associated with the first two PCs (reported in Table 1). The estimated variance of the measurement error for ϵ_{ijk} is 0.2247, and the estimate for the proportion of genes that are differentially expressed is about $\hat{\pi} = 0.816$, no matter the replicate effect v_{il} is included or not.

For this data set, Wang and Kim [22] identified 1984 out of the 2430 genes, which is about 81.64% of the genes and close to our estimate of 81.61%. However, there is a substantial difference between the two methods in terms of the selected genes (see Table 2) and this is further elaborated later. The genes identified by both methods as differential expression constitute the majority of the initial list with 2430 genes. This is high but could be explained by the fact that the initial 2430 genes already show significant change over time during the dauer exit process so we only need to exclude those with expression changes induced by food, which might be a small portion of the genes. After we adjust for an FDR = 0.01 and 0.05, corresponding to a critical value of 0.9192 and 0.6667 respectively for $\hat{d}_i$, the FPCA approach selected 1380 and 1767 genes.

Table 2 Identified gene numbers by the proposed method, two-way mixed ANOVA, EDGE and Liu and Yang's method, adjusted at FDR levels of 0.01 and 0.05. For the proposed method, the FDR is controlled using the procedure in Newton et al. [13]. The FDR procedure used in ANOVA, EDGE and Liu and Yang's method is based on Benjamini and Hochberg [4]. The number of genes overlapped with the proposed method are listed in the parentheses

	Identified Gene Number	
	FDR = 0.05	FDR = 0.01
Proposed Method	1767	1380
ANOVA	1949 (1602)	1681 (1199)
EDGE	1747 (1464)	1551 (1100)
Liu and Yang's Method	366 (362)	0 (0)

It is also of interest to compare the proposed method with the gene-specific B-spline method, EDGE [9, 19], and the gene-specific FPCA method in [12]. In Table 2, we check the overlap of our results with these two methods, together with the ANOVA approach in [22] when adjusting at FDR levels 0.01 and 0.05. Figure 2 shows the gene expression profiles chosen as differentially expressed by our method but not by the others, while Fig. 3 reflects the opposite scenario showing the profiles of genes chosen by our method only. The conventional mixed-effects approach tends to select genes with little between-replicate variation within groups, since this between-replicate variation is estimated only by the information from the gene itself. With large variation present, the real difference between the groups might be obscured by the variation of replications. EDGE seems to fail in some of the cases with big difference in the two groups. The method in [12] by Liu and Yang has substantially different selection from the other methods, which only claimed 366 DE genes with an FDR of 0.05 and 0 with an FDR of 0.01. In most of cases in Fig. 3, the expression profiles in L1 and dauer groups cross with each other, which suggests that this pattern is probably not easily identified by Liu and Yang's method. The comparison confirmed that the proposed method can pick up patterns the other methods failed to identify.

4 Simulation Studies

Initially, we aimed at comparing the performance of our approach to EDGE, a two-way mixed ANOVA model based on B-splines, and the gene-specific FPCA method in [12] via simulation. Because the method in [12] took too much computing time, we were not able to include its results. The rest of this section focuses on the comparison of the three remaining methods. The simulated data are generated either under the FPCA model described in (4) or under a two-way mixed ANOVA model in (6). Under either model, the data were simulated to mimic the *C. elegans* data set. For each of the simulation, a sample of $J = 8$ replicates (four replicates for each one of the two groups), and each replicate with a comparable number of $n = 2500$ genes as in *C. elegans* data set, was generated.

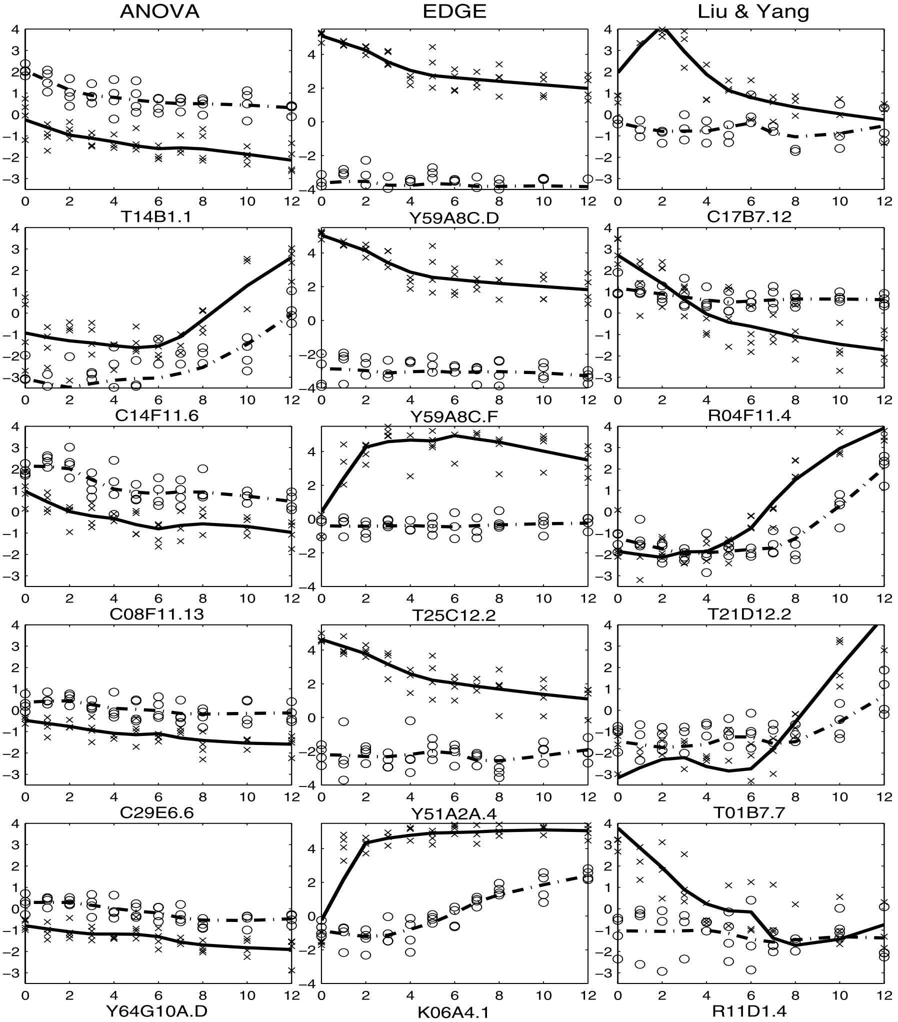


Fig. 2 Genes in *C. elegans* data selected by the proposed method but not by two-way mixed-effects ANOVA (*left column*), EDGE (*middle column*) and Liu and Yang's method (*right column*) when adjusting at FDR = 0.05. *Solid line*: estimated mean expression trajectories for dauer group; *cross mark*: observed expression values for dauer group; *dashed line*: estimated mean expression trajectories for L1 group; *circle*: observed expression values for L1 group. The *x-axis* is time in hours and the *y-axis* is the normalized gene expression intensity

