

# Identifying Differentially Expressed Genes in Time Course Microarray Data

Ping Ma · Wenxuan Zhong · Jun S. Liu

Received: 12 August 2009 / Accepted: 28 September 2009 / Published online: 10 October 2009  
© International Chinese Statistical Association 2009

**Abstract** Identifying differentially expressed (DE) genes across conditions or treatments is a typical problem in microarray experiments. In time course microarray experiments (under two or more conditions/treatments), it is sometimes of interest to identify two classes of DE genes: those with no time-condition interactions (called parallel DE genes, or PDE), and those with time-condition interactions (nonparallel DE genes, NPDE). Although many methods have been proposed for identifying DE genes in time course experiments, methods for discerning NPDE genes from the general DE genes are still lacking. We propose a functional ANOVA mixed-effect model to model time course gene expression observations. The fixed effect of (the mean curve) of the model decomposes bivariate functions of time and treatments (or experimental conditions) as in the classic ANOVA method and provides the associated notions of main effects and interactions. Random effects capture time-dependent correlation structures. In this model, identifying NPDE genes is equivalent to testing the significance of the time-condition interaction, for which an approximate  $F$ -test is suggested. We examined the performance of the proposed method on simulated datasets in comparison with some existing methods, and applied the method to a study of human reaction to the endotoxin stimulation, as well as to a cell cycle expression data set.

**Keywords** Time course microarray · Differentially expressed gene · Functional data analysis · Mixed-effect model · Smoothing spline · ANOVA

---

P. Ma (✉) · W. Zhong  
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA  
e-mail: [pingma@illinois.edu](mailto:pingma@illinois.edu)

W. Zhong  
e-mail: [wenxuan@illinois.edu](mailto:wenxuan@illinois.edu)

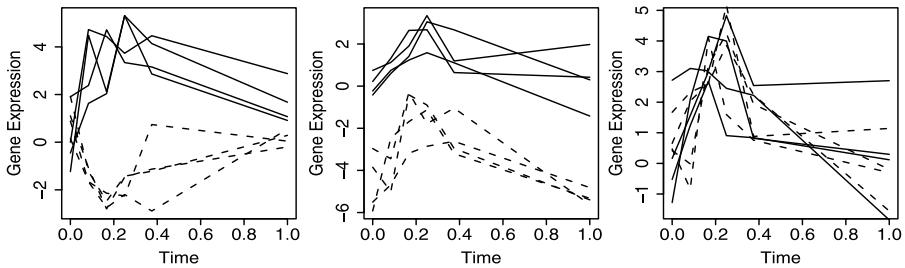
J.S. Liu (✉)  
Department of Statistics, Harvard University, Cambridge, MA 02138, USA  
e-mail: [jliu@stat.harvard.edu](mailto:jliu@stat.harvard.edu)

## 1 Introduction

To study the dynamics of genome-wide mRNA expression levels during a biological process, researchers often conduct time course microarray experiments during the process. Genes with different “expression profiles” across different conditions or treatments are called differentially expressed (DE) genes. Identifying differentially expressed genes in time course microarray experiments has some interesting challenges. For example, the need to account for intrinsic time dependence of gene expression measurements of the same gene at different time points renders classical hypothesis testing methods used in static microarray data (array taken irrespective of time), such as SAM [30], invalid. Moreover, it is often of biological interest to classify and detect different types of DE genes in such settings.

It is possible to cast the gene expression observations as multivariate data with a certain correlation structure, but in this case the same time points across conditions since multivariate approach ignores the time interval and time order of the sampling [29]. As evidenced in our analysis of mutant vs wild-type cell cycle data [24], time interval is key to properly aligning the mutant and wild-type data and to interpreting the results. Alternatively, hidden Markov models have been employed to model time course gene expressions in Yuan and Kendziorski [32]. Unfortunately, such models require the Markov property, which is unlikely to hold for most time course microarray data. Recently, modeling time course microarray data in the form of curves, i.e., functional data, is subject to active research. To identify time course DE genes, Storey et al. [28] introduce a curve-based method, implemented in EDGE, in which the mean gene expressions under two conditions were modeled as two curves, and the two curves were compared using a likelihood ratio test. A refined functional data approach was developed in Hong and Li [14]. In these two curve-based methods, smoothing parameters, e.g., knots and effective degrees of freedom, are the same across all genes and must be specified a priori. Therefore, it would be difficult for the methods to handle drastically different expression patterns among different genes, which could lead to high false discovery rate. An improvement was developed in Ma et al. [21].

An important limitation of aforementioned methods is that they are direct generalizations of the methods designed to analyze time course microarray data in a single condition. Issues such as the condition-time interaction that are unique to time course array taken at two or more conditions have not been well studied. With emergence of multi-factor time course microarray, e.g., factorial designed time course microarray experiment [18], interactions among the factors and/or between factors and time become essential for understanding the source of variation of gene expressions and navigating new biological discovery. However, aforementioned methods identify DE genes including a significant number of DE genes with parallel expression profiles under different conditions (henceforth, PDE genes), i.e., insignificant condition-time interaction, as illustrated in the center panel in Fig. 1. Compared with PDE genes, in many scientific investigations, DE genes with nonparallel expression profiles over time (henceforth, NPDE genes; Center panel in Fig. 1), i.e., significant condition-time interaction, are of primary interest. Focused study of the NPDE genes may provide more information on how cell responds differently to different stimulus or treatments.



**Fig. 1** Simulated longitudinal expression profiles at five time points with four replicates under two biological conditions (*solid and dashed lines*, respectively). *Left*: a DE gene with nonparallel expression profiles; *Center*: a DE gene with parallel expression profiles; *Right*: a non-DE gene

For example, in the cell cycle study reported by Orlando et al. [24], the primary interest is to identify cell cycle regulated genes (in wild-type cells) that drastically change expression patterns in the mutant experiment. Moreover, longitudinal time course microarray experiments are commonly used in case-control study and in clinical trials. In such experiments, mRNA samples are taken from a small number of subjects over time in the treatment group and from another small number of subjects in the control group. Because each group only consists of a small number of subjects, one subject with high baseline gene expression can cause a high average gene expression of the whole group. Thus, there are many PDE genes between groups, but they are biologically uninteresting.

