Statistics and Its Interface Volume 10 (2017) 643–655

Detection of threshold points for gene expressions under multiple biological conditions

Dianliang Deng*, Hong-Bin Fang, Kian Razeghi Jahromi, Jiuzhou Song, and Ming Tan

Temporal gene expression data is of importance in the classifications of gene functions and have been extensively used in biomedical studies, such as cancer diagnostics. However, since temporal gene expressions vary over time, after the initial time periods, many genes exhibit some kind of stability, which means that gene expressions keep constant or fluctuate slightly after those time points. Thereby, this threshold point is a key in the study of behaviours of gene expressions, which can be used to decide the measuring time period and to distinguish the gene expressions. In this paper three methods are presented to detect the threshold points for the gene expressions. In particular, the first-order and second-order change rates are used to construct the test statistics for detecting the threshold points. The simulation study shows that the proposed methods have a good performance for the detection of threshold points. A real dataset with 21 genes in *P. aeruginosa* expressed in 24 biological conditions is used to illustrate the proposed methodology.

AMS 2000 subject classifications: Primary 62H15, 62G20; secondary 62P10.

Keywords and phrases: Relative change rate, Temporal gene expression, Empirical distribution, High dimensional data, Threshold point.

1. INTRODUCTION

Gene expressions are widely used in biological and biomedical studies as they contain rich information for human beings. By using high throughput methods such as oligonucleotide and DNA microarray, serial analysis of gene expression (SAGE) and RNA-sequencing, discrete functional data can be generated on gene expressions (Bjarnason *et al.*, 2003; Cho *et al.*, 1998; Spellman *et al.*, 1998; Yuan and Lin, 2007). From the observed measurements of gene expressions, we would be able to classify gene expression patterns, and find gene regulatory network and gene environmental interaction (Duan *et al.*, 2012). As a result, the longitudinal observations with appropriate number of time points are significantly useful for studying the change of individual gene over time and the effects of other factors. In recent decade, many statistical techniques have been developed to interpret gene expression data. From the point of gene classification, the frequently used methods are various clustering procedures, fold changes, ANOVA, etc. (Draghici *et al.*, 2003; Eisen *et al.*, 1998; Li *et al.*, 2002). Most of these methods are related to the classical and parametric statistical methods. However, when the number of observed values for each time point is less than the number of observed time points, the classical parametric methods of classification are no longer performing appropriately and cannot be used to analyse the gene data. To solve this problem, the reduction of dimensionality for the vector or function of observed measurements is a key in the discrete time points for gene expressions.

For this reason many methods were developed for the reduction of dimensionality to random vector or random function such as principal component analysis (PCA) (Anderson, 2003; Yeung and Ruzzo, 2001), kernel principal component analysis (Kernel PCA) (Schölkopf and Müller, 1998) or smooth spline (Hastie and Tibshirani, 1990). Further, one can be able to define a specific model using smooth spline method to estimate the mean and corresponding variance functions of gene expressions (Fang *et al.*, 2012). All of the aforementioned methods are to indirectly reduce the dimensionality by using the small number of variables or parameters to represent the high-dimensional vector or function. On the other hand, in functional data analysis, if we can directly eliminate some abundant negligible measurements from some time points, the remaining observed values can be efficiently analyzed by the classical statistical methods, which is also a kind of strategies of high throughput data analysis.

In many cases, gene expressions show some changeability within a certain time period and then keep stable or change faintly after that period. In other words, the observing time can be partitioned into two time periods such that for the first period the behaviour of gene expressions exhibit some variability over time and then gene expressions maintain a stable state or vary slightly during the second period. As a result, we are able to eliminate from second time period the observed measurements, which contain less effective information for the gene expressions, and keep the measurements observed at the first time period.

As a motivating example, we consider the data set with 18 genes in *P. aeruginosa* expressed in 24 conditions (see Table 1). For each condition, each gene was measured every 30

*Corresponding author.

Table 1. 18 genes in P. aeruginosa *expression*

Code	**Name**	**Protein**	**Ratio**	**Remarks**
A6	PA5283	Probable transcriptional regulator	99.68 %	48 % similar to putative transcriptional regulator (Bacillus subtilis)
B3	PA2975 (rluC)	Ribosomal large subunit pseudouridine synthase C	99.68 %	Transcription, RNA processing &degradation
B4	PA4991	Hypothetical protein	100 %	Unknown
B5	PA5237	Conserved hypothetical protein	100 %	87 % similar to hypothetical yigC gene product of E. coli
C4	PA0287 (gpuP)	3-guanidinopropionate transport protein	100 %	Transport of small molecules
D1	PA3115 (fimV)	Motility protein FimV	100 %	Membrane proteins; Motility & Attachment
D2	PA3879 (narL)	Two-component response regulator NarL	99.67 %	74 % similar to E.coli NarL protein
D3	PA0894	Hypothetical protein	99.02 %	Unknown
E5	PA1875	Probable outer membrane protein precursor	100 %	41 % similar to alkaline pro-tease secretion protein AprF
E6	PA0573	Hypothetical protein	100 %	Unknown
F2	PA3902	Hypothetical protein	100 %	Unknown
F3	PA3212	Probable ATP-binding component of ABC transporter	100 %	65 % similar to putative amino acid abc transporter, ATP-binding protein (Helicobacter pylori J99)
F5	PA2997 (nqrC)	Na+translocating NADH: ubiquinone oxidoreductase subunit Nrq3	100 %	Energy metabolism

G2	PA0649 (trpG)	Anthranilate synthase component II	100 %	Energy metabolism;
				Biosynthesis of co-factors, prosthetic groups & carriers;
				Amino acid biosynthesis & metabolism

G5	PA1748	Probable enoyl-CoA hydratase/isomerase	98.2 %	61 % similar to putative enoyl-coA hydratase EchA3 (Mycobacterium tuberculosis)
G6	PA3771	Probable transcriptional regulator	99.22 %	54 % similar to a region of putative regulatory protein (Streptomyces coelicolor)
H3	PA1841	Hypothetical protein	100 %	43 % similar to hypothetical yeaK gene product of (E. coli)
σ70	σ70	σ factor		As a control

minutes for 21 hours and, finally, each gene has 43 observations. From the gene expression data we can see that under most of the conditions, observations for many genes look stable after some time points. Figure 1 shows the gene expressions for genes PA4491 and PA6287 under 24 conditions. It is observable that under most of the conditions, observations for gene PA4491 look stable after seven and half hours. Similar to PA4491, gene PA6287 has similar behaviour after six and half hours. Theoretically, the visualizations of gene expressions demonstrate that for each gene after a threshold time point, which is between starting time point and the ending time point, the measurements under the conditions are rather stable or fluctuate slightly. Therefore, after the threshold time point the gene expression has less information on the gene behaviour and the most information for the gene can be captured by the measurements observed before the threshold time point. In other words, the measurements observed after the threshold point is not as important as that observed before the threshold point and thus can be eliminated for further analysis. Furthermore, from the point of application, if we detect the threshold time point such that after that time the gene expression keeps stable or fluc-

