The Annals of Applied Statistics
2021, Vol. 15, No. 1, 509–526
https://doi.org/10.1214/20-AOAS1403

AN EMPIRICAL BAYES CHANGE-POINT MODEL FOR TRANSCRIPTOME TIME-COURSE DATA

BY TIAN TIAN[1,*], RUIHUA CHENG[2] AND ZHI WEI[1,†]

[1]*Department of Computer Science, New Jersey Institute of Technology,* [*]*tt72@njit.edu;* [†]*zhiwei@njit.edu*
[2]*Big Data Statistics Research Center, Tianjin University of Finance and Economics, r.cheng665@gmail.com*

Time-course experiments are commonly conducted to capture temporal changes. It is generally of interest to detect if any changes happen over time, which we define as a detection problem. If there is a change, it is informative to know when the change is, which we define as an identification problem. It is often desired to control Type I error rate at a nominal level while applying a testing procedure to detect or identify these changes. Quite a few analytic methods have been proposed. Most existing methods aim to solve either the detection problem or, more recently, the identification problem. Here, we propose to solve these two problems using a unified multiple-testing framework built upon an empirical Bayes change-point model. Our model provides a flexible framework that can account for sophisticated temporal gene expression patterns. We show that our testing procedure is valid and asymptotically optimal in the sense of rejecting the maximum number of null hypotheses, while the Bayesian false discovery rate (FDR) can be controlled at a predefined nominal level. Simulation studies and application to real transcriptome time-course data illustrate that our proposed model is a flexible and powerful method to capture various temporal patterns in analysis of time-course data.

1. Introduction. Gene expression temporal patterns can reveal regulation networks (Bar-Joseph, Gitter and Simon (2012), Calvano et al. (2005)). The transcriptome time-course experiment measures thousands of gene expression levels during relatively few time points. In such experiments, the number of time points is usually small, such as three to 20. Many such experiments are not replicated due to cost and other limitations, and, when replicates are available, the numbers are typically also small (Tai and Speed (2005)). Technologies for measuring genes include microarrays and RNA-seq. These technologies have been used to capture the temporal gene expression fluctuations in many biological processes, such as regulation of development (Arbeitman et al. (2002)), immune responses (Calvano et al. (2005)) and tissue inflammation programs (Tian, Nowak and Brasier (2005)).

Most transcriptome time-course experiments aim to detect variably expressed genes. Such genes are often related to the biological processes motivating the experiments, for example, the genes related to the immune response. The statistical challenge of transcriptome time-course experiment arises from the fact that the number of time points is relatively small, and the number of genes is usually large (e.g., about 20,000 genes in human). Methods specially designed to deal with this situation are developed, such as ANOVA (classical ANOVA or modified model) (Diggle et al. (2002)), regression approach (Zhao, Prentice and Breeden (2001)), contrasts (Lönnstedt et al. (2005)) and hidden Markov models (Sun and Wei (2011)). The Gaussian process (GP) regression has proven to be more flexible and powerful compared to the linear or spline regression models and has been widely applied for the analysis of time series. In fact, spline regression can be considered as one special case of GP (Kimeldorf and Wahba (1970)). However, it has two main issues: typically, GP needs many observations to obtain acceptable performance (usually tens to hundreds); GP regression time-course

Received April 2019; revised September 2020.
Key words and phrases. Transcriptome, time series, empirical Bayes, change-point model.

methods use a log-likelihood ratio to rank the differentially expressed genes which makes it impossible to report the p-values and to control the false positive rate at the nominal level (Kalaitzis and Lawrence (2011), Yang et al. (2016)). Usually, there are thousands of genes in the context of the transcriptome time-course experiments, which motivates the use of the empirical Bayes approach to make inference (Efron et al. (2001), Tai and Speed (2006), Kendziorski et al. (2003)). The empirical Bayes approach can pool and exploit information across many genes under investigation (Efron (2010)), so it is a good choice to develop a statistical method for the analysis of time-course gene expression data. Empirical Bayes has been widely used in the analysis of other types of genomic data, including the differential expression gene (DEG) analysis (Robinson, McCarthy and Smyth (2010), Smyth (2005), Love, Huber and Anders (2014)) and genomic variants identification (Zhao, Wang and Wei (2013)).

One characteristic of transcriptome time-course experiments is that, if the expression level of a gene changes at a given time point, it is very likely to return to its initial expression level later if we could measure it for longer intervals; that is, it will only appear to remain at that level if the measurement interval is short. The assumption that expression levels will return to normal is reasonable given the need to maintain homeostasis in all organisms. The change-point model assumes a sequence of data can be broken into segments, and observations from each segment follow a same statistical distribution. In the context of transcriptome time-course experiments, because of a relatively small number of time points measured, this experiment may be unable to capture the continuously changing process of expression levels. The change-point model can be a good choice to describe these observations. Considering the change and the recovery of temporal gene expression levels in this context, we build a change-point model with two change points, one for change and the other one for recovery. The change-point model has been proved to be a useful approach to analyze next generation sequencing data (Wang, Wei and Li (2014), Zhang and Wei (2016)), and this method can answer two kinds of questions simultaneously: detection, to detect which genes' expression levels have been changed; identification, to identify when the change happens. Most previous statistical methods focus on the first detection question, such as (Smyth (2005), Tai and Speed (2006), Kalaitzis and Lawrence (2011)), but can't answer the second identification question. Yang et al. proposed a GP regression method for transcriptome time-course data that was able to answer the two questions, but the method essentially needed two conditions to be compared, and FDR levels could not be controlled (Yang et al. (2016)). Here, we present a method based on the change-point model that can deal with both two questions on very few time points and control the FDR at a nominal level. Our model is flexible and applicable to more complicated applications. Meanwhile, our model can deal with both microarray data and RNA-Seq data with a proper normalization technique.

2. Model.

2.1. *Notations of the transcriptome time-course change-point problem.* In time-course experiments, the expression levels of genes are measured longitudinally. For each gene G, $G = G_1, G_2, \ldots, G_N$, we have N time-series sequences of expression levels. By following the previous change-point model (Barry and Hartigan (1993), Denison et al. (2002), Xuan and Murphy (2007)), we assume one gene could have two possible patterns, expressed constantly or expressed differentially. Furthermore, we assume the genes that are expressed differentially could have one or two change points: for genes with one change point, which means gene expression levels change at one time point and remain at that level until the end of experiment; for genes with two change points, expression levels change at one time point and then are restored later (Figure 1).