To define and identify PDE and NPDE genes, we use a functional ANOVA mixed-effect model to model time course gene expressions. The general framework was introduced in Ma and Zhong [22] to cluster multifactor time course microarray data. A distinguishing feature of functional ANOVA analysis is that it decomposes bivariate functions of time and treatments (experiment conditions) similarly as in the classical ANOVA analysis and provides the associated notions of main effects and interactions, see Wahba [31] and Gu [8] for an introduction. Moreover, the nonparametric function has a great flexibility in modeling expression taken at a variety of sampling time points, and the parametric random effects are used to capture time dependent correlation structures. In this general framework, identifying NPDE genes is equivalent to testing whether the interaction term is significant in the functional ANOVA mixed-effect model. We develop a new test for such a purpose and examine the performance of the proposed method on simulated datasets and on the data obtained in a study of human reaction to the endotoxin stimulation, as well as on a cell cycle expression data.

The remainder of the article is organized as follows. In Sect. 2, we present a functional ANOVA mixed-effect model representation for time course gene expression data. A statistical test is considered in Sect. 3 and tested by simulations in Sect. 4. The applications of the method to a study of human reaction to the endotoxin stimulation and to a cell cycle expression data are presented in Sect. 5. A few remarks in Sect. 6 conclude the article. The details of estimation of functional ANOVA mixed-effect model is given in the [Appendix](#).

## 2 Functional ANOVA Mixed-effect Model for Modeling Time Course Gene Expressions

In this section, we describe the functional ANOVA mixed-effect models for modeling longitudinal replicated time course expression data developed in Ma and Zhong [22]. The cross-sectional replicated time course gene expression data can be analyzed using functional ANOVA models without random effects and is illustrated in the real data analysis.

### 2.1 The Model

The expression profile of a gene at time  $t_{ki}$  of subject  $k$  in condition (treatment group)  $g(k)$  can be modeled with a nonparametric mixed-effect model as

$$Y_{ki} = \eta(t_{ki}, g) + \mathbf{z}_k^T \mathbf{b}_k + \epsilon_{ki}, \quad (1)$$

where  $i = 1, \dots, n_k, k = 1, \dots, K, g = 1, \dots, G$  (for convenience of description,  $k$  is suppressed from  $g$ ), population mean time course profile is described by the bivariate function  $\eta$ , which is assumed to be a smooth function of time  $t$  for each group  $g$ , the smoothness of  $\eta$  is quantified through a quadratic functional  $J$  (the details are given in the [Appendix](#)),  $\mathbf{b}_k$  are the subject specific random intercepts to model intra-subject variation with  $\mathbf{b}_k \sim N(\mathbf{0}, B)$ ,  $\mathbf{z}_k$  are the corresponding design vectors for random effect, and random errors  $\epsilon_{ki} \sim N(0, \sigma^2)$  are independent of  $\mathbf{b}_k$  and of each other. To be applicable in a wide range of settings, we allow that  $\eta$  has isolated discontinuities, i.e., jump points. Thus  $\eta$  is only required to be piecewise smooth. The random effect covariance matrix  $B$  and random error variance  $\sigma^2$  are to be estimated from the data. By using different specifications of  $\mathbf{b}$  and associated design vector  $\mathbf{z}$ , model (1) can accommodate various correlation structures. A simple example is to set  $\mathbf{b}_k = b_k$  and  $\mathbf{z}_k^T \mathbf{b}_k = b_k$ ; we have  $\mathbf{B} = \sigma_b^2$  and the same correlation across time.

In model (1), the bivariate function  $\eta$  may be further decomposed as

$$\eta(t, g) = \eta_{\emptyset} + \eta_1(t) + \eta_2(g) + \eta_{1,2}(t, g), \quad (2)$$

where  $\eta_{\emptyset}$  is the overall mean,  $\eta_1(t)$  is the time effect at time  $t$ ,  $\eta_2(g)$  is the treatment effect of  $g$ th group, and  $\eta_{1,2}(t, g)$  is the effect of the interaction between time and treatment. Both time effect and treatment effect are defined as deviation from the overall mean, so  $\int_0^T \eta_1(t) dt = 0$  and  $\sum_{g=1}^G \eta_2(g) = 0$ . Similarly, the time-treatment interaction are defined as  $\int_0^T \eta_{1,2}(t, g) dt = 0$  for all  $g$ , and  $\sum_{g=1}^G \eta_{1,2}(t, g) = 0$  for all  $t$ . Such decomposition extends the classical ANOVA decomposition on discrete domains to generic domains and is referred to as functional ANOVA decomposition [8, 31]. When the time-treatment interaction term  $\eta_{1,2}(t, g)$  is significant, we have different trajectory for population mean time course profiles in different treatment groups, i.e.,  $\eta(t, g_1) - \eta(t, g_2) = \eta_2(g_1) - \eta_2(g_2) + \eta_{1,2}(t, g_1) - \eta_{1,2}(t, g_2)$  for every  $t$ , where the first two terms in the right-hand side of the equation are constants, and the last two terms the right-hand side of the equation vary with  $t$ .

If the treatment-time interaction  $\eta_{1,2}(t, g)$  is not significant in functional ANOVA model (2), one may adequately fit the additive model

$$\eta(t, g) = \eta_{\emptyset} + \eta_1(t) + \eta_2(g), \tag{3}$$

which yields parallel population mean time course profiles in different treatment groups, i.e.,  $\eta(t, g_1) - \eta(t, g_2) = \eta_2(g_1) - \eta_2(g_2)$  for every  $t$ , where the right-hand side of the equation is a constant.

To compare the expression profiles, we refer to the genes with significant time-treatment interaction term in (2), i.e.,  $\eta_{1,2}(t, g) \neq 0$ , as NPDE genes; the genes with significant main effect in treatment  $g$  but no time-treatment interaction in (3), i.e.,  $\eta_2(g) \neq 0$  and  $\eta_{1,2}(t, g) = 0$ , is referred to as PDE genes. The methods to distinguish NPDE genes from PDE genes are still lacking.

### 2.2 Estimation

Model (1) is estimated using the penalized Henderson’s likelihood [10, 25] through minimizing

$$\sum_{k=1}^K \sum_{i=1}^{n_k} (Y_{ki} - \eta(t_{ki}, g) - \mathbf{z}_k^T \mathbf{b})^2 + \sum_{k=1}^K \sigma^2 \mathbf{b}_k^T B^{-1} \mathbf{b}_k + N\lambda J(\eta), \tag{4}$$

where  $N = \sum_{k=1}^K n_k$ , the quadratic functional  $J(\eta)$  quantifies the roughness of  $\eta$ , and the smoothing parameter  $\lambda$  controls the trade-off between the goodness-of-fit and the smoothness of  $\eta$ .