4.1 Simulation under the FPCA Model

We first generated the expression trajectories according to the FPCA model (see (4)) under a range of settings, for $K = 5, 20$ measurement time points and DE proportion $\pi = 0.1, 0.4, 0.8$. The time grid is placed with equal distances from 0 to 1. The population mean expression functions for the two groups are simulated to mimic those

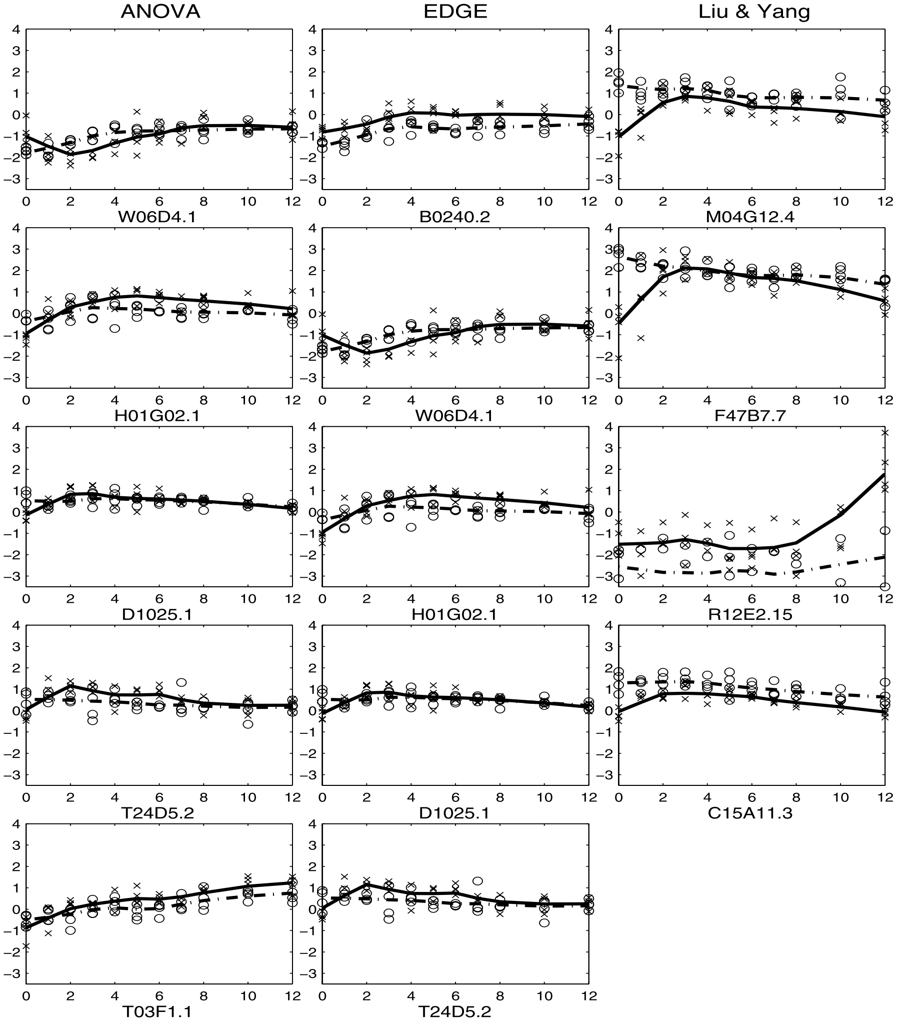


Fig. 3 Genes in *C. elegans* data NOT selected by the proposed method but by two-way mixed-effects ANOVA (*left column*), EDGE (*middle column*) and Liu and Yang's method (*right column*) when adjusting at FDR = 0.05. *Solid line*: estimated mean expression trajectories for dauer group; *cross mark*: observed expression values for dauer group; *dashed line*: estimated mean expression trajectories for L1 group; *circle*: observed expression values for L1 group. The *x-axis* is time in hours and the *y-axis* is the normalized gene expression intensity

estimated from the *C. elegans* data,

$$\mu_1(t) = 0.07\bigl(t + 0.15\sin(2\pi t)\bigr)$$

and

$$\mu_2(t) = 0.04\bigl(t + 0.15\sin(2\pi t)\bigr), \quad t \in [0, 1].$$

Only two principal component functions are considered. They are orthonormalized Legendre polynomials of degrees 1 and 2:

$$\phi_1(t) = \sqrt{3}(2t-1) \tag{7}$$

and

$$\phi_2(t) = \sqrt{5}\big(6t^2 - 6t + 1\big), \tag{8}$$

with similar shapes as in *C. elegans* data. The random effects we considered in the simulated model include gene effects u_{il}, differential effects w_{il} and gene-specific replicate effects e_{ijk}, as well following the structure of (4). We set the variance of the random noise imposed on the expression function as $\sigma_\epsilon^2 = 0.25$, and the variance components as listed in the second column of Table 3.

We perform 100 simulations and summarize the results of parameter estimation in Table 3. The FPCA approach successfully identified two major principal components

Table 3 Biases and rooted mean square errors (RMSEs) for the proposed method under different scenarios. The results are based on 100 simulations for $n = 2500$ genes each with 4 replicates in both control and treatment groups. The number of time points in the simulation is set to $K = 5$ and $K = 20$, and the probability of differential expression takes value of $\pi = 0.1$, $\pi = 0.4$ and $\pi = 0.8$ respectively

Parameter	True Value	$K = 5$		$K = 20$	
		Bias	RMSE	Bias	RMSE
π	0.1	0.001	0.010	0.002	0.010
σ_{u1}^2	5.0	−0.015	0.155	−0.013	0.147
σ_{u2}^2	1.0	−0.003	0.033	0.000	0.032
σ_{w1}^2	3.5	0.044	1.010	−0.011	0.963
σ_{w2}^2	2.5	−0.020	0.348	−0.009	0.374
σ_{e1}^2	2.0	0.003	0.036	0.003	0.034
σ_{e2}^2	0.1	0.032	0.033	−0.005	0.008
π	0.4	0.000	0.015	0.001	0.016
σ_{u1}^2	5.0	−0.009	0.155	0.009	0.150
σ_{u2}^2	1.0	−0.004	0.034	0.000	0.033
σ_{w1}^2	3.5	−0.014	0.341	0.019	0.340
σ_{w2}^2	2.5	−0.001	0.150	−0.004	0.156
σ_{e1}^2	2.0	0.008	0.041	0.000	0.039
σ_{e2}^2	0.1	0.032	0.033	−0.005	0.010
π	0.8	0.001	0.015	0.000	0.014
σ_{u1}^2	5.0	0.003	0.157	−0.014	0.168
σ_{u2}^2	1.0	−0.002	0.037	−0.001	0.037
σ_{w1}^2	3.5	−0.010	0.224	−0.009	0.220
σ_{w2}^2	2.5	−0.006	0.104	−0.007	0.105
σ_{e1}^2	2.0	0.003	0.047	0.002	0.046
σ_{e2}^2	0.1	0.033	0.035	−0.003	0.013