We denote expression level of gene G_i at time point j $(j = 1, 2, \ldots, T)$ as x_{ij}, so that the time sequence of G_i can be written as $X_i = x_{i1}, x_{i2}, \ldots, x_{iT}$ for T observations. With

the proposed change-point model, the sequence can be considered as one, two or three homogeneous sequences, $\Pi = \Pi_1, \ldots, \Pi_Q$, with $Q = 1, 2, 3$, namely, represent situations that the gene has 0, 1 or 2 change points. We assume that measurements within the same homogeneous sequence share a common mean expression level μ_q and arise independently and identically from an observation component $f(X_{i,\Pi_q}|\mu_q)$. We consider μ_q as arising from a common genome-wide distribution $\pi(\mu_q)$ which represents fluctuations in mean expression levels among genes. Since μ_q is latent and not our primary interest, we integrate it away and have

$$f(X_{i,\Pi_q}) = \int \Big(\prod_{t \in \Pi_q} f(X_{i,t}|\mu_q) \Big) \pi(\mu_q)\, d\mu_q.$$

We will specify $f(X_{i,t}|\mu_q)$ and $\pi(\mu_q)$ in the next subsection. Since one gene can have two change points at most, we introduce a variable $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$ to denote the change-point pattern for gene G_i, where $\rho_{i1} \in \{0, 1, 2, \ldots, T-1\}$ and $\rho_{i2} \in \{0, 1, 2, \ldots, T-1\}$ indicate the positions of change points for gene G_i. To specify, if G_i doesn't have any change point, then $\rho_{i1}, \rho_{i2} = 0$, and the whole sequence of G_i will be homogeneous with one latent mean μ_1 ($Q = 1$ and $\Pi_1 = \{1, \ldots, T\}$). If G_i only has one change point, then we let $\rho_{i1} = 0$ and $\rho_{i2} = \tau_i$, $\tau_i \in \{1, 2, \ldots, T-1\}$, which implies expression levels of G_i before time point τ_i ($X_{i,1:\tau_i} = x_{i1}, x_{i2}, \ldots, x_{i\tau_i}$) follow a homogeneous sequence with one latent mean μ_1 and expression levels after τ_i ($X_{i,(\tau_i+1):T} = x_{i(\tau_i+1)}, \ldots, x_{iT}$) follow another homogeneous sequence with a different latent mean μ_2. Thus, $Q = 2$, $\Pi_1 = \{1, \ldots, \tau_i\}$ and $\Pi_2 = \{\tau_i + 1, \ldots, T\}$. Meanwhile, if G_i has two change points, we let $\rho_{i1} = \tau_{i1}$ and $\rho_{i2} = \tau_{i2}$, with $\tau_{i1}, \tau_{i2} \in \{1, 2, \ldots, T-1\}$ and $\tau_{i1} < \tau_{i2}$, represent gene G_i having change points at τ_{i1} and τ_{i2} time points. These two change points divide the gene-expression sequence into $Q = 3$ homogeneous segments, $\Pi_1 = \{1, \ldots, \tau_{i1}\}$, $\Pi_2 = \{\tau_{i1} + 1, \ldots, \tau_{i2}\}$ and $\Pi_3 = \{\tau_{i2} + 1, \ldots, T\}$, with latent means μ_1, μ_2 and μ_3, respectively. In biology, most genes will restore to their original expression levels. Having a new μ_3 will help to characterize some genes more precisely, at the price of a more complex model, while for many other genes restoring to original levels, it may not be necessary. Therefore, we let $\mu_3 = \mu_1$ and merge Π_3 with Π_1. Namely, before time point τ_{i1} and after time point τ_{i2}, expression levels of G_i ($X_{i,1:\tau_{i1} \cup (\tau_{i2}+1):T} = x_{i1}, x_{i2}, \ldots, x_{i\tau_{i1}}, x_{i(\tau_{i2}+1)}, \ldots, x_{iT}$) follow a homogeneous sequence; expression levels of G_i after time point τ_{i1} and before time point τ_{i2} ($X_{i,(\tau_{i1}+1):\tau_{i2}} = x_{i(\tau_{i1}+1)}, \ldots, x_{i\tau_{i2}}$) follow another homogeneous sequence. We think this setting is more representative of biological scenarios and provides a reasonable trade-off between generality and efficiency gain.

2.2. *Normal Normal-Gamma model.* It is intuitive to characterize relative or absolute gene expression levels by a normal distribution, especially when considering log scaled data. Therefore, we consider a normal distribution for defining $f(X_{i,t}|\mu_q)$. As illustrated in Figure 1, if gene G_i expresses constantly, then observed expression levels across all time points follow one normal distribution with a common mean μ_{i1}, namely, $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$. If G_i has change points, based on the number and positions of change points of G_i, we can characterize expression levels by two different normal distributions: if G_i has one change point at time point τ_i, then expression levels before τ_i follow one normal distribution $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$ with mean μ_{i1} and follow another normal distribution $\mathcal{N}(\mu_{i2}, \sigma_{i2}^2)$ with a different mean μ_{i2} after time point τ_i; if G_i has two change points at time points τ_{i1} and τ_{i2}, then we can govern expression levels before τ_{i1} (include τ_{i1}) and after τ_{i2} by $\mathcal{N}(\mu_{i1}, \sigma_{i1}^2)$ and expression levels from $\tau_{i1} + 1$ to τ_{i2} by $\mathcal{N}(\mu_{i2}, \sigma_{i2}^2)$. Formally:

1. No change point, then $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$, where $\rho_{i1} = 0$, $\rho_{i2} = 0$

$$x_{ij}|\boldsymbol{\rho}_i \sim \mathcal{N}(\mu_{i1}, \sigma_{i1}^2); \tag{2.1}$$