Piecewise smooth  $\eta$  can be estimated using (4) through replacing  $\eta$  by  $\tilde{\eta} = \sum_{l=1}^q \beta_l \psi_l + \eta$  in the first term in (4), where the  $\psi_l$  are basis functions representing jump points, and  $\beta_l$  are unknown parameters.

Let  $\mathbf{Y} = (Y_{11}, \dots, Y_{Kn_K})^T$ ,  $Z$  be the matrix stacked by  $\mathbf{z}_i^T$ , and  $\Omega = \sigma^2 B^{-1}$ . The fitted values  $\hat{\mathbf{Y}} = \hat{\boldsymbol{\eta}} + Z\hat{\mathbf{b}}$  of (4) can be written as  $\hat{\mathbf{Y}} = A(\lambda, \Omega)\mathbf{Y}$ , where  $A(\lambda, \Omega)$  is a so-called smoothing matrix.

Treating the correlation parameters  $\Omega$  as extra smoothing parameters, we adopt the approach of Gu and Ma [10] to estimate  $\lambda$  and the correlation parameters  $\Omega$  simultaneously through generalized cross-validation (GCV). The details of the estimation can be found in the [Appendix](#).

### 3 Significance Testing

To identify NPDE genes, we can test

$$H_0 : \eta_{1,2}(t, g) = 0 \quad \text{against} \quad H_1 : \eta_{1,2}(t, g) \neq 0 \tag{5}$$

in the nonparametric model (1) with functional ANOVA (2). In classic ANOVA and linear regression model, the likelihood ratio tests are popular tools for the purpose. In linear mixed-effect models, Self and Liang [26] show that asymptotic sampling

distributions of likelihood ratio statistics are mixture chi-square distributions under standard conditions. Since penalized least squares estimation of smoothing spline model is equivalent to fitting a linear mixed-effect model, Guo [11] developed a likelihood ratio test in nonparametric settings and showed that the asymptotic sampling distribution of this likelihood ratio statistic is still a mixture chi-squared distribution. However, it was pointed out in Crainiceanu and Ruppert [4] that such an approximation does not work very well in finite sample and bootstrap may be used to better approximate the null distribution. Some diagnostic tool was developed in Gu [9] for smoothing spline ANOVA model, whereas test statistics were not available.

Define

$$S_R = \mathbf{Y}^T (I - A_R(\lambda_R, \Omega_R))^2 \mathbf{Y}$$

and

$$S_F = \mathbf{Y}^T (I - A_F(\lambda_F, \Omega_F))^2 \mathbf{Y},$$

where  $A_R(\lambda_R, \Omega_R)$  is the smoothing matrix for estimating (1) with additive ANOVA decomposition (3), and smoothing parameters  $\lambda_R$  and  $\Omega_R$  are chosen by GCV.  $A_F(\lambda_F, \Omega_F)$  is the smoothing matrix for estimating (1) with full ANOVA decomposition (2), and smoothing parameters  $\lambda_F$  and  $\Omega_F$  are chosen by GCV.  $S_R$  and  $S_F$  are the corresponding residual sums of squares in two ANOVA decompositions. They are quadratic forms in normal variable  $\mathbf{Y}$ . When smoothing parameters  $\lambda$  and  $\Omega$  are fixed, the sampling distribution of  $S_R$  and  $S_F$  are linear combinations of noncentral  $\chi^2$  variables under the null hypothesis. The methods and algorithms are available to calculate the relevant probabilities of linear combinations of  $\chi^2$  variables [6].

Since we can approximate a linear combination of  $\chi^2$  variables by a  $\chi^2$  distribution adequately [2], we consider a generalization of the standard  $F$  statistic

$$F = \frac{(S_R - S_F)/(p_F - p_R)}{S_F/(N - p_F)}, \quad (6)$$

where  $p_R$  and  $p_F$  quantify the “number of parameters” in the additive model and full model, respectively. Since the numerator and denominator in (6) are correlated, the sampling distribution of  $F$  does not belong to any known distribution family. To make it practical, we approximate the sampling distribution of (6) by the  $F$  distribution with degrees of freedom of  $p_F - p_R$  and  $N - p_F$ , where  $p_R = [\text{tr}(A_R(\lambda_R, \Omega_R))]$  and  $p_F = [\text{tr}(A_F(\lambda_F, \Omega_F))]$ , and  $[x]$  denotes the rounding integer of  $x$ . The trace of the smoothing matrix is also referred to as effective degrees of freedom, which provides an intuitive analog to the degrees of freedom in parametric models. The approximation we provide here was developed by Hastie and Tibshirani [12] in simple smoothing spline regression. Further careful calibrations of degrees of freedom can be found in Zhang [33] and Liu and Wang [20], which become impractical due to their high computational cost. Furthermore, they fixed the smoothing parameters to derive the calibrations in simple smoothing spline model, which makes the calibrations less appealing in our setting.

Given that the genes are not significantly NPDE, we may further investigate whether they are significant PDE genes. That is, in model (1) with functional ANOVA

(3), we are also interested in testing

$$H_0 : \eta_2(g) = 0; \quad H_1 : \eta_2(g) \neq 0. \quad (7)$$

We can easily generalize test (6) with proper degrees of freedom for that purpose.

For comparison, we can also model the data simply using a two-way ANOVA model, i.e., disregarding the time dependence at all and only testing for time-treatment interactions. That is, we consider the simple linear two-factor model  $Y_{ki} = \mu + \alpha_i + \beta_g + (\alpha\beta)_{ig} + \epsilon_{ki}$  with  $\epsilon_{ki} \sim N(0, \sigma^2)$  and test for the significance of the interaction term. This gives rise to the classic  $F$ -test of the form,

$$F = \frac{SS(\text{time} \times \text{condition}) / (n-1)(G-1)}{SSE / (N-nG)}, \quad (8)$$

where we assume that  $n_k = n$ ,  $SS(\text{time} \times \text{condition})$  is the sum of squares of the interaction, and  $SSE$  is the sum of squares of the error.

#### 4 Simulation Study

To assess the performance of the proposed method, we carried out extensive analysis on simulated datasets. We report some of them in this section.