Table 4 Comparison of the proposed method, two-way mixed-effects ANOVA and EDGE by achieved false positive rates (FPR, proportion of true negatives identified as positives) and false negative rates (FNR, proportion of true positives identified as negatives) at chosen cut-offs (κ's). For the proposed method, two practical cut-offs described in Sect. 2.3 were used. For ANOVA and EDGE, the algorithm, illustrated in Benjamini and Hochberg [4], was used to control FDR at 0.01. The data are simulated from the FPCA model in (4) and (5)

	$K=5$						$K=20$					
	$\pi=0.1$		$\pi=0.4$		$\pi=0.8$		$\pi=0.1$		$\pi=0.4$		$\pi=0.8$	
	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR
Proposed Method[a]	0.006	0.405	0.049	0.274	0.397	0.098	0.007	0.417	0.052	0.280	0.413	0.097
Proposed Method[b]	0.001	0.530	0.004	0.436	0.027	0.318	0.001	0.548	0.004	0.452	0.026	0.333
ANOVA	0.067	0.497	0.087	0.445	0.103	0.413	0.158	0.379	0.178	0.359	0.198	0.343
EDGE	0.025	0.644	0.040	0.576	0.052	0.536	0.411	0.197	0.408	0.201	0.418	0.196

[a] $\kappa=0.5$

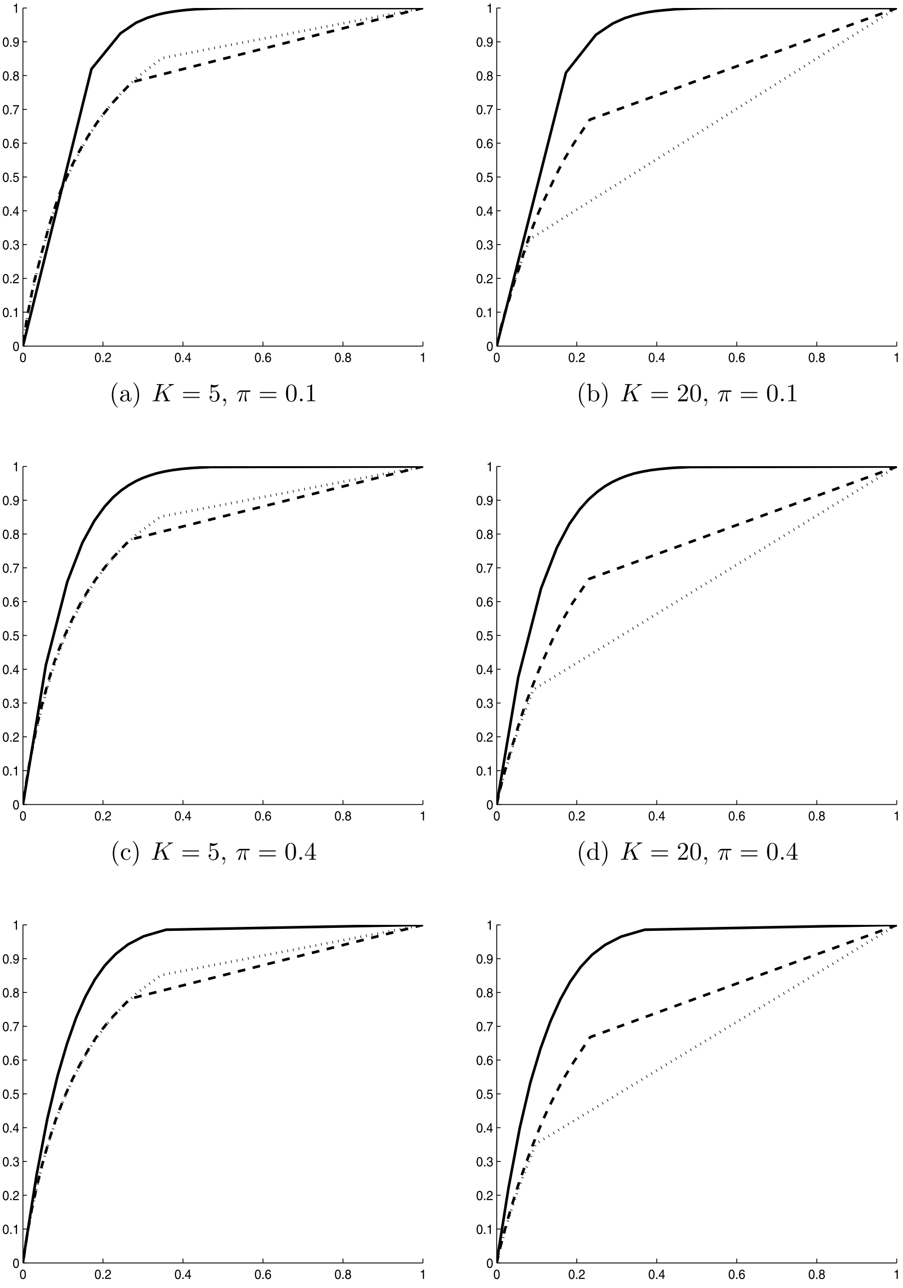
[b] κ is chosen to control FDR at level 0.01 by method described in [13]

as we used to simulate the data. The estimates of all variance components and the differential probability π appear to be virtually unbiased. There is some variance increase for the estimators of $\sigma^2_{w,l}$, $l=1,2$ as the differential probability π decreases. This is presumably because a lesser number of genes can be used to estimate the $\sigma^2_{w,l}$ when the proportion of differentially expressed genes, π, shrinks. As evident form this table, our approach achieves the same level of accuracy and precision even when the measurements were taken at as few as 5 time points. Our explanation is that the underlying expression trajectories we simulated are fairly smooth (polynomials of degree 2) and the time points are evenly spaced in the entire interval, and thus five points can capture most of the curve dynamics, suggesting that under certain situations FPCA method may work as well as when measurement points are only a few for each gene.

Next, we compare the performance of our FPCA method in detecting differentially expressed genes with the two-way mixed ANOVA method and the EDGE method, based on the practical false positive rate (FPR) and false negative rate (FNR) in Table 4, and on the Receiver Operating Characteristic (ROC) curves in Fig. 4 under the same six different settings above. Here the ROC curve of a test shows its performance as a trade-off between specificity (true negative rate) and sensitivity (true positive rate). Typically a curve of sensitivity versus (1—specificity) is plotted while a false positive rate or threshold parameter is varied. The tests with better performance are those closer to the zero false positive axis or with a higher true positive rate. The proposed FPCA method clearly outperforms the two-way mixed ANOVA method and EDGE in all settings tried here.