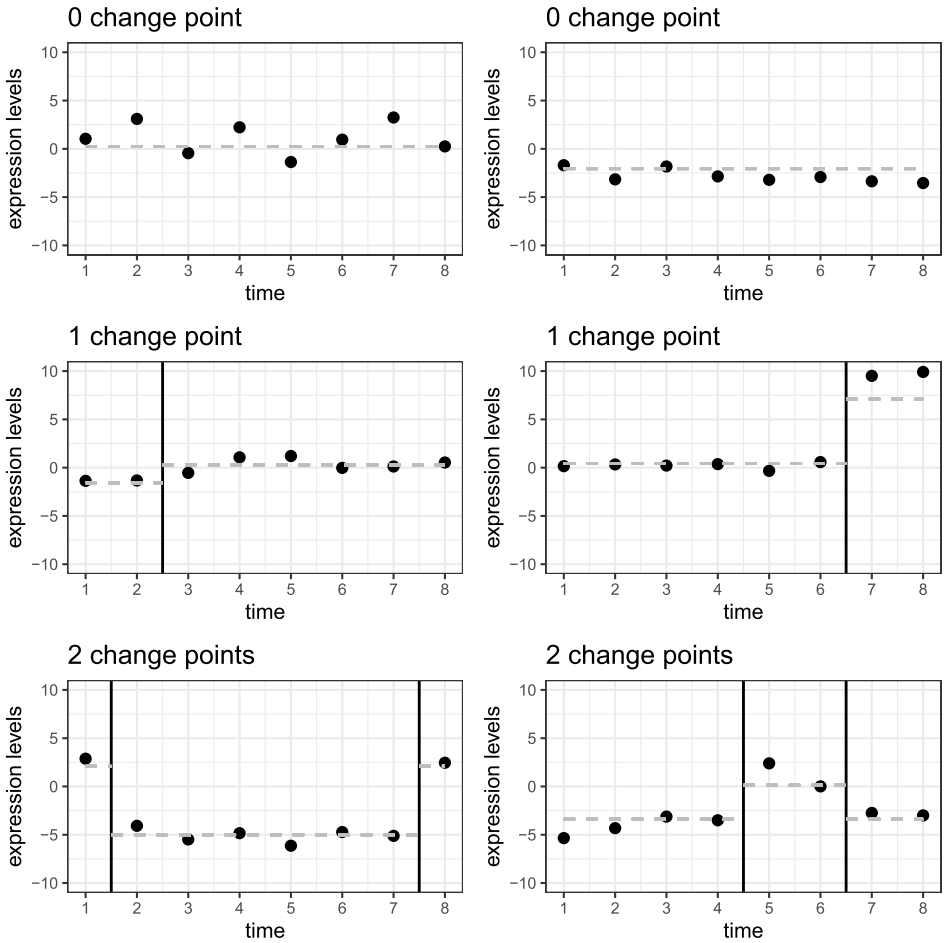

FIG. 1. *Illustration of different time series patterns. The horizontal dashed lines are expected expression levels and the vertical lines are the positions of change points. The first row are examples of zero change point, in which expression levels remain constant across time points. The second row are examples of one change point, whose expression levels have a expected mean at the beginning and then change to another expected mean at a time point.The third row are examples of two change points, expression levels change at one time points and restore later. In our notations, gene G_i without change point corresponds to $\rho_{i1} = 0$ and $\rho_{i2} = 0$; having one change point corresponds to $\rho_{i1} = 0$ and $\rho_{i2} = \tau_i$ $(\tau_i = 1, 2, \ldots, T-1)$; having two change points corresponds to $\rho_{i1} = \tau_{i1}$, $\rho_{i2} = \tau_{i2}$, and $\tau_{i1}, \tau_{i2} = 1, 2, \ldots, T-1$ $(\tau_{i1} < \tau_{i2})$.*

2. One change point at time point τ_i $(\tau_i = 1, 2, \ldots, T-1)$, then $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$, where $\rho_{i1} = 0$, $\rho_{i2} = \tau_i$

$$x_{ij}|\boldsymbol{\rho}_i \sim \begin{cases} \mathcal{N}(\mu_{i1}, \sigma_{i1}^2) & j \le \tau_i, \\ \mathcal{N}(\mu_{i2}, \sigma_{i2}^2) & j \ge \tau_i + 1; \end{cases} \tag{2.2}$$

3. Two change points at time points τ_{i1} and τ_{i2} $(\tau_{i1}, \tau_{i2} = 1, 2, \ldots, T-1$ and $\tau_{i1} < \tau_{i2})$, then $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$, where $\rho_{i1} = \tau_{i1}$, $\rho_{i2} = \tau_{i2}$

$$x_{ij}|\boldsymbol{\rho}_i \sim \begin{cases} \mathcal{N}(\mu_{i1}, \sigma_{i1}^2) & j \le \tau_{i1} \text{ or } j \ge \tau_{i2} + 1, \\ \mathcal{N}(\mu_{i2}, \sigma_{i2}^2) & \tau_{i1} + 1 \le j \le \tau_{i2}. \end{cases} \tag{2.3}$$

We further assume that latent means μ_{iq} and variance σ_{iq} $(q = 1, 2)$ for G_is follow a Normal-Gamma distribution, which is a conjugate prior of the normal distribution, to characterize different latent means and variances for different genes and sequences. We name this model

as the hierarchical Normal Normal-Gamma model (NNG model for short), because we use the hierarchical Normal and Normal-Gamma distributions to characterize gene expression levels. Under the NNG model we assume that the observed individual gene expressions are independent, given its mean, while the means of all individual genes follow a common distribution. The marginal gene expression levels can be dependent under such a hierarchical model. Similar hierarchical models have been used in previous studies (Kendziorski et al. (2003), Newton et al. (2001), Yuan and Kendziorski (2006)) for characterizing microarray gene expression data.