In the first setting, one hundred NPDE time course gene expressions are generated from the following model:

$$Y_{ki} = 2.5 \sin(3\pi t_i)(1 - t_i) + \delta I_{[t_i > 0.5]} I_{[g=1]} + b_k + \epsilon_{ki}, \quad (9)$$

where  $t_i = (i - 1)/14$ ,  $i = 1, \dots, 15$ , and  $k = 1, \dots, 6$ ,  $g = 0, 1$ , the subject random effect  $b_k$  has standard normal distribution, and the random error  $\epsilon_{ki}$  is also standard normal distributed. We set  $\delta = 0, 0.5, 1, 1.5, 2$ .

In the second setting, we generated time course gene expressions of two hundred genes. We first generated one hundred PDE time course gene expressions from the following model:

$$Y_{ki} = 2.5 \sin(3\pi t_i)(1 - t_i) - 4 I_{[g=1]} + b_k + \epsilon_{ki}. \quad (10)$$

We then generated one hundred nondifferentially expressed time course gene expression from

$$Y_{ki} = 2.5 \sin(3\pi t_i)(1 - t_i) + b_k + \epsilon_{ki}. \quad (11)$$

We repeated each setting 100 times and applied the proposed test (6) to the each dataset and identified the NPDE genes at the significance level 0.05.

In the third setting, one hundred NPDE time course gene expressions are generated from the following model:

$$Y_{ki} = (1 + 2\delta I_{[g=1]})t_i + b_k + \epsilon_{ki}, \quad (12)$$

where  $t_i = (i - 1)/14$ ,  $i = 1, \dots, 15$ , and  $k = 1, \dots, 6$ ,  $g = 0, 1$ , the subject random effect  $b_k$  has standard normal distribution, and the random error  $\epsilon_{ki}$  is also standard

**Table 1** Comparison of the proposed method, EDGE, and ANOVA in identifying NPDE genes in simulated datasets. The power is averaged over 100 independent and identically simulated datasets. Note: the standard deviation of each result can be calculated using a binomial distribution where the mean is the proportion of rejections

Method	First Setting					Second Setting	Third Setting	
	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$		$\delta = 0.5$	$\delta = 1$
Proposed	0.046	0.192	0.342	0.653	0.944	0.046	0.064	0.234
EDGE	0.049	0.191	0.304	0.612	0.904	0.343	0.079	0.299
ANOVA	0.042	0.084	0.213	0.494	0.801	0.042	0.095	0.322

normal distributed. We set  $\delta = 0.5, 1$ . The power of the proposed  $F$ -test for all genes were calculated. For comparison, we also applied EDGE (Storey et al. [28]) and the standard two-way ANOVA  $F$ -test (8) to the same data and calculated power. The result is summarized in Table 1.

Through the range of  $\delta > 0$  in the first setting, the powers of the proposed method are consistently higher than those of EDGE and ANOVA. For  $\delta = 0$ , the type-I error rate of the proposed method is very close to 0.05, which indicates that the  $F$ -distribution approximation works well for the proposed test statistics. Since the two-way ANOVA did not take into account the correlation presented in the data, its type-I error rate was lower than its nominal level when  $\delta = 0$ . In the second setting, both the proposed method and ANOVA have consistent performances as in the first setting, whereas the type-I error rate of EDGE is much higher since it cannot distinguish parallel and nonparallel differentially expressed genes. In the third setting, the data was generated from a linear ANOVA model. The power of the proposed method is slightly smaller than that of ANOVA and EDGE.

In all three settings, we also noticed that the numbers of NPDE genes identified by EDGE were different in different runs of the EDGE software even for the same dataset. The fluctuation is caused by the fact that the null distribution of EDGE's statistics is calculated through a resampling method.

## 5 Real Data Analysis

### 5.1 Human Reaction to Endotoxin Stimulation Gene Expression

To study the genome-wide response to acute inflammation, a time course gene expression experiment was reported in Calvano et al. [1]. Among eight human subjects involved in the study, four randomly selected subjects were treated with LipoPolySacharide (LPS) endotoxin, which induces inflammation, and the remaining four were assigned placebo. Blood samples were obtained from all subjects in the study prior to the treatment, and at 2, 4, 6, 9, and 24 hours after the LPS or placebo infusion was administered. From the collected blood samples RNA was extracted and hybridized to Affymetrix HG-U133A microarrays. For one of the control subjects, two samples (namely, those collected at 4 and 6 hours) were missing. The data was



processed using dChip software Li and Wang [19]. For one of the LPS treated subjects, two of the microarrays were detected as outliers (namely, those corresponding to 2 and 24 hours) and were removed from further analysis.

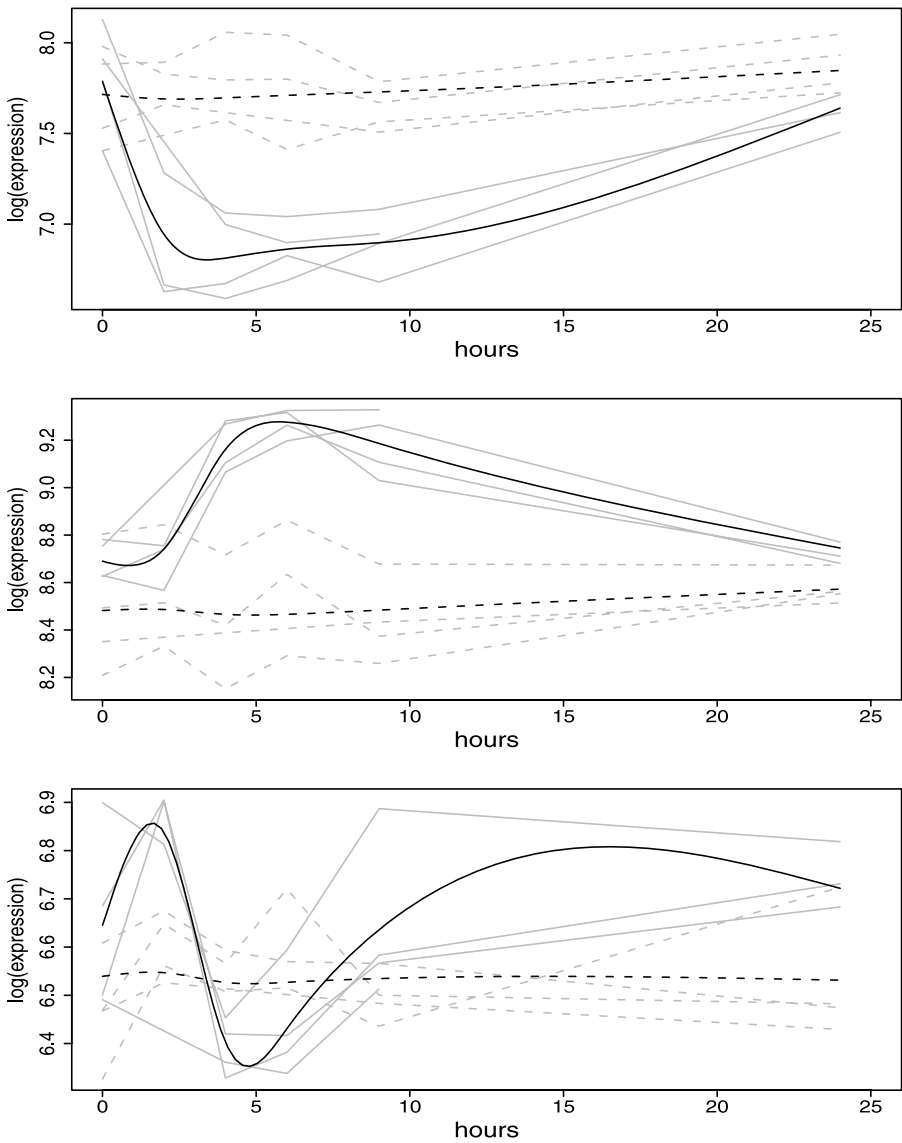
Using the criteria that the coefficient of variation has to be in (0.1, 1000) for pre-screening, we kept 15642 genes for further analysis. We applied our method using model (1) with penalty (15) and identified 1875 NPDE genes at false discovery rate corrected significance level 0.05, i.e.,  $q$ -value [27] less than 0.05. Medical Subject Headings (MeSH) analysis and pathway analysis were performed to test their biological significance.