4.2 Simulation under Two-way Mixed-effects ANOVA Model

In the previous section, the data were generated from model (4) with added noises, and this would seem to favor the proposed FPCA approach. In this section, we compare the three approaches based on the two-way mixed ANOVA model in (6), giving



(a) $K = 5$, $\pi = 0.1$

(b) $K = 20$, $\pi = 0.1$

(c) $K = 5$, $\pi = 0.4$

(d) $K = 20$, $\pi = 0.4$

(e) $K = 5$, $\pi = 0.8$

(f) $K = 20$, $\pi = 0.8$

Fig. 4 Receiver Operating Characteristic curves (or ROC curves) of proposed FPCA model (*solid lines*), two-way mixed ANOVA model (*dashed lines*) and EDGE (*dotted lines*) under the six settings listed in Table 3. The *x-axis* is (1—specificity) and the *y-axis* is the sensitivity

the ANOVA approach an advantage over other procedures. Because of space limitation, we investigate only one of the six cases in Sect 4.1, when $n = 2500$, $K = 5$ and $\pi = 0.1$. This setting contains fewer time points and smaller proportion of DE genes, both seem to favor the ANOVA approach. We also use this setting to further explore the robustness performance of all three approaches.

We randomly picked 10% of 2500 genes as differentially expressed, and then generated the combined terms in ANOVA model ((6) in Sect. 2.4) for each gene by $\mu_i + \beta_{ik} = \sum_{l=1}^{2} a_{il}\phi_l(t_k)$, $\alpha_{iz_j} + \gamma_{iz_jk} = \sum_{l=1}^{2} c_{il}\phi_l(t_k)$, where $a_{il} \sim N(0, \sigma_{al}^2)$, $c_{il} \sim N(0, \sigma_{cl}^2)$, and $\phi_l(t)$ is the orthonormalized Legendre polynomial of degree l as in (7) and (8). If gene i is chosen as a non-DE gene, we let $c_{il} = 0$. The variances for random variables a_{il} and r_{il}, $l = 1, 2$, are $(\sigma_{a1}^2, \sigma_{a2}^2) = (2, 0.5)$ and $(\sigma_{c1}^2, \sigma_{c2}^2) = (3, 1)$. The random effect r_{ij} is generated from $N(0, 0.25)$ and the error term e_{ijk} from $N(0, 0.25)$.

To check the robustness of our method and other methods when the normality assumption for the expression values is violated, we simulated an additional 100 non-normal data sets by first generating normal variables which were then taken to the power of (4/3). This power was chosen as the power transformation (0.75) found to be best for *C. elegans* data to fit normal distribution. We also tested the proposed method in the presence of outliers by generating another set of data containing 1% outliers. The outliers were generated by first randomly picked 1% of expression values and then increasing the selected values by twice standard deviation with their sign preserved (i.e., a value of $2 * \text{sign}(y_{ijk})\text{sd}(y_{ijk})$ was added to the selected y_{ijk}).

The false positive and false negative rates with FDR controlled at 0.01 are reported in Table 5. As expected, the ANOVA model has the best overall performance, while the proposed method does not suffer a great deal of loss in terms of FPR and FNR. The non-normal setting triggers the most increase in FPR and affects all the methods, but the FNR does not seem to be affected. The outliers have a greater influence on the proposed method than the ANOVA approach (when compared to the normal data without outliers) in terms of FPR, but the opposite holds for FNR. All the increases

Table 5 Comparison of the proposed method, two-way mixed-effects ANOVA and EDGE by achieved FPR and FNR using 100 simulations under two-way mixed-effects ANOVA model. We considered the scenarios of whether expression values satisfy normality assumption, and whether the outliers or the missing values in specific genes exist. Two practical cut-offs described in Sect. 2.3 were used for the proposed method. The cut-offs for ANOVA and EDGE were chosen to control FDR at 0.01 using method in Benjamini and Hochberg [4]. All data are simulated from the ANOVA model in (6)

$\pi = 0.1$	normal		non-normal		outlier 1%	
	FPR	FNR	FPR	FNR	FPR	FNR
Proposed Method[a]	0.0008	0.156	0.185	0.145	0.066	0.168
Proposed Method[b]	0.0003	0.174	0.127	0.198	0.0054	0.210
ANOVA	0.006	0.088	0.148	0.099	0.004	0.160
EDGE	0.057	0.663	0.091	0.620	0.121	0.651

[a] $\kappa = 0.5$

[b] κ is chosen to control FDR at level 0.01 by method described in [13]

in FPR and FNR are within an acceptable tolerance range and both the ANOVA and the proposed approach generally outperform EDGE. We thus conclude that the FPCA approach is suitable for the *C. elegans* data and more generally for other time-course data as well.

5 Discussion

We proposed a new FPCA approach to analyze time-course gene expression data, with two important ingredients: first, it is data-adaptive and allows a flexible and natural way to model the data; second, it combines information across all genes to compensate for the small number of replicates per gene, and provides a powerful test for identification of DE genes.

FPCA is a nonparametric method well suited for continuous trajectory data that are increasingly common [10, 11]. Extending the current FPCA method [24] that extracts dominant dynamics of the temporal process first and then the FPC scores, we have developed a method to identify a subset of genes as differentially expressed, based on the FPC scores. In the study presented here, we propose a mixture of normal distributions on the FPC scores, to explain the extra dispersion of FPC scores associated with the treatment group caused by a subset of DE genes. Further extension of the model includes a nonparametric approach, rather than relying on the normal assumption, to model the different distributions of FPC scores in the control and treatment groups.

The idea of information borrowing has been exploited for a long time in microarray studies of cross-sectional design [6, 21]. Currently, there are several effective methods to implement this general idea. Some explicitly fulfill the aim, e.g., by imposing a hierarchical Bayesian model on the basis-induced coefficients [7] or on the original multivariate observations [20], and others do so implicitly, e.g. through a Hidden Markov model as in [25]. A key difference between our approach and these existing methods is that we use a flexible statistical models without specifying any prior structures on the shape of the gene trajectories and correlations between time points. We estimate the correlation structure from the data directly by nonparametric method.

Our model accommodates the correlations among genes from the same replicate through shared random effects. But because the correlations do not come from whether they are differentially expressed, $\hat{d}_i$ does not use the expression profiles from the other genes to decide whether the ith gene is differentially expressed. We can extend our model to such that if we know how the d_i's are correlated then we can use this information to improve the estimate.