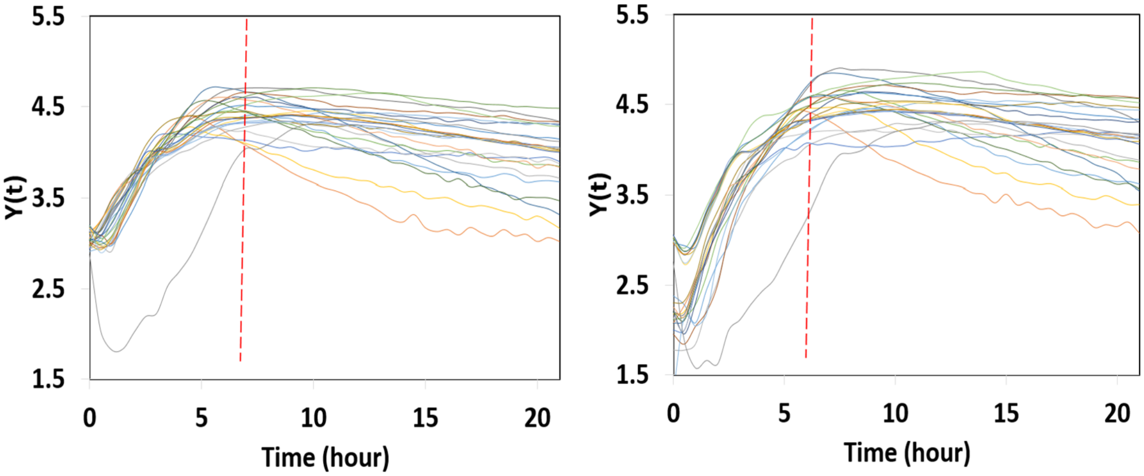

Figure 1. Gene expression for gene PA1841 *(left) and* PA0287 *(right) and the dash line in both figures indicates the time point which after that point curves either look stable or fluctuate slightly.*

tuates very slightly, the observations can be terminated after this threshold time point and thus the observing time can be shortened and cost of observations can be saved. Meanwhile, genes can be classified based on the threshold time points they take to get stable. Moreover, for the purpose of comparison for gene expressions we only need to collect or use few time points instead of collecting information for a longer period. Also, when two genes have the significantly different threshold time points, it is obvious that the varied behaviours of two genes are different from each other. We found that two gene expressions may have the similar behaviour when they have the same or close threshold time points. Anyway, two gene expressions can be compared by using the classical methods if the reduced vectors of measurements for these two gene expressions have less dimensions than sample sizes. Furthermore, it means that having the similar values of threshold time points for two genes is the necessary (not sufficient) condition for the similarity of two gene expressions. However, it does not means that two gene expressions with same or close threshold time points have the similar behaviour. On the other hand, by detecting the threshold time points for all gene expressions, some singular gene expressions can also be found if most of gene expressions have the similar values of threshold time points but the threshold time points for specific gene expressions have the very small/large values or these gene expressions exhibit some other singularity. For example, we found gene *PA4491* shows some singularity because it exhibits some kind linearity but others keep constant after threshold time points.

It should be pointed out that the threshold point discussed in the current paper is different from the change-point, at which, the sequential data exhibit the *abrupt* changes in the generative parameters of the static or dynamic stochastic system. The change-point problem, first introduced in the quality control context, has since developed into a fundamental problem in the areas of statistical control theory, stationarity of a stochastic process, estimation of the current position of a time series, testing and estimation of change in the patterns of a regression model, and most recently in the comparison and matching of DNA sequences in microarray data analysis. Numerous methodological approaches have been implemented in examining change-point models. Maximum-likelihood estimation, Bayesian estimation, isotonic regression, piecewise regression, quasi-likelihood and non-parametric regression are among the methods which have been applied to resolving challenges in change-point problems. Grid-searching approaches have also been used to examine the change-point problem. The pioneer work for the detection of change points can be retrospected in Page (1954, 1955) and Lorden (1971). On the contrary the threshold point problem arises in the functional data, in which the observed curve exhibits some variability before the threshold point and then keep stable until the last observing time point. There is no abrupt change for the curve at the threshold point.

In the current paper several algorithms are constructed to detect the threshold time points based on the classical Hotelling statistic, high dimensional test statistic and empirical distribution based statistic of sample derivatives for gene expressions. The remainder of this paper is organized as follows. Three fundamental methods and corresponding algorithms are introduced in Section 2 for detecting the threshold time points in case of finite number of time points. Also, in Section 3, simulation studies are used to check the efficiency of each method and compare the results for all methods proposed in Section 2. The aforementioned gene expression data are analyzed by using proposed methods in Section 4, which follows with a concluding remarks in Section 5. The theoretical proofs are given in Appendix.

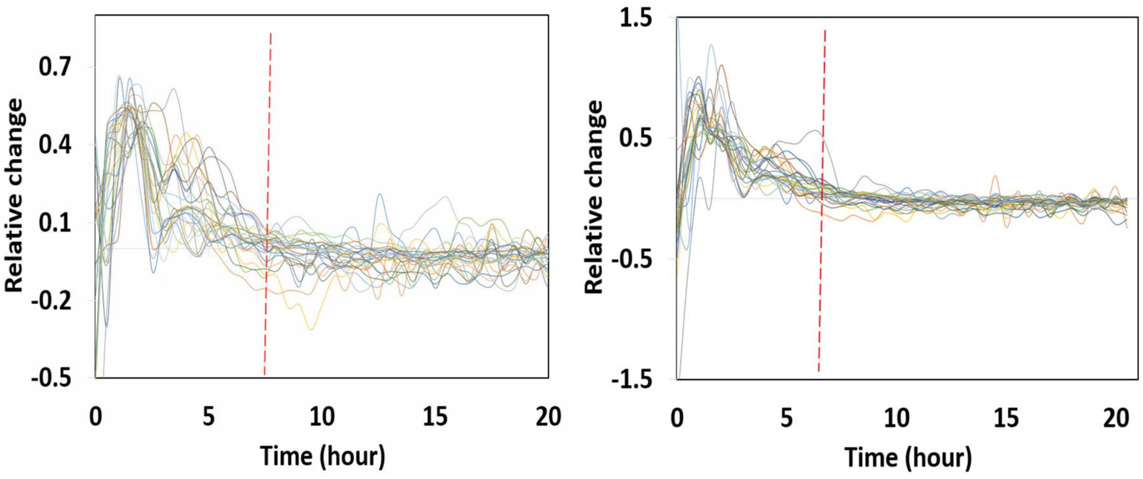

Figure 2. Gene expressions for genes PA1841 *(left) and* PA0287 *(right) and the dash lines in both figures indicate the time points after which curves either look stable or fluctuate slightly.*

2. METHODS FOR DETECTION OF THRESHOLD TIME POINT

Let $Y_{ij}(t_{ik})$ denote the realization of gene expression of the i^{th} gene at time t_{ik} ($t_{ik} \leq T$) under condition j which $i = 1, \cdots, g$, $j = 1, \cdots, c$; and $k = 1, \cdots, p_i + 1$. Eq. (1) shows the considered model to realize the gene expressions.

$$Y_{ij}(t_{ik}) = \mu_{ij}(t_{ik}) + \epsilon_{ij}(t_{ik}) \tag{1}$$

In Eq. (1), $\epsilon_{ij}(t_{k_i})$ is the random error with mean zero and variance $\sigma_i^2(t_{ik})$. For a given gene (fixed value of i), Eq. (1) is expressed as follows

$$Y_j(t_k) = \mu(t_k) + \epsilon_j(t_k) \tag{2}$$

where $k = 1, 2, .., p + 1$ and $j = 1, ..., c$.

We say a gene expression has the threshold time point, denoted by τ if this gene expression demonstrates some kind variability before τ and keeps stable after τ. Ideally, the mean function $\mu(t)$ of this gene expression changes over the time period $[0, \tau]$ and keeps a constant when $t \geq \tau$. Hence, the model for gene expression with the threshold time τ can be written as

$$Y(t) = \mu(t)I(t \leq \tau) + \mu_\tau I(t > \tau) + \epsilon(t) \tag{3}$$

where μ_τ is a constant and τ is the threshold time point. Now, finding reasonable estimate for τ is the main attempt in this paper. Since the relative change rate of gene expression is zero for $t > \tau$ from the model (3), the threshold point can be detected by modeling the relative change rate $Z(t)$ for the gene expression $Y(t)$, which can be defined as the derivative of $Y(t)$. Now from the observations $Y(t_k)(k = 1, 2, ..., p + 1)$, the observations of change rate $Z(t)$ can obtained as

$$Z(t'_k) = \frac{Y(t_{k+1}) - Y(t_k)}{t_{k+1} - t_k}.$$

where $t'_k = (t_{k+1} + t_k)/2$ for $k = 1, \cdots, p$. Therefore, when the mean function is stable for any $t > \tau$ and, relative change rate $Z(t)$ for the gene expression with the threshold point τ can be modeled as follows:

$$Z(t) = \mu_Z(t)I\{t \leq \tau\} + \epsilon_Z(t) \tag{4}$$

where $\mu_Z(t)$ is the mean function of $Z(t)$ and $\epsilon_Z(t)$ is the error variable with zero mean and variance $\sigma_Z^2(t)$. From this model, one can see that the mean function of relative change rate $Z(t)$ for gene expression equals zero when $t > \tau$.