MeSH is the National Library Medicine's controlled vocabulary for indexing articles in the PubMed database. Using Hypergeometric test, we test the significance of the hypothesis that a particular MeSH term occurs in the nonparallel differentially expressed gene set by chance [3]. At the significance level 0.05, the significant MeSH terms are "Ribosomal Proteins", "Cell Nucleolus", "Regulatory Sequences", "Ribonucleic Acid", "RNA", "Viral", "Proteome", "Proteomics", "Ribosomes", "RNA", "Messenger", "Evolution", "RNA-Binding Proteins", "Proteins", "RNA", "Small Nuclear", "Cell Nucleus". These significant MeSH terms suggest that many biological processes are enriched in the NPDE gene set.

Interestingly, nine pathways are statistically significantly over-represented at level 0.05 in the selected gene sets when identified using Pathway-Express [15]. They are: Amyotrophic lateral sclerosis (ALS,  $P$ -value = 0.0122), Antigen processing and presentation ( $P$ -value = 0.0165), mammalian TOR (mTOR,  $P$ -value = 0.0229), TGF-beta signaling pathway ( $P$ -value = 0.0468), Axon guidance ( $P$ -value = 0.0090), neurodegenerative Disorders ( $P$ -value = 0.0283), Dentatorubropallidolusian atrophy ( $P$ -value = 0.0365), long-term potentiation ( $P$ -value = 0.04298), Huntington's disease ( $P$ -value = 0.0230). These pathways can be classified into three categories: (1) inflammation disease associated pathway: ALS is a common adult-onset inflammatory disease, which will typically lead to paralysis and death [17]; (2) neural network associated pathways: neurodegenerative disorders, Huntington's disease, and Dentatorubropallidolusian atrophy are three neuron diseases. Moreover, long-term potentiation and axon guidance are important processes for the action of a neuronal network; immune-response-associated pathways: antigen processing and presentation cell process which coordinates the immune response to inflammation. In addition, TGF-beta signaling pathway regulates many cellular processes including the immune response.

The gene expression profiles and cubic spline fits of three represented genes, SOD1, TAPBP, and CAMK2G, are given in Fig. 2. For all three genes, the expressions in the placebo group did not change significantly. SOD1 is the well-known ALS-associated gene, which modulates the ALS progression [17]. The estimated expression of SOD1 in LPS treated group drops immediately after time zero and increases slowly after 4 hours. Since the primary function of SOD1 is to detoxify superoxide, making oxygen and hydrogen peroxide, we infer from the estimated expression that these functions were weakened at the beginning and went back to normal function gradually after 4 hours.

The expression of TAPBP (TAP binding protein) in the LPS treatment group increases after 2 hours of treatment and drops slowly after 6 hours. This suggests that



**Fig. 2** Cubic spline fits of log-transformed gene expressions. From *top to bottom*: the SOD1 (ALS pathway), TAPBP (Antigen processing and presentation), and CAMK2G (Long-term potentiation). The fitted  $\eta(t, g)$ , with  $g = \text{LPS}$  or placebo, are in *solid* and *dashed lines*, respectively, with the corresponding data superimposed as *faded solid* and *faded dashed lines*

TAPBP encodes more transmembrane glycoprotein to mediate interaction between newly assembled molecules and the transporter associated with antigen processing, which is essential for the transport of antigenic peptides across the endoplasmic reticulum membrane [23].

It is very interesting to examine the gene expression changes of CAMK2G, which is a calcium/calmodulin-dependent protein kinase (CaM kinase) gene in the long-term potentiation pathway. The expression of CAMK2G in the LPS treatment group had a sharp peak at 4 hours, dropped significantly after 4 hours, and then increased slowly afterwards. Once treated, we hypothesize that it is necessary for CaMK2G to be quickly activated so as to induce long-term potentiation. After that, CaMK2G was suppressed and returned to its normal state [23].