The designation of a gene as differential expression depends on the variance of trajectories as well as the difference of trajectories between the two groups. Two-way ANOVA model estimates gene-specific variance, while the FPCA model estimates the variance by pooling all the trajectories and smoothing over time points, thus borrowing the information across genes and time points. Due to the small sample size, the gene-specific variance is usually not estimated very well and the test statistic can perform poorly. The influence of different variance estimates can be seen from their respective tendency of picking genes as differentially expressed. The benefit of borrowing information across all genes as compared to other gene-specific approaches, such as EDGE and the gene-specific FPCA approach in [12], have been amply illustrated.

Through the simulations, we found that the violation of normality assumption resulted in a greater increase of false positive rate for the proposed method as well as other methods. Our suggestion is to check the normality assumption first and to transform the data appropriately to resemble a normal distribution. The traditional Box–Cox family of transformation, defined by $(y^\theta - 1)/\theta$, can be applied to the data, with the parameter θ selected to achieve normality.

To obtain a benchmark on computing (i.e., CPU) time, we ran the proposed method, two-way mixed ANOVA, the EDGE method and Liu and Yang's method for *C. elegans* data on a dual quad core processor 2.27 GHz Windows 7 64-bit system with 12 GB RAM. Approximately 75 seconds for the proposed method, 5 minutes for ANOVA, 25 minutes for EDGE and over 60 hours for Liu and Yang's method were required for completion. Our algorithm is computationally fast and thus allows the analysis of very large genome wide data sets with thousands of genes. Perhaps the most important advantage is that the temporal variation in the data is biologically meaningful and can be captured by only a few data-adaptive bases. All the other methods used predetermined basis function which may not have a biological interpretation and may increase the number of basis functions needed if the gene expression profiles over time are not smooth enough. For the *C. elegans* data studied in this paper, the expression profiles are fairly smooth, so the FPCA method needs only two eigenfunctions to explain over 90% of the temporal variation. We expect the FPCA method to be broadly useful in more general settings.

In summary, we have developed a fast algorithm for the selection of genes whose expression differentiates in two groups. After extracting meaningful dimensions from a time-course data set using FPCA, we pick a subset of genes that retain the information carried in the full data set. We expect that FPCA-based algorithms will be an invaluable tool for the analysis of ever-increasing genome-wide data.

Acknowledgement We thank two referees and the editor, Xihong Lin, for insightful comments that led to substantial improvement of the paper.

Appendix

A.1 Parameter Estimation

To apply current FPCA approaches, we first take the average of the observed profiles over the J_g replications within group g, $g = 0, 1$. This yields the mean expression values in group g, $\bar{y}_{igk} = \frac{1}{J_g}\sum_{\{z_j=g\}} y_{ijk}$, where z_j is the indicator of group membership of the jth replicate. This averaging requires each replicate be sampled at the same time points. In case of missing measurements in any of the J_g replicates we can impute the missing values, or in case the time-course data are sampled at different times for the replicates we can estimate the mean profiles through a nonparametric scatter plot smoother similarly to the one described below for the mean expression profile $\mu_g(t)$ for each group. Then model (5) becomes

$$\bar{y}_{igk} = \bar{X}_{ig}(t_{ik}) + \bar{\epsilon}_{igk},$$

where $\bar{X}_{ig}(t) = \frac{1}{J_g}\sum_{\{z_j=g\}} X_{ijk}(t)$ is the mean expression process, and $\bar{\epsilon}_{igk} = \frac{1}{J_g}\sum_{\{z_j=g\}} \epsilon_{ijk}$. The mean expression function based on model (1) can thus be written as

$$\bar{X}_{ig}(t) = \mu_g(t) + \sum_{l=1}^{\infty} \bar{b}_{igl}\phi_l(t), \quad a \le t \le b, \tag{9}$$

where $\bar{b}_{igl} = \frac{1}{J_g}\sum_{\{z_j=g\}} b_{ijl}$.

The following shows that the mean expression process over replicates possesses the same eigenfunctions as individual expression process. The covariance function of the average observed profile, by definition, is $\operatorname{cov}\{\bar{y}_{igk_1}, \bar{y}_{igk_2}\} = \operatorname{cov}\{\bar{X}_{ig}(t_{igk_1}), \bar{X}_{ig}(t_{igk_2})\} + \frac{\sigma_\epsilon^2}{J_g}\delta_{k_1k_2}$, where $\delta_{k_1k_2}$ is the Kronecker delta function taking value 1 if $k_1 = k_2$ or value 0 if $k_1 \neq k_2$. It is easy to see that the covariance function of mean expression function is $\operatorname{cov}\{\bar{X}_{ig}(t), \bar{X}_{ig}(s)\} = \sum_{l=1}^{\infty} \frac{\lambda_l}{J_g}\phi_l(t)\phi_l(s)$. The eigenvectors of covariance function for the original expression function $X_{ij}(t)$ and for the averaged expression function $\bar{X}_{ig}(t)$ are the same. They only differ at the variation size of random coefficients as a result of averaging the profiles. Thus consistent estimators for eigenvectors $\{\phi_l(t)\}$ based on averaged expression profiles will also be consistent estimators for eigenvectors of original profiles. The percentage of variation explained by each principal component also remains the same. This shows that we can use the averaged profiles to estimate eigenvectors in model (1) and choose appropriate number of components.

By local weighted least squares (LWLS) smoothing method, which minimizes

$$\sum_{i,g,k} K\left(\frac{t - t_{igk}}{h}\right)\bigl[\bar{y}_{igk} - \bigl(\beta_0 + \beta_1(t - t_{igk})\bigr)\bigr]^2$$

over $\beta_0,\ \beta_1$, we obtain estimator for the mean function $\hat{\mu}(t) = \hat{\beta}_0$. In the above formula, t_{igk} is the time point with at least one observation for all replications in group g, $\bar{y}_{igk}$ is the average curve, h is the bandwidth for smoothing and $K(\cdot)$ is the kernel function which weights neighboring points. Subsequently, the individual deviation from the mean function is estimated by subtracting $\hat{\mu}_g(t_{igk})$ from $\bar{y}_{igk}$. With raw covariance $G_{ig}(t_{igk_1}, t_{igk_2}) = (\bar{y}_{igk_1} - \hat{\mu}(t_{igk_1}))(\bar{y}_{igk_2} - \hat{\mu}(t_{igk_2})), k_1 \neq k_2$, the covariance function $\hat{G}(s,t)$ is estimated by two-dimensional linear LWLS, minimizing

$$\sum_{i,g,k_1\neq k_2} K\left(\frac{s - t_{igk_1}}{h}, \frac{t - t_{igk_2}}{h}\right)\bigl[G_{ig}(t_{igk_1}, t_{igk_2}) \\ - \bigl(\beta_0 + \beta_1(s - t_{igk_1}) + \beta_2(t - t_{igk_2})\bigr)\bigr]^2$$

over β_0, β_1, β_2, with $\hat{G}(s,t)=\hat{\beta}_0$. The technical details can be seen in Yao et al. [24]. The bandwidths used to smooth the mean and covariance functions are chosen by generalized cross-validation (GCV) method, which minimizes the generalized leave-one-out prediction error. The covariance function $\hat{G}(s,t)$ is then evaluated at a fine grid and the eigenvectors of the resulting matrix, $\hat{\phi}_l(t)$, can then be computed numerically. In practice, a finite number of principal components are used. There are several ways to select the appropriate number of principal components, such as AIC, BIC or use the scree plots to decide the suitable number of L so that a high percentage of the variation in the data can be explained by these L components. Here we recommend against the use of cross-validation method for such a method selection purpose, because cross-validation tends to overfit the data with more random effects resulting in a larger L than practically needed. The L eigenvectors truncated therein form an orthonormal basis to represent the true gene expression trajectories.