Formally, if we define the precision $\lambda_{iq} = \sigma_{iq}^{-2}$, then

$$\mu_{iq}, \lambda_{iq} \sim \mathrm{NG}(\mu, \lambda|\nu_0, \kappa_0, \alpha_0, \beta_0) = \mathcal{N}\big(\mu|\nu_0, (\kappa_0\lambda)^{-1}\big)\Gamma(\lambda|\alpha_0, \beta_0). \tag{2.4}$$

We use an empirical Bayes approach to estimate hyperparameters ν_0, κ_0, α_0 and β_0 from data which estimate parameters by maximizing the likelihood. From the hierarchical Normal Normal-Gamma model, we have

$$\begin{aligned} f(x_{ij}|\mu_{iq}, \lambda_{iq}) &= \sqrt{\frac{\lambda_{iq}}{2\pi}} \exp\Big(-\frac{\lambda_{iq}}{2}(x_{ij} - \mu_{iq})^2\Big), \\ f(\mu_{iq}, \lambda_{iq}|\nu_0, \kappa_0, \alpha_0, \beta_0) &= \frac{1}{Z_{\mathrm{NG}}} \lambda_{iq}^{\alpha_0 - \frac{1}{2}} \exp\Big(-\frac{\lambda_{iq}}{2}\big(\kappa_0(\mu_{iq} - \nu_0)^2 + 2\beta_0\big)\Big), \end{aligned} \tag{2.5}$$

where $Z_{\mathrm{NG}} = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}(\frac{2\pi}{\kappa_0})^{\frac{1}{2}}$.We assume that gene-expression levels across time points are independent and identically distributed (i.i.d.), so the marginal likelihood of a homogeneous sequence $X_{i,m:n} = x_{im}, x_{i(m+1)}, \ldots, x_{in}$ can be calculated as the product of likelihoods at every time point

$$\begin{aligned} &\ell(X_{i,m:n}|\nu_0, \kappa_0, \alpha_0, \beta_0) \\ &\quad = \int_{-\infty}^{\infty}\int_{0}^{\infty} \Big\{\prod_{j=m}^{n} \mathcal{N}(x_{ij}|\mu, \lambda)\Big\} \mathcal{N}\big(\mu|\nu_0, (\kappa_0\lambda)^{-1}\big)\Gamma(\lambda|\alpha_0, \beta_0)\, d\mu\, d\lambda. \end{aligned} \tag{2.6}$$

By applying calculus results of the normal distribution (Murphy (2007)), we can integrate out latent variables μ and λ,

$$\ell_0(X_{i,m:n}|\nu_0, \kappa_0, \alpha_0, \beta_0) = \frac{\Gamma(\alpha_s)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_s^{\alpha_s}} \Big(\frac{\kappa_0}{\kappa_s}\Big)^{\frac{1}{2}} (2\pi)^{-\frac{s}{2}}, \tag{2.7}$$

where

$$\begin{aligned} \kappa_s &= \kappa_0 + s, \\ \alpha_s &= \alpha_0 + s/2, \\ \beta_s &= \beta_0 + \frac{1}{2}\sum_{i=m}^{n}(x_i - \bar{x}_{i,m:n})^2 + \frac{\kappa_0 s(\bar{x}_{i,m:n} - \nu_0)^2}{2(\kappa_0 + s)}. \end{aligned}$$

Here, s is the length of homogeneous sequence: $s = n - m + 1$, and $\bar{x}_{i,m:n}$ is the mean of sequence $X_{i,m:n}$: $\bar{x}_{i,m:n} = (\sum_{j=m}^{n} x_{ij})/s$.

Now, it is straightforward to calculate the likelihood of genes with zero, one or two change points as following:

1. No change point, $\boldsymbol{\rho}_i = (0, 0)$, the expression levels of G_i are one homogeneous sequence:

$$\ell(X_{i,1:T}|\nu_0, \kappa_0, \alpha_0, \beta_0, \boldsymbol{\rho}_i) = \ell_0(X_{i,1:T}|\nu_0, \kappa_0, \alpha_0, \beta_0);$$

2. One change point at time point τ_i, then $\boldsymbol{\rho}_i = (0, \tau_i)$, the likelihood is the product of two homogeneous sequences

$$\ell(X_{i,1:T}|\nu_0, \kappa_0, \alpha_0, \beta_0, \boldsymbol{\rho}_i) = \ell_0(X_{i,1:\tau_i}|\nu_0, \kappa_0, \alpha_0, \beta_0) \times \ell_0(X_{i,(\tau_i+1):T}|\nu_0, \kappa_0, \alpha_0, \beta_0);$$

3. Two change points at time points τ_{i1} and τ_{i2}, then $\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2})$, the likelihood is the product of two homogeneous sequences, similar to the one change point case,

$$\begin{aligned}\ell(X_{i,1:T}|\nu_0, \kappa_0, \alpha_0, \beta_0, \boldsymbol{\rho}_i) = {} & \ell_0(X_{i,1:\tau_{i1}\cup(\tau_{i2}+1):T}|\nu_0, \kappa_0, \alpha_0, \beta_0) \\ & \times \ell_0(X_{i,(\tau_{i1}+1):\tau_{i2}}|\nu_0, \kappa_0, \alpha_0, \beta_0).\end{aligned}$$

On a null hypothesis there is no change across all of the time points. All the observed data values share the same mean expression value, and the data for a given gene G_i arise from a joint probability density function (pdf) $f_0(X_i)$. Alternatively, we consider different change-point patterns and denote the joint pdf as $f_k(X_i)$, $k > 0$. We don't know a priori for the underlying pattern of gene G_i and introduce discrete mixing parameters p_k to denote the unknown probabilities of expression pattern k. So, we can have the marginal distribution of the data in a mixture of the form

$$f(X_i) = p_0 f_0(X_i) + \sum_{k=1}^{K} p_k f_k(X_i). \tag{2.8}$$

In our change-point model the number of all possible different change-point patterns is $K = \binom{T}{2}$ for a T time-points experiment. By default, we assume each pattern having an equal prior probability of happening. Suppose that the prior probability for each gene G_i to have change points is P; then, symbolically,

$$\Pr(\boldsymbol{\rho}_i; P) = \begin{cases} 1 - P & (\rho_{i1}, \rho_{i2}) = (0, 0), \\ \dfrac{P}{\binom{T}{2}} & \text{at least one of } (\rho_{i1}, \rho_{i2}) \text{ is not } 0. \end{cases} \tag{2.9}$$

Thus, $p_0 = 1 - P$ and $p_k = P/K$ for $k = 1, \ldots, K$. We use a single parameter P to distinguish nonnulls from nulls but do not further differentiate nonnull patterns. This is a reasonable setting because we do not want to be biased toward any nonnull pattern a priori. In contrast, if needed we may allow each pattern to have its own parameter p_k, subject to $\sum_{k=0}^{K} p_k = 1$. This is a more complex model with K more parameters. It may be appropriate if we expect all patterns will involve a significant number of genes.