Figure 2 shows the relative change rates for genes *PA1841* and *PA6287* under 24 conditions. As it is shown in this figure, the relative change rate for $t \geq 7.5$ hours is approximately zero on gene *PA1841*. Also, gene *PA6287* has similar patter and its relative change gets near to zero after $\tau = 6.5$ hours.

Based on the discussion above, to detect the threshold point for the gene expression, we need to find the time point $0 < \tau < T$ such that $\mu_Z(t) = 0$ for $t > \tau$. This question can be transferred into the hypothesis testing. Towards this end, the appropriate value for the threshold point τ should satisfy that at the significance level α, the null hypothesis $H_0 : \mu_z(t) = 0$ for $t > \tau$ cannot be rejected and at the same significance level $H_0 : \mu_Z(t) = 0$ is rejected for some $t \leq \tau$. Since the gene expression are observed at discrete time points, the threshold time point τ for gene expression can be decided if the null hypothesis $H_0 : \mu_Z(t'_k) = \mu_Z(t'_{k+1}) = ... = \mu_Z(t'_p) = 0$ can not be rejected at the significance level α and $H_0 : \mu_Z(t'_{k-1}) = \mu_Z(t'_k) = \mu_Z(t'_{k=1}) = ... = \mu_Z(t'_p) = 0$ is rejected at the same significance level for the time point t'_k. The backward procedure can be used to find t'_k. At first, the null hypothesis $H_0 : \mu_Z(t'_p) = 0$ is tested for the last observed time point t'_p of a specific gene expression. If H_0 is not rejected at this stage then $H_0 : \mu_Z(t'_p) = \mu_Z(t'_{p-1}) = 0$ will be tested. This procedure is continued as long as such null hypothe-

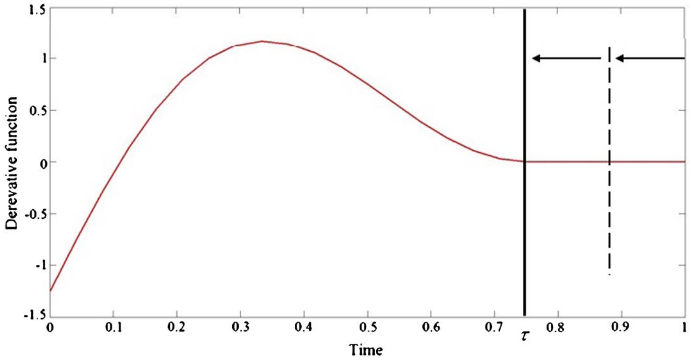

Figure 3. This figure indicates the general scheme of the method that can be used to find the approximate value for τ.

sis is not rejected. Otherwise, the first time point t'_k such that $H_0 : \mu_Z(t'_{k-1}) = ... = \mu_Z(t'_p) = 0$ is rejected could be an appropriate estimate for τ. A scheme of this method is indicated in Figure 3.

In what follows we propose several procedures for the detection of threshold time point for the gene expression.

2.1 Hotelling's statistic

Now based on the model (3), if it is assumed that the error variable $\epsilon(t)$ is normally distributed, then the error term $\epsilon_Z(t)$ can also be normally distributed. Hence, Hotelling's T^2 can be used as test statistic for detecting the threshold time point. Note that under the normal assumption, $Z_j(t'_k)$ for $k = 1, ..., p$ and $j = 1, .., c$ are normally distributed. Then for $l = 1, 2, ..., p$, $\mathbf{Z}_j^{(l)} = (Z_j(t'_l), ..., Z_j(t'_p))^T \sim MVN(\boldsymbol{\mu}_Z^{(l)}, \boldsymbol{\Sigma}_Z^{(l)})$ where $\boldsymbol{\mu}_Z^{(l)}$ and $\boldsymbol{\Sigma}_Z^{(l)}$ are the mean and the variance of $\mathbf{Z}_j^{(l)}$, respectively. Hence, under the null hypothesis $H_0 : \boldsymbol{\mu}_Z^{(l)} = 0$, the statistic S_H given in Eq. (5) follows F distribution with $p - l + 1$ and $c - (p - l + 1)$ degrees of freedom, provided that $(p - l) < c$.

$$S_H = \frac{T_l^2}{c-1} \frac{c - (p - l + 1)}{p - l + 1} \tag{5}$$

where

$$T_l^2 = c\bar{\mathbf{Z}}^{(l)^T} \mathbf{S}_l^{-1} \bar{\mathbf{Z}}^{(l)}$$

and

$$\mathbf{S}_l = \frac{1}{c-1} \sum_{j=1}^{c} \left(\mathbf{Z}_j^{(l)} - \bar{\mathbf{Z}}^{(l)}\right) \left(\mathbf{Z}_j^{(l)} - \bar{\mathbf{Z}}^{(l)}\right)^T,$$

$$\bar{\mathbf{Z}}^{(l)} = \frac{1}{c} \sum_{j=1}^{c} \mathbf{Z}_j^{(l)}.$$

Now, for given l the hypothesis $H_0 : \mu_Z(t'_l) = ... = \mu_Z(t'_p) = 0$ can be tested using Hotelling's T^2 for a specific gene. Therefore, the approximate value of τ is calculated using Algorithm 1.

Algorithm 1: Detecting τ using Hotelling's T^2

```
1. Let $l = p - 1$.
2. Compute $T_l^2$.
3. If $S_H$ is greater than $F_{\alpha, p-l+1, c-(p-l+1)}$, go to step 7.
     Else go to step 4.
4. Let $l := l - 1$.
5. If $(p - l) > c$ go to step 6. Else return to step 2.
6. This method can not be used any more as number of
     time points $(p - l + 1)$ is greater than the sample size
     $(c)$.
7. The value of estimate for $\tau$ is $t_{l-1}$.
```

Algorithm 2: Detecting τ using orthogonal transformations

```
1. Let $l := p - 1$.
2. Test $H_l : \mu_X^{(l)}(t'_l) = 0$, ..., $H_p : \mu_X^{(l)}(t'_p) = 0$ simultaneously using $t$-test. The significance level for each test
     is equal to $\alpha_0 = 1 - (1 - \alpha)^{(p-l+1)}$ where $\alpha$ is the significance level to test $H_0 : \mu_Z(t'_l) = ... = \mu_Z(t'_p) = 0$.
3. If for all value of $l \leq r \leq p$ $H_r : \mu_X(t'_r) = 0$ is not
     rejected for $\alpha_0$ go to step 7; else go to step 4.
4. $l := l - 1$.
5. If $p - l > c$ go to step 6; else return to step 2.
6. This method can not be used any more as number of
     time points $(p - l + 1)$ is greater than the sample size
     $(c)$.
7. The value of estimate for $\tau$ is $t_l$.
```

Meanwhile, it is possible to make this method more simple using orthogonal transformations. Let $\lambda_l, .., \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}_Z^{(l)}$ and $\boldsymbol{\beta}_r = (\beta_{rl}, ..., \beta_{rp})^T$ are the corresponding eigenvectors of λ_r $(r = l, ..., p)$. Without loss of generality we can assume that norm of each eigenvector is 1. Also, let $\boldsymbol{\beta}^{(l)} = (\boldsymbol{\beta}_l, ..., \boldsymbol{\beta}_p)$ be the matrix which consists of eigenvectors and $\boldsymbol{X}_j^{(l)} = \boldsymbol{\beta}^{(l)^T} Z_j^{(l)}$. The mean of $\boldsymbol{X}_j^{(l)}$ is $\mu_X^{(l)} = \boldsymbol{\beta}^{(l)^T} \mu_Z^{(l)}$ and the variance is

$$\Sigma_X^{(l)} = \boldsymbol{\beta}^{(l)} \Sigma_Z^{(l)} \boldsymbol{\beta}^{(l)^T} = \begin{pmatrix} \lambda_l & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

Since $\boldsymbol{\beta}_{(l)}$ contains orthogonal vectors $\boldsymbol{\beta}^{(l)} \boldsymbol{\beta}^{(l)^T} = I$. Therefore, $\boldsymbol{\beta}^{(l)^T} \boldsymbol{\mu}_Z^{(l)} = \mathbf{0}$ is equivalent to $\boldsymbol{\mu}_Z^{(l)} = 0$. Hence, instead of testing $H_0 : \boldsymbol{\mu}_Z^{(l)} = 0$ we can test $H_0 : \boldsymbol{\mu}_X^{(l)} = 0$. Further, since $\Sigma^{(l)}$ is diagonal, $X_j(t'_l), ..., X_j(t'_p)$ are independently distributed and we can test $H_l : \mu_X^{(l)}(t'_l) = 0, ..., H_p : \mu_X^{(l)}(t'_p) = 0$ simultaneously using t-test for each of the hypothesizes. Therefore, using Algorithm 2 for a given gene we are able to find the suitable value of τ.