After removing the smoothed group mean and projecting the centered observed gene trajectories onto the corresponding eigenvector, we obtain the prediction of random coefficient, $\hat{b}_{ijl}$, on lth eigenfunction. Since $b_{ijl}=\int_a^b (X_{ij}(t)-\mu_{z_j}(t))\phi_l(t)\,dt$, the estimate of b_{ijl} can be obtained by substitute the respective estimators into the above formula,

$$\hat{b}_{ijl}=\int_a^b \big(X_{ij}(t)-\hat{\mu}_{z_j}(t)\big)\hat{\phi}_l(t)\,dt,$$

where $\hat{\mu}_{z_j}(t)$ and $\hat{\phi}_l(t)$ are the estimated functions described previously. But the trajectory $X_{ij}(t)$ is not completely observed for all t. Therefore, this inner product is approximated by trapezoid rule. For sparse data, the scores can be estimated by functional principal component analysis through conditional expectation (PACE) [24]. Using the PC scores, we can then predict the entire expression profile for each gene.

Then we estimate the variances of each random component, $\sigma^2_{u,l}$, $\sigma^2_{w,l}$, $\sigma^2_{e,l}$ based on $\hat{b}_{ijl}$ by the least squares method, minimizing

$$C_{LS}=\sum_{(i,j,l),(i',j',l)} \big(\hat{b}_{ijl}\hat{b}_{i'j'l}-\mathrm{Cov}(\hat{b}_{ijl},\hat{b}_{i'j'l})\big)^2, \qquad (10)$$

since the target $\mathrm{Cov}(\hat{b}_{ijl},\hat{b}_{i'j'l})$ is a function of $\sigma^2_{u,l}$, $\pi\sigma^2_{w,l}$, $\sigma^2_{e,l}$ and σ^2_{ϵ} (see Appendix A.2). However, the indicator of whether a gene is differentially expressed, d_i, has to be known in order to estimate variance $\sigma^2_{w,l}$. For this reason, we treat d_i as the missing data and use EM algorithm to estimate the variance explained by differential expression, $\sigma^2_{w,l}$, after we estimate all the other variance components through least squares method. The details can be seen in Appendix A.3. Given the estimated variance components, we then estimate the probability of gene i being differentially expressed given observed expression profiles

$$\hat{d}_i=\hat{E}[d_i|\mathbf{y}]=\widehat{\mathrm{Prob}}[d_i=1|\mathbf{y}],$$

where $\mathbf{y}$ denotes all the observed expression data. This provides a criterion to select a list of most significantly differentially expressed genes.

A.2 Least Squares Method

Suppose we have truncated the number of principal components to a finite number L in practice. Denote $\mathbf{b}_{il}^T = (b_{i1l}, \dots, b_{iJl})$ as the vector containing all the principal component scores on lth eigenfunction associated with ith gene. Write the model for $\mathbf{b}_{il}$ in the matrix form

$$\mathbf{b}_{il} = u_{il}\mathbf{1}_J + w_{il}\mathbf{z} + \mathbf{v}_l + \mathbf{e}_{il}, \tag{11}$$

where $\mathbf{1}_J$ is a $J \times 1$ vector of all ones, $\mathbf{v}_l^T = (v_{1l}, \dots, v_{Jl})$, $\mathbf{z}^T = (z_1, \dots, z_J)$ and $\mathbf{e}_{il}^T = (e_{i1l}, \dots, e_{iJl})$. The coefficients of all n genes corresponding to lth eigenfunction are vectorized such that $\mathbf{b}_l^T = (\mathbf{b}_{1l}^T, \dots, \mathbf{b}_{nl}^T)$ is a vector of dimension $nJ \times 1$. Then

$$\mathbf{b}_l = \mathbf{u}_l \otimes \mathbf{1}_J + \mathbf{w}_l \otimes \mathbf{z} + \mathbf{1}_n \otimes \mathbf{v}_l + \mathbf{e}_l, \tag{12}$$

where $\otimes$ is the Kronecker product, $\mathbf{u}_l^T = (u_{1l}, \dots, u_{nl})$, $\mathbf{w}_l^T = (w_{1l}, \dots, w_{nl})$, and $\mathbf{e}_l^T = (e_{1l}^T, \dots, e_{nl}^T)$. Let $B_l = \mathbf{b}_l\mathbf{b}_l^T$, then

$$E(B_l) = \mathrm{Var}(\mathbf{b}_l) = \sigma_{u,l}^2 I_n \otimes \mathbf{1}_{J\times J} + \pi\sigma_{w,l}^2 I_n \otimes \mathbf{z}\mathbf{z}^T + \sigma_{v,l}^2 \mathbf{1}_{n\times n} \otimes I_J + \sigma_{e,l}^2 I_n \otimes I_J. \tag{13}$$

Here, $\mathbf{1}_{m\times n}$ denotes an $m \times n$ all-ones matrix.