2.3. *Hyperparameter estimator.* Empirical Bayes is an approach of statistical inference that combines Bayesian and frequentist reasoning, in which the prior distribution is estimated from data (Efron (2010)). It estimates hyperparameters by the approach of maximum marginal likelihood. In our Normal Normal-Gamma model, formally, we denote Φ as the set of parameters with $\Phi = (P, \nu_0, \kappa_0, \alpha_0, \beta_0)$.The locations of change points ρ_{i1} and ρ_{i2} are latent, and we sum them out. Then, the maximum marginal likelihood estimation of Φ, applied to total N genes, can be written as

$$\hat{\Phi} = \arg\max_{\Phi} \log\left(\prod_{i=1}^{N} f(X_i; \Phi)\right) = \arg\max_{\Phi} \log\left(\prod_{i=1}^{N} \sum_{\boldsymbol{\rho}_i} \Pr(\boldsymbol{\rho}_i; \Phi)\ell(X_i|\boldsymbol{\rho}_i; \Phi)\right).$$

Here, $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$. $(\rho_{i1}, \rho_{i2}) = (0, 0)$, or $(\rho_{i1}, \rho_{i2}) = 0, 1, \ldots, T - 1$ and $\rho_{i1} < \rho_{i2}$.

We apply the Adam optimization algorithm (Kingma and Ba (2014)) to estimate parameters Φ with some reasonable constraints (e.g., $0 \le P \le 1$). We implement this procedure using the Tensorflow R package (Abadi et al. (2016), Allaire and Tang (2020)).

2.4. *Empirical Bayes testing and decision procedure to detect change points in time series data*. In the context of transcriptome time-course experiments, we have two questions aimed to answer:

Q1: *Detection*, which genes' expression levels have changed across time series? So, we need to detect which genes have change points.

Q2: *Identification*: If any genes with a change, at which time point(s) have the expression levels changed? Namely, we need to identify the positions of change points.

For Q1, we detect the existence of change points for a gene. For Q2, we aim to find the accurate location of change points; that is, we need not only to detect the existence of change points but also to identify the exact locations of change points. Given the estimated parameters $\hat{\Phi}$, we can calculate the posterior probability of each change-point pattern for gene G_i,

$$\widehat{\Pr}(\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2})|X_i) = \frac{\ell(X_i|\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2}), \hat{\Phi}) \times \Pr(\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2})|\hat{\Phi})}{\sum_{\tau_{i1}, \tau_{i2}} \ell(X_i|\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2}), \hat{\Phi}) \times \Pr(\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2})|\hat{\Phi})}.$$

Here, $\tau_{i1}, \tau_{i2} = 0$, or $\tau_{i1}, \tau_{i2} = 0, 1, \ldots, T-1$ and $\tau_{i1} < \tau_{i2}$.

We let

$$\begin{aligned}\widehat{\pi_{i0}} &= \widehat{\Pr}(\boldsymbol{\rho}_i = (0,0)|X_i),\\ \widehat{\pi_i^*} &= \max\{\widehat{\Pr}(\boldsymbol{\rho}_i = (\tau_{i1}, \tau_{i2})|X_i\}\\ &\qquad \tau_{i1}, \tau_{i2} = 0, 1, \ldots, T-1 \text{ and } \tau_{i1} < \tau_{i2}.\end{aligned}$$

Here, $\widehat{\pi_{i0}}$ is the probability of G_i and has no change points; and $\widehat{\pi_i^*}$ means the probability of the most likely change-point pattern for G_i, under the condition of estimated hyperparameters.

For both two questions we will develop a procedure for controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)) at a nominal level α and find as many genes as possible with change points. Symbolically, our empirical Bayes test and decision procedures for Q1 and Q2 are described, respectively.

Empirical Bayes testing and decision procedure for Q1:

1. Order genes by $\widehat{\pi_{i0}}$ in an ascending order and denote them as $\widehat{\pi_0^{(1)}}, \widehat{\pi_0^{(2)}}, \ldots, \widehat{\pi_0^{(N)}}$.
2. Let $m = \max\{n : \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi_0^{(i)}} \le \alpha\}$.
3. Report gene G_i to have change points if $G_i \in \mathcal{G}^{\text{Detection}}$, where $\mathcal{G}^{\text{Detection}} = \{i : \widehat{\pi_{i0}} \le \widehat{\pi_0^{(m)}}\}$.

Empirical Bayes testing and decision procedure for Q2:

1. Order genes by $1 - \widehat{\pi_i^*}$ in an ascending order and denote them as $\widehat{\pi_*^{(1)}}, \widehat{\pi_*^{(2)}}, \ldots, \widehat{\pi_*^{(N)}}$.
2. Let $m = \max\{n : \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi_*^{(i)}} \le \alpha\}$.
3. Report gene G_i to have changes points at τ_{i1}^* and τ_{i2}^* if $G_i \in \mathcal{G}^{\text{Identification}}$, where $\widehat{\Pr}(\boldsymbol{\rho}_i = (\tau_{i1}^*, \tau_{i2}^*)|X_i) = \widehat{\pi_i^*}$ and $\mathcal{G}^{\text{Identification}} = \{i : (1 - \widehat{\pi_i^*}) \le \widehat{\pi_*^{(m)}}\}$.

In the following sections we illustrate that the proposed empirical Bayes testing and decision procedures for Q1 and Q2 are both powerful and attain the performance of the oracle procedures asymptotically in controlling FDR at α level by asymptotic analysis and simulation experiments.