2.2 High-dimensional test statistic

Based on the Algorithms 1 and 2 for each step one time point is added until the threshold is detected. It means after a while the number of time points will be more than the sample size (c), which means we are dealing with high-dimensional data. As a result, Hotelling's T^2 statistic does not work when the number of selected time points for hypothesis test are more than the sample size. To overcome this problem the high-dimensional test statistic used to test $\mu_z(t'_l) = ... = \mu_z(t'_p) = 0$ is

$$T_D^l = \frac{c\bar{\mathbf{Z}}^{l^T}\bar{\mathbf{Z}}^l}{\operatorname{tr}(\mathbf{S}_l)}.$$

Further, under the null hypothesis with normal assumption, T_D^k approximately is normally distributed (Nishiyama *et al.*, 2013). Meanwhile, the density function of T_D^l can be approximated as Eq. (6) and the approximate upper percentile of T_D^l using Eq. (7).

$$(6)\quad P(\frac{\tilde{T}_D^l}{\tilde{\sigma}} \le z) = \Phi(z) - \frac{1}{2}\phi(z)\left[\frac{1}{\sqrt{p'}}C_3h_2(z) + \frac{1}{p'}C_4h_3(z) + \frac{1}{p'}C_6h_5(z) + \frac{1}{c}\right]h_1(z) + O(p'^{-3/2})$$

$$(7)\quad \widehat{Z}(\alpha) = \Phi(\alpha) + \frac{1}{\sqrt{p'}}\frac{\sqrt{2}\widehat{a}_3}{3\sqrt{\widehat{a}_2^3}}(\phi^2(\alpha) - 1) + \frac{1}{p'}\left\{\frac{\widehat{a}_4}{2\widehat{a}_2^2}\Phi(\alpha)(\Phi^2(\alpha) - 3) - \frac{2\widehat{a}_3^2}{9\widehat{a}_2^3}\Phi(\alpha)(2\Phi^2(\alpha))\right\} + \frac{1}{2c}\Phi(\alpha) + O(p'^{-3/2})$$

where p', C_3, C_4, C_6 and $\tilde{\sigma}$ are defined as $p - l + 1$, $\frac{\sqrt{2}a_3}{3\sqrt{a_2^3}}$, $\frac{a_4}{2a_2^2}$, $\frac{a_3^2}{9a_2^2}$, and $\sqrt{\frac{2a_2}{a_1^2}}$, respectively. Also, $\widehat{a}_1$, $\widehat{a}_2$, $\widehat{a}_3$, $\widehat{a}_4$, b_0, b_1, b_2, and b_3 are defined as follows

$$\begin{aligned}
\widehat{a}_1 &= \frac{\operatorname{tr}(S_l)}{p'} \\
\widehat{a}_2 &= \frac{c^2}{p'(c-1)(c+2)}\left[\operatorname{tr}(S_l^2) - \frac{\operatorname{tr}(S_l)^2}{c}\right] \\
\widehat{a}_3 &= \frac{c}{(c-1)(c-2)(c+2)(c+4)} \\
&\qquad \times\left\{\frac{\operatorname{tr}(S^3)}{p'} - 3(c+2)(c-1)\widehat{a}_1\widehat{a}_2 - cp'^2\widehat{a}_1^3\right\} \\
\widehat{a}_4 &= \frac{1}{b}\left(\frac{\operatorname{tr}(S^4)}{p'} - p'b_1\widehat{a}_1 - p'^2b_2\widehat{a}^2\widehat{a}_2 - p'b_3\widehat{a}_2^2 - cp'^3\widehat{a}_1^4\right) \\
b_0 &= c(c^3 + 6c^2 + 21c + 18), \quad b_1 = 2c(2c^2 + 6c + 9) \\
b_2 &= 2c(3c + 2), \quad b_3 = c(2c^2 + 5c + 7).
\end{aligned}$$

Eventually, using Algorithm 3 the value of threshold point can be determined.

Algorithm 3: Detecting τ using high-dimensional test statistic

```
1. Let $l := p - 1$.
2. Calculate $T_D^l = \frac{c\bar{\mathbf{Z}}^{l^t}\bar{\mathbf{Z}}^{l^t}}{\operatorname{tr}(\mathbf{S}_l)}$
3. Let $\widehat{Z}(1 - \frac{\alpha}{2})$ be equal to
       $\Phi(1 - \frac{\alpha}{2}) + \frac{1}{\sqrt{p'}}\frac{\sqrt{2}\widehat{a}_3}{3\sqrt{\widehat{a}_2^3}}(\phi^2(1 - \frac{\alpha}{2}) - 1)$
       $+ \frac{1}{p'}\left\{\frac{\widehat{a}_4}{2\widehat{a}_2^2}\Phi(1 - \frac{\alpha}{2})(\Phi^2(1 - \frac{\alpha}{2}) - 3) -$
       $\frac{2\widehat{a}_3^2}{9\widehat{a}_2^3}\Phi(1 - \frac{\alpha}{2})(2\Phi^2(\alpha))\right\} + \frac{1}{2c}\Phi(1 - \frac{\alpha}{2}).$
4. If $|T_D^l|$ is greater than $\widehat{Z}(1 - \frac{\alpha}{2})$ go to step 8; Else go
   to step 5.
5. Let $l := l - 1$.
6. If $l = 0$ go to step 7; Else return to step 2.
7. No threshold was detected.
8. The value of estimate for $\tau$ is $t_l$.
```

2.3 Empirical distribution based test

Another way to find the approximate value of τ is to use the properties of empirical distribution for relative change rate of gene expression. From model (4), it is known that $\mu_Z(t) = 0$ for $t \ge \tau$. Further, if $Z(t)$ is normally distributed, then $F_{Z(t)}(0)$ is equal to $\frac{1}{2}$ for $t \ge \tau$. Note that $F_{Z(t)}(x)$ can be estimated by the empirical distribution $\widehat{F}_{Z(t)}(x) = \frac{1}{c}\sum_{j=1}^{c} I(Z_j(t) < x)$ and strong law of large number implies that $\widehat{F}_{Z(t'_k)}(x)$ converges to $F_{Z(t'_k)}(x)$ almost surely for any value of x. Also, from Central Limit Theorem, $\sqrt{4c}(\widehat{F}_{Z(t)}(0) - 1/2) \to N(0,1)$ in distribution. Furthermore, to detect the threshold time point for the gene expression, we need the following theorem, the proof of which is given in Appendix.

Theorem. *Let $Z_j(t'_k)$ be normally distributed for $j = 1, ..., c$ and $k = 1, ..., p$. Also, for $t'_k \ge \tau$, $\mu_Z(t'_k) = 0$. Then, the distribution of Eq. (8) converges to standard normal distribution as $c \to \infty$.*

$$(8)\quad Z_\tau = \frac{\sum_{\{r:t'_r \ge \tau\}}(\widehat{F}_{Z(t'_k)}(0) - \frac{1}{2})}{\sqrt{\frac{1}{c}[\frac{1}{4}(2c_\tau - c_\tau^2) + \sum_{(r,s)\in A} F_{Z(t'_r),Z(t'_s)}(0,0)]}}$$

where $A = \{(r,s)|r \neq s, t'_r \ge \tau$ and $t'_s \ge \tau\}$ and $c_\tau = \#\{r; t'_r \ge \tau\}$.