Similarly, we vectorize the estimated coefficients on lth eigenfunction $\hat{b}_{ijl} = \int (y_{ij}(t) - \hat{\mu}_{z_j}(t))\hat{\phi}_l(t)\,dt$, for $i = 1, \dots, n$ and $j = 1, \dots, J$, and denote the vector by $\hat{\mathbf{b}}_l$. If $\hat{\phi}_l(t)$ is a consistent estimator for $\phi_l(t)$, then $\hat{b}_{ijl} = b_{ijl} + \int \epsilon(t)\phi_l(t)\,dt + o(1)$ and $E(\hat{b}_{ijl}^2) = E(b_{ijl}^2) + \sigma_\epsilon^2 + o(1)$. Thus, the expectation of $\hat{B}_l = \hat{\mathbf{b}}_l\hat{\mathbf{b}}_l^T$ is

$$E(\hat{B}_l) = \mathrm{Var}(\hat{\mathbf{b}}_l) = \mathrm{Var}(\mathbf{b}_l) + \sigma_\epsilon^2 I_n \otimes I_J + o(1). \tag{14}$$

The term $o(1)$ in (14) is negligible. We will ignore the influence of this term in the following computation. Now the minimization criterion in (10) in Appendix A.1 could be rewritten as

$$\begin{aligned} C_{LS} &= \mathrm{tr}\{(\hat{B}_l - E(\hat{B}_l))(\hat{B}_l - E(\hat{B}_l))^T\} \\ &= \mathrm{tr}\{\hat{B}_l\hat{B}_l^T - 2E(\hat{B}_l)\hat{B}_l^T + E(\hat{B}_l)E(\hat{B}_l)^T\}. \end{aligned} \tag{15}$$

Note that under the least squares approach, only $\pi\sigma_{w,l}^2$ is identifiable, but not π and $\sigma_{w,l}^2$. We thus first proceed to estimate $\pi\sigma_{w,l}^2$ and then use the EM-algorithm in Appendix A.3.

Taking derivative with respect to the $\sigma_{u,l}^2, \pi\sigma_{w,l}^2, \sigma_{v,l}^2, \sigma_{e,l}^2$ and setting them to be 0, we have linear system to solve

$$\frac{\partial C_{LS}}{\partial \sigma_{u,l}^2} = -2[\mathrm{tr}\{I_n \otimes \mathbf{1}_{J\times J}\hat{B}_l^T\} - \mathrm{tr}\{I_n \otimes \mathbf{1}_{J\times J}E(\hat{B}_l)^T\}]$$

$$= -2\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\hat{b}_{ijl}\right)^{2} - n\big(J^2\sigma_u^2 + J_2^2\big(\pi\sigma_{w,l}^2\big) + J\sigma_{v,l}^2 + J\sigma_{e,l}^2 + J\sigma_\epsilon^2\big)\right]$$

$$= 0, \tag{16a}$$

$$\frac{\partial C_{LS}}{\partial(\pi\sigma_{w,l}^2)} = -2\big[\mathrm{tr}\big\{I_n \otimes ZZ^T\hat{B}_l^T\big\} - \mathrm{tr}\big\{I_n \otimes ZZ^T E(\hat{B}_l)^T\big\}\big]$$

$$= -2\left[\sum_{i=1}^{n}\left(\sum_{\{j:z_j=1\}}\hat{b}_{ijl}\right)^{2}\right.$$

$$\left. - n\big(J_2^2\sigma_{u,l}^2 + J_2^2\big(\pi\sigma_{w,l}^2\big) + J_2\sigma_{v,l}^2 + J_2\sigma_{e,l}^2 + J_2\sigma_\epsilon^2\big)\right] = 0, \tag{16b}$$

$$\frac{\partial C_{LS}}{\sigma_{v,l}^2} = -2\big[\mathrm{tr}\big\{\mathbf{1}_{n\times n} \otimes I_J\hat{B}_l^T\big\} - \mathrm{tr}\big\{\mathbf{1}_{n\times n} \otimes I_J E(\hat{B}_l)^T\big\}\big]$$

$$= -2\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{b}_{ijl}\right)^{2}\right.$$

$$\left. - n\big(J\sigma_{u,l}^2 + J_2\big(\pi\sigma_{w,l}^2\big) + nJ\sigma_{v,l}^2 + J\sigma_{e,l}^2 + J\sigma_\epsilon^2\big)\right] = 0, \tag{16c}$$

$$\frac{\partial C_{LS}}{\sigma_{e,l}^2} = -2\big[\mathrm{tr}\big\{I_n \otimes I_J\hat{B}_l^T - I_n \otimes I_J E(\hat{B}_l)^T\big\}\big]$$

$$= -2\left[\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{b}_{ijl}^2 - n\big(J\sigma_{u,l}^2 + J_2\big(\pi\sigma_{w,l}^2\big) + J\sigma_{v,l}^2 + J\sigma_\epsilon^2\big)\right] = 0, \tag{16d}$$

subject to $\sigma_{u,l}^2,\ \sigma_{w,l}^2,\ \sigma_{v,l}^2,\ \sigma_{e,l}^2 > 0$. Note that π and $\sigma_{w,l}^2$ are not identifiable by the least squares method alone. We will use EM algorithm to solve them in the next section.

A.3 EM Algorithm

The least squares method yields the estimates $\hat{\sigma}_{u,l}^2, \widehat{\pi\sigma_{w,l}^2}, \hat{\sigma}_{v,l}^2$ and $\hat{\sigma}_{e,l}^2, l = 1, \ldots, L$. The EM algorithm here is used to disentangle the π and $\sigma_{w,l}^2$. Denote the vector of all random coefficients by $\mathbf{b} = (\mathbf{b}_i, i = 1, \ldots, n)$, where $\mathbf{b}_i = (\mathbf{b}_{il}, l = 1, \ldots, L)$ with $\mathbf{b}_{il}$ defined in (11). We assume a normal distribution for $\mathbf{b}_{il}$ conditional on d_i, $\mathbf{b}_{il}|d_i \sim N(0, (\sigma_{u,l}^2 + d_i\sigma_{w,l}^2)\mathbf{1}_{J\times J} + (\sigma_{v,l}^2 + \sigma_{e,l}^2)I_J)$. In this step, we will work on the principal component scores $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_i, i = 1, \ldots, n)$ obtained from the FPCA step, instead of $\mathbf{b}$ directly. The distribution of $\hat{\mathbf{b}}_{il}$ is approximated by $N(0, (\sigma_{u,l}^2 + d_i\sigma_{w,l}^2)\mathbf{1}_{J\times J} + (\sigma_{v,l}^2 + \sigma_{e,l}^2 + \sigma_\epsilon^2)I_J)$.

With $\hat{\sigma}^2_{u,l}$, $\widehat{\pi\sigma^2_{w,l}}$, $\hat{\sigma}^2_{v,l}$ and $\hat{\sigma}^2_{e,l}$, $l = 1, \dots, L$, held fixed, we initialize $\pi^{(0)}$ by a random value between 0 and 1, and $\sigma^{2(0)}_{w,l} = \widehat{\pi\sigma^2_{w,l}}/\pi^{(0)}$. At the E-step of $(m+1)$-th iteration, the missing d_i is filled by