## 5.2 Wild-type and Cyclin-mutant Cell Cycle Gene Expression

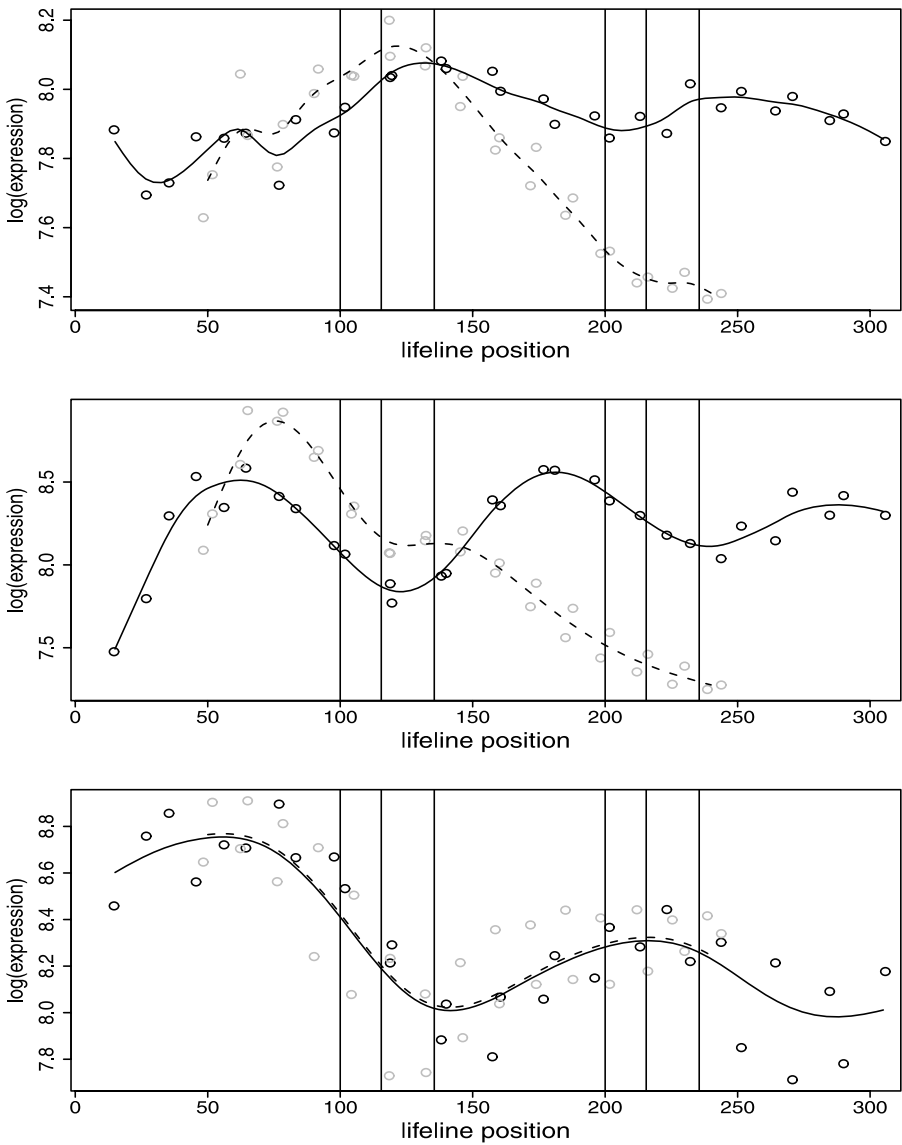
It is well known that the activity of cyclin-dependent kinases (CDKs) have a significant impact on the cell cycle program in yeast. To determine the extent to which the cell cycle is controlled by CDKs, Orlando et al. [24] conducted a time course microarray at 15 equally spaced time points with two replicates in synchronized populations of both wild-type cells and cyclin-mutant cells using Affymetrix Yeast 2.0 Arrays. The 15 time points cover roughly two cell cycles in wild-type cells, and one and a half cell cycle in cyclin-mutant cells. In addition, the synchronization times were different for four arrays at each time point. Thus Orlando et al. [24] re-aligned arrays to the so-called “lifeline position” to make arrays comparable. The resulting lifeline position correspond to 60 distinct nonequal interval positions. Due to this special data structure, it is very difficult to employ the multivariate Gaussian method [29] and hidden Markov model [32] for data analysis.

Orlando et al. [24] selected 1275 genes, which were reported to be cell cycle associated in the literature. We focus on these 1275 genes in our study. Since the experiment is a cross-sectional replicated experiment, we applied our method using model (1) with random effect design vector  $\mathbf{z}$  setting at 0. Using penalty (15), we identified 879 pattern-differentially expressed genes at significance level 0.05, i.e.,  $p$ -value less than 0.05. Among the 879 genes, 549 genes were also identified as NPDE genes by Orlando et al. [24].

In Fig. 3, the estimated gene expressions for RAPI, FAA3, and PLB1 are plotted. RAPI is a cell-cycle regulated gene [13]. We can see that the estimated mean expression in cyclin-mutant cells has a peak in first S phase and drops down continuously till the 2nd G2/M phase, which is significant from that in wild-type cells ( $P$ -value =  $1.67 \times 10^{-10}$ ). However, Orlando et al. [24] did not find it differentially expressed.

The estimated expression of FAA3 in wild-type cells exhibits two strong periodic patterns from recovery phase to second S phase and shows a mild half periodic pattern in the second G2/M phase. In contrast, the estimated expression in cyclin-mutant cells does not have obvious second period pattern ( $P$ -value =  $5.44 \times 10^{-21}$ ). Orlando et al. [24] also find FAA3 expressions differentially expressed in two populations.

The estimated expressions of PLB1 have almost the same pattern in both wild-type and cyclin-mutant cells. Our test suggests that the expressions are not significantly nonparallel differentially expressed ( $P$ -value = 0.3868). However, Orlando et al. [24] found PLB1 expressions differentially expressed in two populations.



**Fig. 3** Cubic spline fits of log-transformed gene expressions. From *top to bottom*: the RAP1, FAA3, and PLB1. The fitted  $\eta(t, g)$ , with  $g =$  wild-type or cyclin-mutant, are in *solid and dashed lines*, respectively, with the corresponding data superimposed as *solid and faded circles*. The cell cycle phases separated by *vertical lines* are, from *left to right*, recovery from synchrony, 1st G1, 1st S, 1st G2/M, 2nd G1, 2nd S, 2nd G2/M

### 6 Discussion

In this article, we propose a statistical method for identifying DE genes in time course microarray experiment. Functional ANOVA mixed-effect models are employed to

model the time course gene expressions. We develop a test for testing the pattern differentially expressed gene. The simulation analysis suggest that the proposed method performs very well. Although it was motivated for identifying the differentially expressed genes in time course expression data, our proposed method has a wide spectrum of applications, including, for example, comparison of growth curves. The calculations reported in this article were performed in R. Open-source code is available in the R package MFDA.

**Acknowledgements** The authors thank Tanya Logvinenko for providing the LPS microarray data and assisting the data analysis in the early stage of this research. Ping Ma’s research was supported by NSF DMS-0800631. Wenxuan Zhong’s research was supported through the NIH Genes, Environment and Health Initiative through award U01ES016011. Jun S. Liu’s research was supported by NIH R01-GM078990 and NSF DMS-0706989.

### Appendix

Model (1) is estimated using the penalized Henderson’s likelihood [10, 25]. For the completeness of the description, we include the details of the estimation here.