3. Asymptotic property. In this section we develop gene-wise hypotheses and demonstrate that the proposed empirical Bayes testing and decision procedure for Q1 and Q2 are both optimal. Our key testing statistics are based on the posterior probability of change-point pattern, which is

$$P(\boldsymbol{\rho}_i = k|X_i) \propto p_k f_k(X_i).$$

3.1. *Multiple testing of detection.* For the problem of detecting genes with change points, let $w_i \in \{0, 1\}$ denote whether gene G_i has change points. Given $w_i = 0$, then gene G_i has no change point. If $w_i = 1$, then gene G_i has one or two change points. The selection task aims to test N hypotheses, that is, for $i = 1, \ldots, N$,

$$H_i^0 : w_i = 0 \quad \text{VS} \quad H_i^\alpha : w_i = 1.$$

Denote d_i as the 0–1 decision rule for the i hypothesis, namely, if $d_i = 1$, then we reject the null hypothesis. In that sense it can be deemed as a classification task. Moreover, theoretically, Yuan and Kendziorski (2006) demonstrate that maximizing a posterior can optimize a classification problem. Sun and Wei (2011) further justify the equality between the multiple testing and weighted classification problems. It is reasonable to find the optimal decision rule of multiple testing through minimizing the classification error. Let $D = \{d_1, \ldots, d_N\}$ and $W = \{w_1, \ldots, w_N\}$. We define a loss function

$$L(W, D) = \frac{1}{N} \sum_{i=1}^{N} \{\lambda_1 (1 - w_i) d_i + w_i (1 - d_i)\}, \tag{3.1}$$

where λ_1 is the ratio of Type I error to Type II error for Q1.

It is known that $w_i = 0$ represents gene G_i having no changes, that is, $\rho_{i1} = 0$, $\rho_{i2} = 0$. Then, $\Pr(w_i = 0) = \Pr(\boldsymbol{\rho}_i = (0, 0)) = 1 - P$. So, $\pi_{i0} = \Pr(w_i = 0|X_i) = \frac{\Pr(X_i|w_i=0)\Pr(w_i=0)}{\Pr(X_i)} = (1 - P) f_0(X_i)/f(X_i)$. Let $\pi_0 = \{\pi_{10}, \ldots, \pi_{N0}\}$. To minimize $E\{L(W, D)\}$, we have the rule $D\{\pi_0, t\mathbb{1}\} = (d_i, i = 1, \ldots, N)$, where $d_i = I(\pi_{i0} < t)$ and $t = 1/(\lambda_1 + 1)$.

In most real scenarios it is of special attention to control the false discovery rate (FDR) at a nominal level α. We consider the FDR and FNR (False Nondiscovery Rate) which are defined as

$$\text{FDR} = E \frac{\sum_{i=1}^{N} (1 - w_i) d_i}{\sum_{i=1}^{N} d_i \vee 1}, \qquad \text{FNR} = E \frac{\sum_{i=1}^{N} w_i (1 - d_i)}{\sum_{i=1}^{N} (1 - d_i) \vee 1}.$$

According to the definition of π_0, we can rewrite FDR in the following form:

$$\begin{aligned} \text{FDR}_{\pi_0} &= E \frac{\sum_{i=1}^{N} (1 - w_i) d_i}{\sum_{i=1}^{N} d_i \vee 1} \\ &= E \left[\frac{1}{\sum_{i=1}^{N} d_i \vee 1} \sum_{i=1}^{N} E_{w_i|X_i} \{(1 - w_i) d_i | X_i\} \right] \\ &= E \frac{\sum_{i=1}^{N} I(\pi_{i0} < t) \pi_{i0}}{\sum_{i=1}^{N} I(\pi_{i0} < t)} \end{aligned}$$

which increases as t increases. Hence, to control the FDR_{π_0} at a nominal level α, we can find a value $t(\alpha)$ that satisfies the decision Bayes rule. Following Theorem 1 and Theorem 2 of Sun and Wei (2011), we can reach the property that: (1) $E\{L(W, D)\}$ can be minimized by $D(\pi_{i0}, t(\alpha)\mathbb{1})$; (2) $D(\pi_{i0}, t(\alpha)\mathbb{1})$ is optimal in the weighted classification problem and optimal in the multiple testing problem. It guarantees that the FNR_{π_0} is optimized at the

smallest level while the FDR_{π_0} is controlled at α level. Suppose that hyperparameters Φ are known, that is, P, P_k, f_k are known. So, the oracle multiple testing procedure is of the form

$$D(\pi_0, t_{\mathrm{OR}}\mathbb{1}) = \big[I(\pi_{i0} < t_{\mathrm{OR}}) : i = 1, \ldots, N\big], \tag{3.2}$$

where the oracle cutoff $t_{\mathrm{OR}} = \mathrm{SUP}\{t \in (0,1) : \mathrm{FDR}_{\pi_0}(t) \le \alpha\}$. Since it is not easy to obtain t_{OR}, as mentioned in Section 2.4, we propose to obtain the estimated cutoff $\widehat{t_{\mathrm{EB}}}$ through the proposed empirical Bayes testing and decision procedure. Next we show its asymptotic consistency with the oracle optimal procedure in choosing the cutoff.

THEOREM 3.1. *Consider the change-point model defined by* (2.1)–(2.9). *Let* $\hat{\Phi}$ *be an estimate of the change-point models* Φ *such that* $\hat{\Phi} \xrightarrow{p} \Phi$. *Let* $\widehat{\mathrm{FDR}}_{\pi_0}$ *and* $\widehat{\mathrm{FNR}}_{\pi_0}$ *be the FDR and FNR level through empirical Bayes procedure for the detection problem, respectively. Then,*

$$\widehat{\mathrm{FDR}}_{\pi_0} = \mathrm{FDR}^{\mathrm{OR}}_{\pi_0} + o(1), \qquad \widehat{\mathrm{FNR}}_{\pi_0} = \mathrm{FNR}^{\mathrm{OR}}_{\pi_0} + o(1),$$

where $\mathrm{FDR}^{\mathrm{OR}}_{\pi_0}$ *and* $\mathrm{FNR}^{\mathrm{OR}}_{\pi_0}$ *are the FDR and FNR level of the oracle procedure* (3.2).