$$\begin{aligned}
\hat{d}_i^{(m+1)} &= E(d_i|\hat{\mathbf{b}}) \\
&= \frac{f(\hat{\mathbf{b}}|d_i = 1)\pi}{f(\hat{\mathbf{b}}|d_i = 1)\pi^{(m)} + f(\hat{\mathbf{b}}|d_i = 0)(1 - \pi^{(m)})} \\
&= \frac{f(\hat{\mathbf{b}}_{-i}|\hat{\mathbf{b}}_i, d_i = 1) f(\hat{\mathbf{b}}_i|d_i = 1)\pi^{(m)}}{f(\hat{\mathbf{b}}_{-i}|\hat{\mathbf{b}}_i, d_i = 1) f(\hat{\mathbf{b}}_i|d_i = 1)\pi^{(m)} + f(\hat{\mathbf{b}}_{-i}|\hat{\mathbf{b}}_i, d_i = 0) f(\hat{\mathbf{b}}_i|d_i = 0)(1 - \pi^{(m)})} \\
&= \frac{f(\hat{\mathbf{b}}_i|d_i = 1)\pi^{(m)}}{f(\hat{\mathbf{b}}_i|d_i = 1)\pi^{(m)} + f(\hat{\mathbf{b}}_i|d_i = 0)(1 - \pi^{(m)})} \qquad (17) \\
&= \frac{\prod_{l=1}^{L} f(\hat{\mathbf{b}}_{il}|d_i = 1)\pi^{(m)}}{\prod_{l=1}^{L} f(\hat{\mathbf{b}}_{il}|d_i = 1)\pi^{(m)} + \prod_{l=1}^{L} f(\hat{\mathbf{b}}_{il}|d_i = 0)(1 - \pi^{(m)})}, \qquad (18)
\end{aligned}$$

where $\hat{\mathbf{b}}_{-i}$ contains all the estimated coefficients except those for the ith gene $\hat{\mathbf{b}}_i$. Note that the density function f of the gene expression values depends on variance estimates $\sigma^{2(m)}_{w,l}$ from last iteration. Equation (17) holds because the dependence of $\mathbf{b}_{-i}$ on $\mathbf{b}_i$ is only through the replicate effects $\mathbf{v}_j$ such that $f^{(m)}(\hat{\mathbf{b}}_{-i}|\hat{\mathbf{b}}_i, d_i) = f^{(m)}(\hat{\mathbf{b}}_{-i}|\hat{\mathbf{b}}_i)$. Denote by Σ_{il1} the covariance matrix for $(\hat{\mathbf{b}}_{il}|d_i = 1)$ and by Σ_{il0} the covariance matrix for $(\hat{\mathbf{b}}_{il}|d_i = 0)$. Then we have $f(\hat{\mathbf{b}}_{il}|d_i = 1) = (\frac{1}{\sqrt{2\pi|\Sigma_{il1}|}})^J \exp\{-\frac{1}{2}\hat{\mathbf{b}}_{il}\Sigma_{il1}^{-1}\hat{\mathbf{b}}_{il}\}$ and $f(\hat{\mathbf{b}}_{il}|d_i = 0) = (\frac{1}{\sqrt{2\pi|\Sigma_{il0}|}})^J \exp\{-\frac{1}{2}\hat{\mathbf{b}}_{il}\Sigma_{il0}^{-1}\hat{\mathbf{b}}_{il}\}$. At the $(m+1)$-th M-step, $\hat{\pi}^{(m+1)}$ can be estimated as $\sum_{i=1}^{n} \hat{d}_i^{(m+1)}/n$ and $\hat{\sigma}^{2(m+1)}_{w,l} = \widehat{\pi\sigma^2_{w,l}}/\hat{\pi}^{(m+1)}$. The iteration converges if $\|\sigma^{2(m)} - \sigma^{2(m-1)}\|_2^2 < 1e - 07$.

A.4 Bootstrap Variance Estimates of Parameters

Denote all the parameters by $\theta = \{\pi, \sigma^2_\epsilon, \sigma^2_{u,l}, \sigma^2_{w,l}, \sigma^2_{v,l}, \sigma^2_{e,l}, l = 1, \dots, L\}$. We use bootstrap method on estimated principal component scores to estimate the variance of θ. Draw a resample

$$\hat{\mathbf{b}}^* = \big(\hat{\mathbf{b}}^*_1, \dots, \hat{\mathbf{b}}^*_n\big),$$

by resampling with replacement from $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_n)$. Repeat the process for B times. For each resample $\hat{\mathbf{b}}^*_m$, we can have an estimate $\hat{\theta}^*_m$ for θ. Thus, the variance estimate for $\hat{\theta}$ is

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{B} \sum_{m=1}^{B} \big(\hat{\theta}^*_m - \hat{\theta}^*\big)^2,$$

where $\hat{\theta}^*$ is the average of the $\hat{\theta}^*_m$'s.

References

1. Ash RB, Gardner MF (1975) Topics in stochastic processes. Academic Press, New York
2. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci USA 100(18):10146–10151
3. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I (2003) Continuous representations of time-series gene expression data. J Comput Biol 10(3–4):341–356
4. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—A practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300
5. Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32:490–495
6. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96(456):1151–1160
7. Hong F, Li H (2006) Functional hierarchical models for identifying genes with different time-course expression profiles. Biometrics 62(2):534–544
8. Joyce E, Popper S, Falkow S (2009) Streptococcus pneumoniae nasopharyngeal colonization induces type i interferons and interferon-induced gene expression. BMC Genomics 10:404
9. Leek JT, Monsen E, Dabney AR, Storey JD (2006) Edge: Extraction and analysis of differential gene expression. Bioinformatics 22:507–508
10. Leng X, Müller HG (2006) Classification using functional data analysis for temporal gene expression data. Bioinformatics 22:68–76
11. Liu X, Müller HG (2003) Modes and clustering for time-warped gene expression profile data. Bioinformatics 19:1937–1944
12. Liu X, Yang M (2009) Identifying temporally differentially expressed genes through functional principal component analysis. Biostatistics 10:667–679
13. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5(2):155–176
14. Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, Lee YS (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 19(6):694–703
15. Ramsay J, Silverman B (1997) Functional data analysis. Springer, New York
16. Rice J, Wu C (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57:253–259
17. Saaf A, Halbleib J, Chen X, Yuen S, Leung S, Nelson W, Brown P (2007) Parallels between global transcriptional programs of polarizing caco-2 intestinal epithelial cells in vitro and gene expression programs in normal colon and colon cancer. Mol Biol Cell 18:4245–4260
18. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100(16):9440–9445
19. Storey JD, Xiao WZ, Leek JT, Tompkins RG, Davis RW (2005) Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA 102(36):12837–12842
20. Tai Y, Speed T (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. Ann Stat 34(5):2387–2412
21. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98(9):5116–5121
22. Wang J, Kim SK (2003) Global analysis of dauer gene expression in Caenorhabditis elegans. Development 130(8):1621–1634
23. Xu XL, Olson JM, Zhao LP (2002) A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. Hum Mol Genet 11(17):1977–1985
24. Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100(470):577–590
25. Yuan M, Kendziorski C (2006) Hidden Markov models for microarray time course data in multiple biological conditions. J Am Stat Assoc 101:1323–1332