Model (1) is estimated using the penalized Henderson’s likelihood via minimizing

$$\sum_{k=1}^K \sum_{i=1}^{n_k} (Y_{ki} - \eta(t_{ki}, g) - \mathbf{z}_k^T \mathbf{b})^2 + \sum_{k=1}^K \sigma^2 \mathbf{b}_k^T B^{-1} \mathbf{b}_k + N\lambda J(\eta), \tag{13}$$

where  $N = \sum_{k=1}^K n_k$ , the quadratic functional  $J(\eta)$  quantifies the roughness of  $\eta$ , and the smoothing parameter  $\lambda$  controls the trade-off between the goodness-of-fit and the smoothness of  $\eta$ .

To minimize (13), we only need to consider smooth functions in the space  $\{\eta : J(\eta) < \infty\}$  or subspace therein. As an abstract generalization of the vector spaces used extensively in multivariate analysis, Hilbert spaces inherit many nice properties of the vector spaces. However, a Hilbert space is too loose to use for functional data analysis since even the evaluation functional  $[x](f) = f(x)$ , the simplest functional one may encounter, is not guaranteed to be continuous in a general Hilbert space. An example is that in the Hilbert space of square-integrable functions defined on  $[0, 1]$ , evaluation is not even well defined. Consequently, one may focus on a constrained Hilbert space for which the evaluation functional is continuous. Such a Hilbert space is referred to as a reproducing kernel Hilbert space. For example, the space of functions with square-integrable second derivatives is a reproducing kernel Hilbert space if it is equipped with appropriate inner products, see Wahba [31] and Gu [8] for details.

The minimization of (13) is performed in a reproducing kernel Hilbert space  $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$  in which  $J(\eta)$  is a square semi norm, and the solution resides in the space  $\mathcal{N}_J \oplus \text{span}\{R_J(s_i, g_j; \cdot, \cdot), i = 1, \dots, n, j = 1, \dots, K\}$ , where  $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$  is the null space of  $J(\eta)$ ,  $R_J(\cdot, \cdot; \cdot, \cdot)$  is the so-called reproducing kernel in  $\mathcal{H} \ominus \mathcal{N}_J$ , and  $\mathbf{s} = (s_1, \dots, s_n)$  is a distinct combination of all  $t_{ki}$  ( $k = 1, \dots, K, i =$

$1, \dots, n_k$ ). The solution has the expression

$$\eta(t, g) = \sum_{v=1}^m d_v \phi_v(t, g) + \sum_{i=1}^n \sum_{j=1}^G c_i R_J(s_i, g_j; t, g), \tag{14}$$

where  $\{\phi_v\}_{v=1}^m$  is a basis of  $\mathcal{N}_J$ .

For model (1) with functional ANOVA (2), we use the following penalty that is analogous to the quadric penalty gives rise to cubic spline estimate,

$$J(\eta) = \theta_1^{-1} \int_0^T (d^2 \eta_1 / dt^2)^2 dt + \theta_{1,2}^{-1} \int_0^T \sum_{g=1}^G (d^2 \eta_{1,2} / dt^2)^2 dt, \tag{15}$$

which has a null space  $\mathcal{N}_J$  of dimension  $m = 2G$ . The  $\theta_1$  and  $\theta_{1,2}$  are extra smoothing parameters, often suppressed in the notation, that adjust the relative penalties on the roughness of different components. See, e.g., Gu [8], Sect. 2.4. A set of  $\phi_v$  is given by

$$\{1, t, I_{[g=j]} - 1/G, (I_{[g=j]} - 1/G)t, j = 1, \dots, G - 1\},$$

and the function  $R_J$  is given by

$$R_J(s, g_1; t, g_2) = \theta_1 \int_0^T (s - u)_+(t - u)_+ du + \theta_{1,2} (I_{[g_1=g_2]} - 1/G) \int_0^T (s - u)_+(t - u)_+ du.$$

See, e.g., Gu [8], Sect. 2.4.4.

For model (1) with functional ANOVA (3), we use the penalty

$$J(\eta) = \int_0^T (d^2 \eta_1 / dt^2)^2 dt, \tag{16}$$

which has a null space  $\mathcal{N}_J$  of dimension  $m = G + 1$ . Correspondingly, we do not have  $(I_{[g=j]} - 1/G)t$  in the set of  $\phi_v$ , and

$$R_J(s, g_1; t, g_2) = \int_0^T (s - u)_+(t - u)_+ du.$$

It is interesting to note that testing the hypothesis on a component in functional ANOVA decomposition

$$H_0 : \eta_{1,2}(t, g) = 0 \quad \text{against} \quad H_1 : \eta_{1,2}(t, g) \neq 0 \tag{17}$$

is equivalent to testing the hypothesis on the corresponding smoothing parameter

$$H_0 : \theta_{1,2} = \infty \quad \text{against} \quad H_1 : \theta_{1,2}(t, g) < \infty. \tag{18}$$

Substituting (14) into (4), the numerical problem becomes the minimization of

$$(\mathbf{Y} - \mathbf{S}\mathbf{d} - \mathbf{R}\mathbf{c} - \mathbf{Z}\mathbf{b})^T(\mathbf{Y} - \mathbf{S}\mathbf{d} - \mathbf{R}\mathbf{c} - \mathbf{Z}\mathbf{b}) + \mathbf{b}^T\Omega\mathbf{b} + n\lambda\mathbf{c}^T\mathbf{Q}\mathbf{c} \quad (19)$$

with respect to  $\mathbf{d} = (d_1, \dots, d_m)^T$ ,  $\mathbf{c} = (c_1, \dots, c_n)^T$ ,  $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_K^T)^T$ , where  $\mathbf{Y} = (Y_{11}, \dots, Y_{KnK})^T$ ,  $S$  is  $N \times m$  with the entry  $\phi_v$ ,  $R$  is  $N \times nG$  with the entry  $R_J(s_i, g_j; t_{km}, gl)$ ,  $Z$  is stacked by  $\mathbf{z}_i^T$ ,  $\Omega = \sigma^2 B^{-1}$ , and  $Q$  is  $nG \times nG$  with the entry  $R_J(s_i, g_j; s_k, gl)$ . The solution of (19) satisfies the normal equation

$$\begin{pmatrix} S^T S & S^T R & S^T Z \\ R^T R & R^T R + (N\lambda)Q & R^T Z \\ Z^T S & Z^T R & Z^T Z + \Omega \end{pmatrix} \begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{Y} \\ R^T \mathbf{Y} \\ Z^T \mathbf{Y} \end{pmatrix}. \quad (20)$$

The normal equation of (20) can be solved by a pivoted Cholesky decomposition followed by backward and forward substitutions; see, e.g., [16] for details.