Based on above theorem, we can construct an algorithm to find the approximate value for τ. For this reason we first set $\tau = t'_p$ and find the asymptomatic test statistic (8) to test $H_0 : \mu_Z(t'_p) = 0$. If H_0 is not rejected at significance level of α, we repeat this test for $\tau = t_{p-1}$ and if $H_1 :\sim H_0$ is not significant, this loop is continued. The appropriate

Algorithm 4 Estimating τ using empirical distribution function for $Z(t)$

```
1. Let $l := p$ and $\tau = t'_l$
2. Compute $Z_\tau$.
3. If $|Z_\tau| > \Phi^{-1}(1-\alpha/2)$, then go to step 5; Else go to
   step 4. ($\Phi(.)$ is the distribution of standard normal
   random variable)
4. Let $l := l - 1$ and go to step 2.
5. The value of estimate for $\tau$ is $t'_l$.
```

value for τ is the time point before the first time point for which H_1 is significant. (See Algorithm 4).

Usually, as the time increases the relative change for each specific gene approaches zero gradually, which affects our method for determining the appropriate estimate of τ. To overcome this problem we can use the sample second-order derivative instead of the sample change rate (first-order derivative) and the method explained in the change rate can be applied to the second-order derivative. The sample second-order derivative at $t''_k = (t'_{k+1} + t'_k)/2$ is defined as

$$W_j(t''_k) = \frac{Z_j(t'_{k+1}) - Z_j(t'_k)}{t'_{k+1} - t'_k}$$

for $k = 1, 2, ..., p-1$. Similar to the first-order derivative we can use the sample second-order derivative for the method introduced above. As a result, instead of testing $H_0 : \mu_Z(t'_k) = 0$ for $t'_k \geq t'_l$ (t'_l is the time point corresponding to the steps in our algorithm), we try to test $H_0 : \mu_W(t''_k) = 0$ for $t''_k \geq \tau'$ for some τ'. Here, whenever H_0 is not rejected, it means the $\mu_Z(t'_k) = C$ for $t'_k \geq \tau'$ where C is a constant, and the next step is to test $H_0 : C = 0$. For this sake we can define $U_j^{\tau'}$ as

$$U_j^{\tau'} = \sum_{t'_r \geq \tau'} Z_j(t'_r)$$

and it is clear that $U_1^{\tau'}, ..., U_c^{\tau'}$ are mutually independent and have identical normal distributions. Therefore, under the null hypothesis, $Z_U^{\tau'} = \frac{\bar{U}^{\tau'}}{S_U^{\tau'}/\sqrt{c}}$ has standard normal distribution, where

$$\bar{U}^{\tau'} = \sum_{j=1}^{c} U_j, \quad S_U^{2\tau'} = \frac{1}{c-1}\sum_{j=1}^{c}(U_j^{\tau'} - \bar{U}^{\tau'})^2$$

Now, using Algorithm 5, we can detect the threshold time point based on the second-order derivative of the gene expression. Simulation results obtained by using the sample second derivative demonstrate that the accurate approximation to true value of τ can be obtained.

Algorithm 5: Estimating τ using empirical distribution function for $W(t)$

```
1. Let $l := p - 1$.
2. Let $\tau' = t''_l$
3. Compute $W_{\tau'}$ by using the expression in (8) where
   instead of sample change rate, sample second-order
   derivatives are used (note that $W_{\tau'}$ has same asymp-
   totic distribution as $Z_\tau$).
4. If $|W_{\tau'}| > \Phi(1-\alpha/2)$, then go to step 6; Else go to
   step 5.
5. Let $l := l - 1$ and get to step 2.
6. If $|Z_U^{\tau'}| > \Phi(1-\alpha/2)$ ($t = t_{l-1}$, then there is no time
   point $\tau'$ such that the mean function $\mu(t)$ keeps stable
   for $t \geq \tau'$; else go to step 7.
7. The value of estimate for $\tau$ is $t''_{l-1}$.
```

3. SIMULATION STUDY

To find the advantages and disadvantages of proposed methods the simulation studies are designed. For this reason, Eq. (9) is considered as the mean function.

$$(9) \qquad \mu(t) = \left(\frac{-\cos(2\pi t)}{2\pi} - \frac{4}{3}(t - \frac{3}{4})^3 + t\right) I(t < \tau)$$

Also, we consider 25 equally spaced time points between 0 and 1 as $t_1, ..., t_{25}$ and τ is fixed as t_{19}. In addition, $\boldsymbol{Z}_j = (Z_j(t_1), ..., Z_j(t_p))$ is normally distributed with mean function

$$\mu_Z(t) = \left(\sin(2\pi t) - 4(t - \frac{3}{4})^3 + 1\right) I(t < \tau)$$

and variance $\boldsymbol{\Sigma}_Z = \sigma^2 \boldsymbol{GRG}$ (this covariance structure explained in Fang *et al.*, 2012, where $\rho = 0.5$, $\alpha = 2$, $\sigma^2 = 0.2$. Figure 4 demonstrates the mean function, first-order derivative, and second-order derivative for $t \in [0, 1]$ and the vertical dashed line indicates the threshold time point.

The simulation results based on 10000 replications for three different numbers of conditions ($c = 20, 40, 60$) are given in Table 2. Results show that the best method for detecting τ is the empirical distribution test based on the sample second-order derivatives of gene expression (Algorithm 5), which gives the better approximation to the true value of τ than the same method based first-order derivatives (Algorithm 4). Moreover, as the number of conditions increases the detected threshold time point based on the second-order derivatives suffers less change than the other methods (see Mean Absolute Deviation(MAD)). Meanwhile, when $c = 20$, the estimate of τ based on empirical distribution test of second-order derivatives is close to the true value of τ for all three different significant values (α). In summary, empirical distribution test based on the second-order derivative can detect the threshold time point most accurately among these methods.

In order to further investigate the performance of Algorithms 3–5, we consider the scenarios in which the number

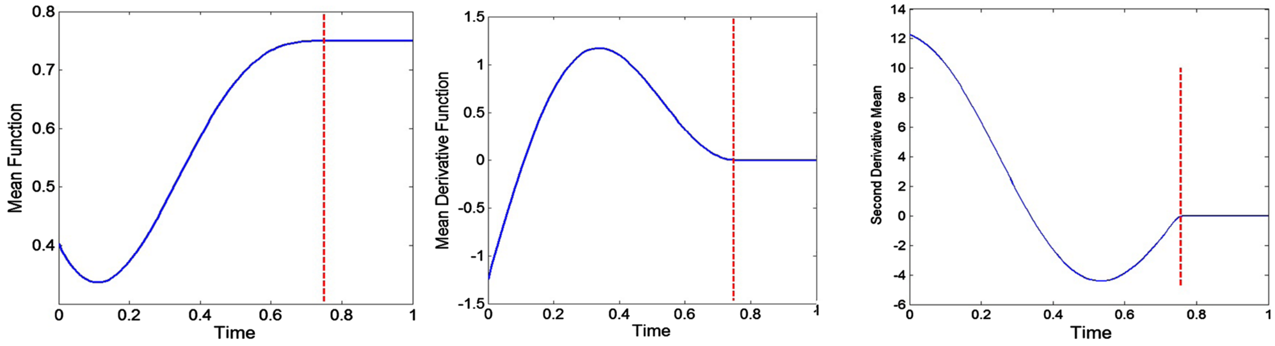

Figure 4. Mean function $\mu(t)$, and its first-order and second-order derivatives $\mu_Z(t), \mu_W(t)$ are indicated by left, center and right, respectively. τ is indicated using dashed line in each plot.