The maximum likelihood estimate (MLE) is used to estimate Φ. Under certain regularity conditions, the MLE is strongly consistent and asymptotically normal. Here, when the number of genes $N \to \infty$, we have $\hat{\Phi} \xrightarrow{p} \Phi$.

3.2. *Multiple testing of identification.* For the identification problem we aim to identify the change-point pattern $\boldsymbol{\rho}_i$ for gene G_i, $i = 1, \ldots, N$. There are $\binom{T}{2}$ possible different change-point patterns after excluding the $(0,0)$ pattern. The selection task requires N hypotheses. For the ith nonnull hypothesis, $i = 1, \ldots, N$, we assume that the change-point pattern for gene G_i is the one who has the largest posterior probability given X_i. Define

$$\Theta_i^{\alpha} = \big\{\boldsymbol{\rho}_i : \operatorname*{argmax}_{\boldsymbol{\rho}_i}\big\{\Pr(\boldsymbol{\rho}_i | X_i)\big\}\big\}.$$

Then, the N hypotheses are given as follows: for $i = 1, \ldots, N$,

$$H_i^0 : \boldsymbol{\rho}_i \notin \Theta_i^{\alpha} \quad \text{VS} \quad H_i^{\alpha} : \boldsymbol{\rho}_i \in \Theta_i^{\alpha}.$$

Define the binary vector $\Gamma = (\gamma_1, \ldots, \gamma_N) \in \{0,1\}^N$, where

$$\gamma_i = \begin{cases} 1 & \text{if } \boldsymbol{\rho}_i \in \Theta_i^{\alpha}, \\ 0 & \text{otherwise.} \end{cases}$$

Let δ_i be the 0–1 decision rule for the ith hypothesis in the identification problem. The null hypothesis is then rejected if $\delta_i = 1$. Here, it turns to be a classification task as well. Let $\Delta = \{\delta_1, \ldots, \delta_N\}$. We define a loss function

$$L(\Gamma, \Delta) = \frac{1}{N}\sum_{i=1}^{N}\big\{\lambda_2(1-\gamma_i)\delta_i + \gamma_i(1-\delta_i)\big\}, \tag{3.3}$$

where λ_2 is the ratio of Type I error to Type II error for Q2.

It is easy to see that $\Pr(\gamma_i = 1 | X_i) = \max_{\boldsymbol{\rho}_i}\{\Pr(\boldsymbol{\rho}_i | X_i)\} = \pi_i^*$ and $\Pr(\gamma_i = 0 | X_i) = 1 - \pi_i^*$. Specifically, $\pi_i^* = \max_{k=1,\ldots,\binom{T}{2}}\{p_k f_k(X_i)\}/f(X_i)$. Let $\pi^* = \{\pi_1^*, \ldots, \pi_N^*\}$. To minimize $E\{L(\Gamma, \Delta)\}$, we have the rule $\Delta\{\pi^*, t\mathbb{1}\} = (\delta_i, i = 1, \ldots, N)$, where $\delta_i = I(1 - \pi_i^* < t)$ and $t = \frac{1}{\lambda_2 + 1}$. The FDR for this problem, therefore, becomes

$$\mathrm{FDR}_{\pi^*} = E\frac{\sum_{i=1}^{N} I(1-\pi_i^* < t)(1-\pi_i^*)}{\sum_{i=1}^{N} I(1-\pi_i^* < t)}.$$

FNR_{π^*} can be generated in a similar way which, to save space, is omitted here. Note that FDR_{π^*} is monotonically increasing with respect to t. Hence, the oracle multiple-testing procedure for Q2 is of the form,

$$\Delta(\pi^*, t_{\mathrm{OR}}\mathbb{1}) = \big[I\big(1-\pi_i^* < t_{\mathrm{OR}}\big) : i = 1, \ldots, N\big], \tag{3.4}$$

where the oracle cutoff $t_{\mathrm{OR}} = \mathrm{SUP}\{t \in (0,1) : \mathrm{FDR}_{\pi^*}(t) \le \alpha\}$. Given the proposed procedure for identification problem Q2 to determine cutoff in Section 2.4, the next theorem shows its asymptotic performance with the oracle procedure in controlling FDR and choosing the cutoff.

THEOREM 3.2. *Consider the change-point model defined by* (2.1)–(2.9). *Let* $\hat{\Phi}$ *be an estimate of the change-point models* Φ *such that* $\hat{\Phi} \xrightarrow{p} \Phi$. *Let* $\widehat{\mathrm{FDR}}_{\pi^*}$ *and* $\widehat{\mathrm{FNR}}_{\pi^*}$ *be the FDR and FNR level through empirical Bayes procedure for the identification problem, respectively. Then,*

$$\widehat{\mathrm{FDR}}_{\pi^*} = \mathrm{FDR}_{\pi^*}^{\mathrm{OR}} + o(1), \qquad \widehat{\mathrm{FNR}}_{\pi^*} = \mathrm{FNR}_{\pi^*}^{\mathrm{OR}} + o(1),$$

where $\mathrm{FDR}_{\pi^*}^{\mathrm{OR}}$ *and* $\mathrm{FNR}_{\pi^*}^{\mathrm{OR}}$ *are the FDR and FNR level of the oracle procedure* (3.4).

Similarly, when the number of genes $N \to \infty$, we have $\hat{\Phi} \xrightarrow{p} \Phi$.

The proofs for Theorem 3.1 and 3.2 are provided in Section A.2 of the Supplementary Material (Tian, Cheng and Wei (2021a)).

3.3. *Asymptotic property under explicit short-ranged dependency.* We assume that the observed individual gene expressions are independent, given its latent mean, while the means of all individual genes follow a common distribution. The marginal gene-expression levels can be dependent under such a hierarchical model. We can further allow individual gene expressions to have short-ranged dependency explicitly, conditional on their means. We show that the corresponding asymptotic properties hold when gene expressions are short-ranged dependent conditional on their latent means (See Section A.1 and A.2 in the Supplementary Material (Tian, Cheng and Wei (2021a))).