The fitted values  $\hat{\mathbf{Y}} = \mathbf{S}\hat{\mathbf{d}} + \mathbf{R}\hat{\mathbf{c}} + \mathbf{Z}\hat{\mathbf{b}}$  of (4) can be written as  $\hat{\mathbf{Y}} = A(\lambda, \Omega)\mathbf{Y}$ , where the smoothing matrix is

$$A(\lambda, \Omega) = (S, R, Z) \begin{pmatrix} S^T S & S^T R & S^T Z \\ R^T R & R^T R + (N\lambda)Q & R^T Z \\ Z^T S & Z^T R & Z^T Z + \Omega \end{pmatrix}^+ \begin{pmatrix} S^T \\ R^T \\ Z^T \end{pmatrix},$$

and  $C^+$  denotes the Moore–Penrose inverse of  $C$  satisfying  $CC^+C = C$ ,  $C^+CC^+ = C^+$ ,  $(CC^+)^T = CC^+$  and  $(C^+C)^T = C^+C$ .

With varying smoothing parameters  $\lambda$  (including  $\theta_1, \theta_{1,2}$ ) and correlation parameters  $\Omega$ , (20) defines an array of possible estimates, in which we need to choose a specific one in practice. A classic data-driven approach for selecting the smoothing parameter  $\lambda$  is generalized cross-validation (GCV), which was proposed in Craven and Wahba [5]. Treating the correlation parameters  $\Omega$  as extra smoothing parameters, we adopt the approach of Gu and Ma [10] to estimate  $\lambda$  and the correlation parameters  $\Omega$  simultaneously through minimizing the GCV score

$$V(\lambda, \Omega) = \frac{N^{-1}\mathbf{Y}^T(I - A(\lambda, \Omega))^2\mathbf{Y}}{\{N^{-1}\text{tr}(I - A(\lambda, \Omega))\}^2}. \quad (21)$$

Since the GCV score  $V(\lambda, \Omega)$  is nonquadratic in  $\lambda$  and  $\Omega$ , one may employ standard nonlinear optimization algorithms to minimize the GCV as a function of the tuning parameters. In particular, we used the modified Newton algorithm developed by Dennis and Schnabel [7] to find the minimizer. It was shown in Gu and Ma [10] that the minimizer of  $V(\lambda, \Omega)$  yields optimal smoothing asymptotically.

## References

1. Calvano S, Xiao W, Richards D et al (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437:1032–1037
2. Cantoni E, Hastie T (2002) Degrees-of-freedom tests for smoothing splines. *Biometrika* 89:251–263
3. Castillo-Davis C, Hartl D (2003) Genemerge: post-genomic analysis, data-mining and hypothesis. *Bioinformatics* 19:891–892

4. Crainiceanu CM, Ruppert D (2004) Restricted likelihood ratio tests in nonparametric longitudinal models. *Stat Sin* 14(3):713–729
5. Craven P, Wahba G (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer Math* 31:377–403
6. Davies RB, (1980) [Algorithm AS 155] The distribution of a linear combination of  $\chi^2$  random variables (AS R53: 84V33 pp 366–369). *Appl Stat* 29:323–333
7. Dennis JE, Schnabel RB (1996) Numerical methods for unconstrained optimization and nonlinear equations. SIAM, Philadelphia. Corrected reprint of the 1983 original
8. Gu C (2002) Smoothing spline ANOVA models. Springer, New York
9. Gu C (2004) Model diagnostics for smoothing spline ANOVA models. *Can J Stat* 32(4):347–358
10. Gu C, Ma P (2005) Optimal smoothing in nonparametric mixed-effect models. *Ann Stat* 33:1357–1379
11. Guo W (2002) Inference in smoothing spline analysis of variance. *J R Stat Soc, Ser B: Stat Methodol* 64(4):887–898
12. Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
13. Hogan C, Serpente N, Cogram P, Hosking CR, Bialucha CU, Feller SM, Braga VMM, Birchmeier W, Fujita Y (2004) Rap1 regulates the formation of e-cadherin-based cell–cell contacts. *Mol Cell Biol* 24:6690–6700
14. Hong F, Li H (2006) Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics* 62:534–544
15. Khatri P, Bhavsar P, Bawa G, Draghici S (2004) Onto-tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res* 32:W449–W456
16. Kim Y-J, Gu C (2004) Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J Roy Stat Soc Ser B* 66:337–356
17. Kunst CB (2004) Complex genetics of amyotrophic lateral sclerosis. *Am J Hum Genet* 75:933–947
18. Leung YF, Ma P, Link BA, Dowling J (2008) Factorial microarray analysis of zebrafish retina development. *Proc Natl Acad Sci* 105:12909–12914
19. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci* 98:31–36
20. Liu A, Wang Y (2004) Hypothesis testing in smoothing spline models. *J Stat Comput Simul* 74(8):581–597
21. Ma P, Castillo-Davis CI, Zhong W, Liu JS (2006) A data-driven clustering method for time course gene expression data. *Nucleic Acids Res* 34:1261–1269
22. Ma P, Zhong W (2008) Penalized clustering of large scale functional data with multiple covariates. *J Amer Stat Assoc* 103:625–636
23. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 33(Database Issue):D45–D58
24. Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JES, Iversen ES, Hartemink AJ, Haase SB (2008) Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* 453:944–947
25. Robinson GK (1991) That BLUP is a good thing: The estimation of the random effects. *Statist Sci* 6:15–51 (with discussions)
26. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 82:605–610
27. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proc Natl Acad Sci* 100:9440–9445
28. Storey JD, Xiao W, Leek JT, Tompkins R, Davis G (2005) Significance of time course microarray experiments. *Proc Natl Acad Sci* 102:12837–12842
29. Tai YC, Speed TP (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann Stat* 34:2387–2412
30. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98:5116–5121
31. Wahba G (1990) Spline models for observational data. CBMS-NSF regional conference series in applied mathematics, vol. 59. SIAM, Philadelphia
32. Yuan M, Kendziora C (2006) Hidden Markov models for microarray time course data under multiple biological conditions. *J Am Stat Assoc* 101:1323–1340
33. Zhang C (2003) Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *J Am Stat Assoc* 98(463):609–628