Table 2. The detected values of $\tau = 19$ with MAD using the proposed algorithms 1–5

Method	α	c=20		c=40		c=60	
		Detected Value	MAD	Detected Value	MAD	Detected Value	MAD
Algorithm 1	0.10	17.0612	1.9539	16.6804	2.3660	16.0435	1.9915
	0.05	17.0172	1.9915	16.3954	2.6502	15.9730	3.0331
	0.01	16.8626	2.1661	16.0342	2.979	15.5013	3.5388
Algorithm 2	0.10	18.8212	2.1778	18.6503	1.8401	18.7409	1.7005
	0.05	18.8089	2.1545	18.6832	1.8544	18.7607	1.7182
	0.01	18.7718	2.1321	18.6756	1.8404	18.7387	1.7148
Algorithm 3	0.10	19.9177	1.3021	19.9073	1.1658	19.371	0.5149
	0.05	19.6042	1.1962	19.8014	0.9884	19.5960	0.7092
	0.01	18.3078	1.1881	19.5960	0.8535	19.3146	0.5092
Algorithm 4	0.10	16.9828	2.6259	17.2211	2.3526	17.3517	2.1981
	0.05	16.6951	2.7269	17.0805	2.3714	17.1648	2.2149
	0.01	16.5043	2.8140	16.6433	2.5593	16.8009	2.3642
Algorithm 5	0.10	19.3012	0.7233	19.2844	0.7021	19.2917	0.7107
	0.05	19.2021	0.5939	19.2080	0.5987	19.1769	0.5498
	0.01	19.1250	0.4675	19.0736	0.3552	19.0611	0.3236

of time points between τ and T is more than the number of conditions. The mean function considered here is as follows:

$$\mu(t) = \left(\sin(4\pi t) - 2(t - \frac{3}{4})^3 - 12.1936t\right) I(t \leq \tau)$$

As a result, the first change rate has the following mean function:

$$\mu_{\mathbf{Z}}(t) = \left(4\pi\cos(4\pi t) - 6(t - \frac{3}{4})^2 - 12.1936\right) I(t \leq \tau)$$

and similar to the previous model we can use the following covariance structure $\mathbf{\Sigma_Z} = \sigma^2 \boldsymbol{GRG}$ where $\rho = 0.5$, $\alpha = 2$, $\sigma^2 = 0.2$. Also, the 50 equally spaced time points between 0 and 1 as $t_1, ..., t_{50}$ are chosen and τ is fixed as t_{25}. Now, we try to compare methods which are more compatible with high-dimensional cases. We know Algorithms 3, 4 and 5 can still be used for detecting the threshold time point even when the number of time points between the threshold point τ and T is greater than the number of conditions. According to Table 2, Algorithms 1 and 2 do not perform as well as Algorithms 3, 4 and 5. Thus we only compare Algorithms 3, 4 and 5. We choose the number of conditions $c = 10, 15, 20$ for Algorithm 3 and $c = 20, 25, 30$ for Algorithms 4 and 5. From the simulation results it can be realized that these three methods perform really well and are close to each other. Generally, we can say Algorithm 5 performs better than the other methods in terms of mean absolute deviation. However, Algorithm 3 is preferable for the small number of conditions.

4. EXAMPLE

Now, we consider the analysis of the data set of 18 genes in *P. aeruginosa* expressed in 24 conditions (see Table 1 in Section 1). For each condition, each gene was measured every 30 minutes for 21 hours and, finally, each gene has 43 observations. To find the appropriate value of τ, at first the empirical test based on the second derivative is used to find τ such that $\forall t \geq \tau$ we have $\mu_W(t) = 0$, which means $\forall t \geq \tau$ $\mu_Z(t) = C$. For the second step as explained before, t-statistics can be used to test $H_0 : C = 0$ for each of these

Table 3. The detected values of $\tau = 25$ with MAD using Algorithms 3, 4 and 5

Method	α	Detected Value	MAD	Detected Value	MAD	Detected Value	MAD
		c=10		c=15		c=20	
Algorithm 3	0.10	25.8344	1.2968	25.2288	0.3684	25.1202	0.1614
	0.05	25.1362	0.2574	24.8628	0.1804	24.7738	0.2290
	0.01	24.8580	0.2628	24.6290	0.3762	24.5454	0.4546
		c=20		c=25		c=30	
Algorithm 4	0.10	25.4335	0.9271	25.4050	0.8681	25.5068	0.8850
	0.05	24.9539	0.4840	25.0891	0.5183	25.1347	0.5101
	0.01	24.8850	0.4271	24.9421	0.1364	25.0179	0.4057
		c=20		c=25		c=30	
Algorithm 5	0.10	25.0854	0.1134	25.0712	0.1037	25.0665	0.0961
	0.05	25.0412	0.0541	25.0506	0.0817	25.0613	0.0571
	0.01	25.0381	0.0421	25.0503	0.0805	25.0475	0.0347

Table 4. Estimated value of τ for 18 genes based on testing sample second derivative

Gene	Estimated τ (in hours)	The value of test statistic for Testing $H_0 : C = 0$	95% Confidence Interval for C		Simillar genes
			Lower Bound	Upper Bound	
PA5283	16.5	0.4045	-0.1129	0.0743	—
PA2975	6.5	1.3752	-0.1027	0.0180	PA0649
PA4991	13	**2.2714**	**-0.0819**	**-0.006**	—
PA5237	14.5	1.1681	-0.0956	0.0242	—
PA0287	6.5	0.4994	-0.0899	0.0534	—
PA3115	12	1.6327	-0.1358	0.0124	—
PA3879	8	0.5422	-0.0718	0.0407	—
PA0894	10.5	1.6587	-0.0602	0.0050	PA3771
PA1875	9.5	0.6356	-0.1297	0.2542	—
PA0573	10	0.4966	-0.1156	0.0689	PA3771, PA3902
PA3902	11.5	0.4663	-0.0859	0.0529	PA3771, PA0573
PA3212	8.5	0.5109	-0.0962	0.1640	—
PA2997	14	1.8172	-0.1243	0.0047	—
PA0649	7.5	0.3682	-0.0816	0.0558	PA1748, PA1941, PA2975
PA1748	6.5	0.6283	-0.0943	0.0485	PA0649
PA3771	9	0.0544	-0.1283	0.1356	PA0894, PA0573, PA3902
PA1841	8.5	1.0206	-0.1064	0.0335	PA0649
σ70	10.5	0.5394	-0.1378	0.0783	—

18 genes (as simulation study indicates even the number of conditions c is close to 20, the algorithm 5 works well). For each step the significance level α is equal to 0.025.

Table 4 gives the results for the detected threshold time points, test statistic for $H_0 : \mu_Z(t) = 0$ and confident intervals for all 18 different genes by using Algorithm 5. The first column of Table 4 lists the names of all 18 genes. The second column in this table gives the estimated threshold time points such that $\mu_W(t) = 0$ (which is equivalent to $\mu_Z(t) = C$) for t greater than or equal that time point for the gene expression in each row. Also, the third column of Table 4 shows the values of test statistic $Z_U^{\tau'}$ for all genes to test $H_0 : C = 0$ (which is equivalent to $\mu_Z(t) = 0$) for $t \geq \tau$ for 18 genes, respectively. Lower bounds and upper bounds of confidence intervals for the constant C are given in Columns 4 and 5 of Table 4, respectively. At first we note that the null hypothesis $H_0 : C = 0$ is rejected at the significance level $\alpha = 0.05$ only for gene *PA4991*, which means the gene *PA4991* increases (decreases) linearly after the time point $t = 13$. Contrastably, the other genes keep stable after the corresponding estimated time point τ. In some sense, gene PA4991 has no threshold time point and thus has the different behaviour from other genes. Also, the number of observed time points cannot be reduced for this gene expression. Further, if we review Table 4 carefully, we can find gene *PA5283* has the largest value 16.5 hours of τ among the other genes. After gene *PA5283*, genes *PA5237* and *PA2997* have the second and third largest values 14.5 hours and 14 hours of τ, respectively. Genes *PA2975*, *PA0287*, *PA1748* have the same smallest value 6.5 hours of τ among the other genes and gene PA0649 has the second smallest value 7.5 hours of τ. Therefore, if we consider the pairwise difference

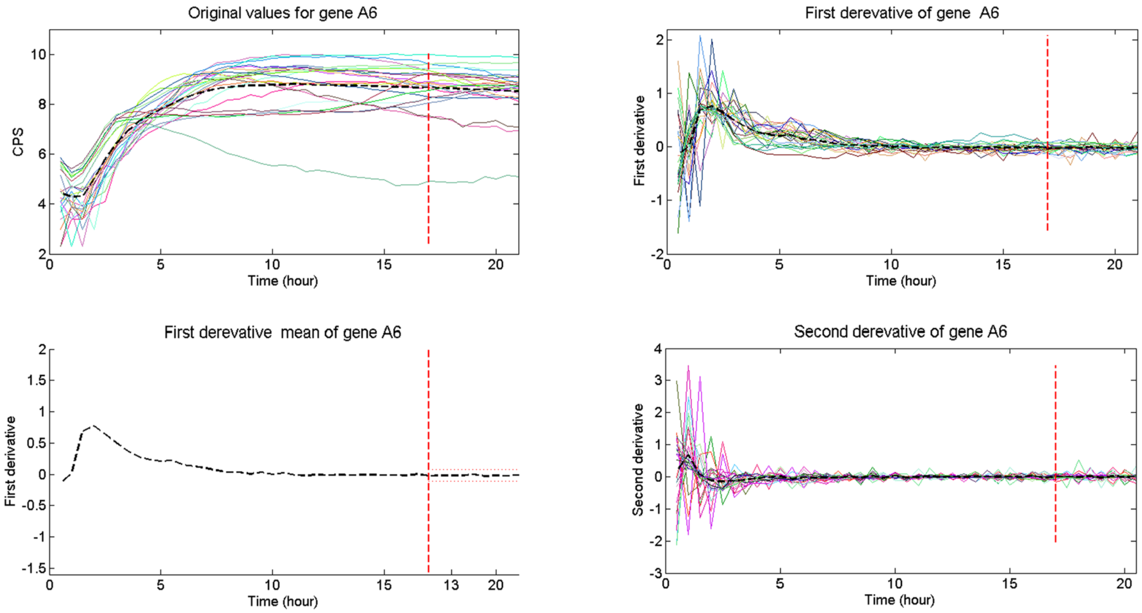

Figure 5. Sample visualised results for gene PA5283.

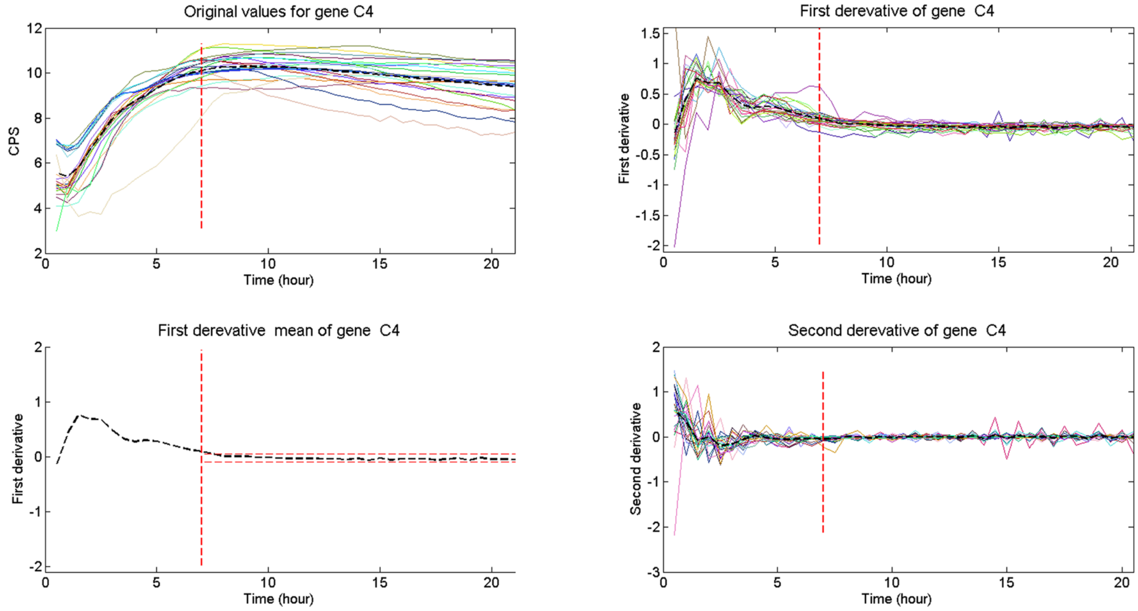

Figure 6. Sample visualised results for gene PA0287.

of τ values, we can find that some of gene pairs such as [*PA5283, PA2975*], [*PA4991, PA0287*], [*PA2997, PA3771*], and etc. are quite significantly different. In this particular case the genes with two hours difference in estimated τ values are considered to have different mean functions. However, when this difference for a pair of genes is less than or equal two hours, the statistical test is performed to compare the mean functions for two genes at 5% significance level using the method, which was introduced by Nishiyama *et al.* (2013). The last column indicates the genes with equivalent mean functions to the specified gene in each row. Eventually, we can find the mean function for *PA0649*, *PA2975*, *PA1748*, and *PA1841* are not significantly different. Also, we have the same situation for *PA3771*, *PA0894*, *PA0573* and *PA3902*.

In particular, we present the results in Figures 5 and 6 for the analysis of genes *PA5283* and *PA0287*, respectively, by using Algorithm 5. Each figure contains four sub-plots. The left top sub-plot presents the original observed gene expressions for all conditions and the vertical dash line shows the approximate value of τ. Also, the thick black dot line represents the sample mean for the specific gene. The right sides sub-plots on first row and second row indicate the same thing as aforementioned plot but for sample first-order derivatives and second-order derivatives, receptively. The left bottom subplot indicates the estimated first-order derivative mean and the confidence band of gene expression. From these plots one can see that genes *PA5283* and *PA0287* have the significantly different threshold points, which shows that genes *PA5383* and *PA0287* exhibit the different behaviors. Gene PA0287 becomes stable only after 6.5 hours and gene PA5283 needs a little longer time 16.5 hours to keep constant.

5. CONCLUDING REMARKS

Finding genes with similar behaviors and classifying them into the same category using statistical methods are the main concern on recent gene expression studies. Since we have high-dimensional data sets for analysing the properties of gene expressions, the methods for dimensional deduction are utilized significantly in this case. The existing methods such as PCA, KPCA, and etc. deal with the reduction of dimensionality using a combination of observations in the data sets. However, the current paper is proposing a strategy to eliminate the observations for some time points with relative change rate close to zero, instead of considering the combination of observations for different time points. To apply this strategy, three methods and their corresponding algorithms are proposed. The first method is based on Hotelling's T^2 which is one of the classical testing methods for multivariate data sets. The second method is similar to the first model. But instead of using T^2 statistic, the test statistic based on high-dimensional data is used. As for the third method, the method based on empirical distribution for the sample first-order and second-order derivatives is proposed to find the threshold time point. As indicated from the simulation study, the method based on the empirical distribution of sample second-order derivative can detect the values of threshold time points more accurately for the gene expressions. Since the multiple threshold would be more common than single threshold homeostasis, we should develop the method to detect the multiple thresholds and apply this method to the high throughput gene expression data. In fact, multiple threshold points could be detected using the modified method from the current algorithms. What we need to do is to slightly change each algorithm and use it to detect multiple thresholds.

Usually, after the number of observations has been reduced, we are able to compare the mean functions together. For this purpose, several existing statistical methods could be available under two situations. First situation is that covariance matrices for two genes are equal and second one is covariance matrices are different from each other. In the second case we may confront Behrens-Fisher problem. Overall, the best case to use the explained strategies for reducing the number of observations is when the number of time points are close to the number of conditions. The proposed method can make the gene expression data appropriate for using classical methods for classification of gene expressions. Also, the threshold points can be used to distinguish among the gene expressions. We have done some research on the classification of gene expressions based on the threshold time points. In general, if two genes have the significantly different threshold points, one can easily conclude that these two gene expressions have obviously different behaviours. If the threshold points of two genes are close each other, we may further consider testing the equality of mean functions for these two gene expressions.

Further, from Section 4, one can see there exists a peak point for the relative change rate of each gene expression. This point is the fastest increasing point for gene expression. We would like to know if this peak time point has its own meaning in the gene research and if this point can help to explain the behavior of gene expression. The potential statistical question is how to detect this peak point. Furthermore, the time interval between the peak point and threshold point has its own interest because this interval shows the time period that the gene expression change from the fastest increasing point to the point with zero change rate.

Finally, for the convenience of users, we upload the Matlab scripts for the detection of threshold point used in the simulation section and example section to GitHub and everyone can have access to these files using the following link: **https://github.com/jahromi/DTPGE**.

Also, users can contact the developer from GitHub directly.

ACKNOWLEDGMENTS

Authors are very grateful to associate editor and one referee for their careful reading and constructive comments. The research of D. Deng is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

APPENDIX. PROOF OF THEOREM 1

Consider $V_j = \sum_{\{r:t'_r>\tau\}} I(Z_j(t'_r) \le 0)$ for $j = 1, ..., c$. Since $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_c$ are mutually independent and identically distributed, $V_1, ..., V_c$ are also mutually independent and identically distributed. Based on central limit theorem, the distribution of $\frac{\bar{V}-E(\bar{V})}{\sqrt{\operatorname{var}(\bar{V})}}$ converges to the standard normal distribution.

Meanwhile, we have

$$\begin{aligned}
\text{(A.1)}\qquad \sum_{\{r:t'_r>\tau\}} \widehat{F}_{Z(t'_r)}(0) &= \sum_{\{r:t'_r>\tau\}} \sum_{i=1}^{c} \frac{1}{c} I(Z_j(t'_r) \le 0) \\
&= \frac{1}{c} \sum_{i=1}^{c} \sum_{\{r:t'_r>\tau\}} I(Z_j(t'_r) \le 0) = \frac{1}{c} \sum_{i=1}^{c} V_j = \bar{V}
\end{aligned}$$

Also,

$$\text{(A.2)}\qquad E(\bar{V}) = E\left(\sum_{\{r:t'_r>\tau\}} \widehat{F}_{Z(t'_r)}(0)\right)$$

$$= \sum_{\{r:t'_r>\tau\}} \frac{1}{2} = \frac{1}{2}\sum_{r=1}^{p} I\{t'_r > \tau\} = \frac{c_\tau}{2}$$

and

$$\operatorname{var}(\bar{Z}) = \operatorname{var}\left(\sum_{\{r:t'_r>\tau\}} \widehat{F}_{Z(t'_r)}(0)\right)$$

$$= \operatorname{cov}\left(\sum_{\{r:t'_r>\tau\}} \widehat{F}_{Z(t'_r)}(0), \sum_{\{r:t'_r>\tau\}} \widehat{F}_{Z(t'_r)}(0)\right)$$

$$= \sum_{\{r:t'_r>\tau\}} \operatorname{var}(\widehat{F}_{Z(t'_r)}(0)) + \sum_{(r,s)\in A} \operatorname{cov}(\widehat{F}_{Z(t'_r)}(0), \widehat{F}_{Z(t'_s)}(0))$$

$$= \sum_{\{r:t'_r>\tau\}} F_{Z(t'_r)}(0)(1 - F_{Z(t'_r)}(0))$$

$$+ \sum_{(r,s)\in A} \operatorname{cov}(\widehat{F}_{Z(t'_r)}(0), \widehat{F}_{Z(t'_s)}(0)).$$

where $A = \{r, s | r \neq s, t'_r > \tau \; and \; t'_s > \tau\}$. Now

$$\operatorname{cov}(\widehat{F}_{Z(t'_r)}(0), \widehat{F}_{Z(t'_s)}(0))$$

$$= \frac{1}{c^2}\sum_{i=1}^{c}\sum_{j=1}^{c} E(I(Z_i(t'_r) < 0)I(Z_j(t'_s) < 0)))$$

$$- \frac{1}{c^2}\sum_{i=1}^{c}\sum_{j=1}^{c} E(I(Z(t'_r) < 0))E(I(Z(t'_s) < 0))$$

$$= \frac{1}{c}\frac{(c-1)}{4} + \frac{1}{c}\int_{-\infty}^{0}\int_{-\infty}^{0} f_{Z(t'_r),Z(t'_s)}(x,y)dxdy - \frac{1}{4}$$

$$= -\frac{1}{4c} + \frac{1}{c}\int_{-\infty}^{0}\int_{-\infty}^{0} f_{Z(t'_r),Z(t'_s)}(x,y)dxdy$$

Therefore,

(A.3)

$$\operatorname{var}(\bar{Z}) = \operatorname{var}\left(\sum_{\{r:t'_r\geq\tau\}} \widehat{F}_{Z_j(t'_r)}(0)\right)$$

$$= \frac{c_\tau}{4c} + \frac{1}{c}\sum_{(r,s)\in A} [F_{Z(t'_r),Z(t'_s)}(0,0) - \frac{1}{4}]$$

$$= \frac{1}{c}\left[\frac{1}{4}(2c_\tau - c_\tau^2) + \sum_{(r,s)\in A} F_{Z(t'_r),Z(t'_s)}(0,0)\right]$$

As we can see Eq. (A.1) indicates the expansion of $\bar{V}$ and Eq. (A.2), (A.3) imply that

$$\frac{\bar{V} - E(\bar{V})}{\sqrt{\operatorname{var}(\bar{V})}} = \frac{\sum_{\{r:t'_r\geq\tau\}} \widehat{F}_{Z(t'_r)}(0) - \frac{c_\tau}{2}}{\sqrt{\frac{1}{c}[\frac{1}{4}(2c_\tau - c_\tau^2) + \sum_{(r,s)\in A} F_{Z(t'_r),Z(t'_s)}(0,0)]}}$$

Hence Eq. (8) converges to standard normal variable in distribution as $c \to \infty$.

Received 14 March 2016

REFERENCES

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley-Interscience. MR1990662

Bjarnason, J., Southward, C. M., and Surette, M. G. (2003). Genomic profiling of iron-responsive genes in salmonella enterica serovar typhimurium by high-throughput screening of a random promoter library. *Journal of Bacteriology*, **185**(16), 4973–4982.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L. Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, **2**(1), 65–73.

Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S., and Tainsky, M. A. (2003). Noise sampling method: an anova approach allowing robust selection of differentially regulated genes measured by dna microarrays. *Bioinformatics*, **19**(11), 1348–1359.

Duan, K., Ware, T., McCullough, W. M., Surette, M. G., and Song, J. (2012). Comprehensive analysis of gene-environmental interactions with temporal gene expression profiles in pseudomonas aeruginosa. *PloS One*, **7**(3), doi:10.1371/journal.pone.0035993.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25), 14863–14868.

Fang, H., Deng, D., Tian, G., Shen, L., Duan, K., and Song, J. (2012). Analysis for temporal gene expressions under multiple biological conditions. *Statistics in Biosciences*, **4**(2), 282–299.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, **43**. CRC Press. MR1082147

Li, H., Luan, Y., Hong, F., and Li, Y. (2002). Statistical methods for analysis of time course gene expression data. *Frontiers in Bioscience: A Journal and Virtual Library*, **7**, 90–98.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, **42**, 1897–1908. MR0309251

Nishiyama, T., Hyodo, M., Seo, T., and Pavlenko, T. (2013). Testing linear hypotheses of mean vectors for high-dimension data with unequal covariance matrices. *Journal of Statistical Planning and Inference*, **143**(11), 1898–1911. MR3095080

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115. MR0088850

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527. MR0072412

Schölkopf, B., Smola, A., and Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**(5), 1299–1319.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9**(12), 3273–3297.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**(9), 763–774.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, **94**(1), 19–35. MR2367824

Dianliang Deng
Department of Mathematics and Statistics
University of Regina
Sask., S4S 0A2
Canada
E-mail address: deng@uregina.ca

Hong-Bin Fang
Department of Biostatistics
Bioinformatics and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
USA
E-mail address: hf183@georgetown.edu

Kian Razeghi Jahromi
Department of Mathematics and Statistics
University of Regina
Sask., S4S 0A2
Canada
E-mail address: jahromi@outlook.com

Jiuzhou Song
Department of Animal and Avian Sciences
University of Maryland
College Park, MD 20742-2311
USA
E-mail address: Songj88@umd.edu

Ming Tan
Department of Biostatistics
Bioinformatics and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
USA
E-mail address: ttan34@georgetown.edu