4. Results.

4.1. *Simulation settings.* We conduct extensive simulation experiments to investigate the performance of the proposed model under various biological scenarios. We generate $N = 5000$ genes at $T = 8$ time points for each setting, and each time point has three replicates. We vary P from small to large, to represent different proportions of genes with changes. We select $P \cdot N$ genes to have change points randomly, then remaining $(1-P)\cdot N$ genes will be set without change point. For the genes with change points, they have $\binom{8}{2}$ different patterns to have change points, and each pattern is selected randomly with an equal probability. We set $\nu_0 = 0$, $\kappa_0 = 0.1$, $\alpha_0 = 1$, $\beta_0 = 10$ and vary P from 0.01 to 0.3 ($P = 0.01, 0.02, 0.5, 0.1, 0.15, 0.2, 0.25, 0.3$). The hyperparameters are close to the ones we estimate from real datasets. We set the nominal FDR level to be 0.1. The simulation is repeated 100 times for each parameter setting. Averaged sensitivity and FDR levels are reported. The source code of the NNG model for replication is availabe in the Supplementary Material (Tian, Cheng and Wei (2021b)).

We investigate the performance of our empirical Bayes method for both detection and identification questions. Tai et al. developed a method to detect genes that are differentially expressed from time-course data by utilizing moderated likelihood ratio statistic and Hotelling T^2-statistic derived from a multivariate normal empirical Bayes model

(Tai and Speed (2006)) (MN model for short). Kalaitzis et al. proposed to rank differentially expressed gene from time series through the log-ratio of marginal likelihood of Gaussian process regressions (Kalaitzis and Lawrence (2011)) (GP model for short). Limma is a linear model for assessing differential expression in the context of microarray experiments (Smyth (2005)). Here, limma is used to test if genes are constantly expressed across time points. We use the MN model, the GP model and limma as the competing methods to be compared with our Normal Normal-Gamma model (NNG model for short). It is noted that the MN model and GP model give only rankings of genes, and cannot control type I error rate at a given nominal level. In order to compare detection sensitivity of the two methods, we select the same number of genes as the NNG model reports and summarize their FDR and sensitivity. The MN model, GP model and limma do not report change-point positions. To make a comparison, we use a frequentist testing procedure to identify the change-point positions of genes detected by the baseline methods. The frequentist method (Wang, Wei and Li (2014), Zhang and Wei (2016)), essentially, scans the whole sequence and selects a position exhibiting the most dramatic difference as potential change points to be determined by a statistical testing. Here, in our time-course context we implement a frequentist method that scans the whole sequence for finding a change-point pattern from the $\binom{8}{2}$ possible different patterns. The target pattern has the most significant difference between the two sequences separated by change points which is quantified by the p-value of the Student's t-test.

It is good to evaluate the robustness of our procedures to model misspecification. Therefore, we also use a Gamma-Gamma (GG) model (Newton et al. (2001), Kendziorski et al. (2003)) to generate data. The GG model assumes data $x_{ij} \sim \Gamma(\alpha, \beta_i)$ and $\beta_i \sim \Gamma(\alpha_0, \nu)$. Following (Kendziorski et al. (2003)), we set the GG model parameters as $\alpha = 12$, $\alpha_0 = 1$, $\nu = 36$. The other simulation settings remain the same (e.g., $N = 5000$, $T = 8$). Our hierarchical model assumes observed gene expressions are independent conditional on their hidden means. Here, we also use simulation to demonstrate the performance of our testing procedures for analyzing explicitly dependent data. We generate dependent data following a multivariate normal distribution with a short-range dependent covariance structure (Xie et al. (2011)) (see Section A.3 in Supplementary Material (Tian, Cheng and Wei (2021a)) for details).

4.2. *Simulation results*. Parameter estimates averaged among 100 simulations are summarized in Table 1. As shown, the estimated parameters are close to their true values with small standard deviations (SD). This observation indicates that the empirical Bayes approach can estimate parameters well.

Averaged sensitivity and FDR are reported for varied simulation settings in Figure 2. For Q1, our NNG model can control FDR precisely at the nominal level 0.1 for all settings. In contrast, The MN and GP model shows much inflated FDR levels when selecting the same number of genes as the NNG model. Limma tends to be conservative with FDR slightly lower than the nominal level as P increases. Regarding sensitivity, the NNG model shows a clear superiority, in comparison with the MN, GP and limma models, which indicates its optimality. All methods show sensitivity increasing with P, which confirms that, when the number of nonnulls (P) is smaller, it is more challenging to detect them. It is interesting that the NNG model shows a larger improvement under the more challenging settings.

For Q2, the NNG model again controls FDR precisely at the nominal level, while all competing methods yield much inflated FDR. This is not surprising because those competing methods are not designed to identify change-point positions, and their FDRs are thus out of control. Regarding sensitivity, the NNG model outperforms all the competing methods with a similar dominating pattern, which confirms its optimality.

TABLE 1

Summary of parameter estimates. Parameter estimates are averaged over 100 *simulations*; *standard deviations are shown in parentheses. For each simulation*, $v_0 = 0$, $\kappa_0 = 0.1$, $\alpha_0 = 1$, $\beta_0 = 10$

	P							
	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3
$\hat{v_0}$	0.00046 (0.14)	−0.027 (0.14)	−0.004 (0.13)	−0.009 (0.14)	0.008 (0.14)	−0.016 (0.12)	0.01 (0.12)	−0.0084 (0.12)
$\hat{\kappa_0}$	0.1 (0.0023)	0.1 (0.0024)	0.1 (0.0021)	0.1 (0.0021)	0.1 (0.0023)	0.1 (0.002)	0.1 (0.002)	0.1 (0.0022)
$\hat{\alpha_0}$	1 (0.019)	1 (0.018)	1 (0.018)	1 (0.018)	1 (0.018)	1 (0.017)	1 (0.017)	1 (0.017)
$\hat{\beta}_0$	10 (0.24)	10 (0.24)	10 (0.24)	10 (0.23)	10 (0.24)	10 (0.24)	10 (0.25)	10 (0.22)
$\hat{P}$	0.01 (0.0012)	0.02 (0.0015)	0.05 (0.0021)	0.1 (0.003)	0.15 (0.0032)	0.2 (0.0037)	0.25 (0.0043)	0.3 (0.0042)

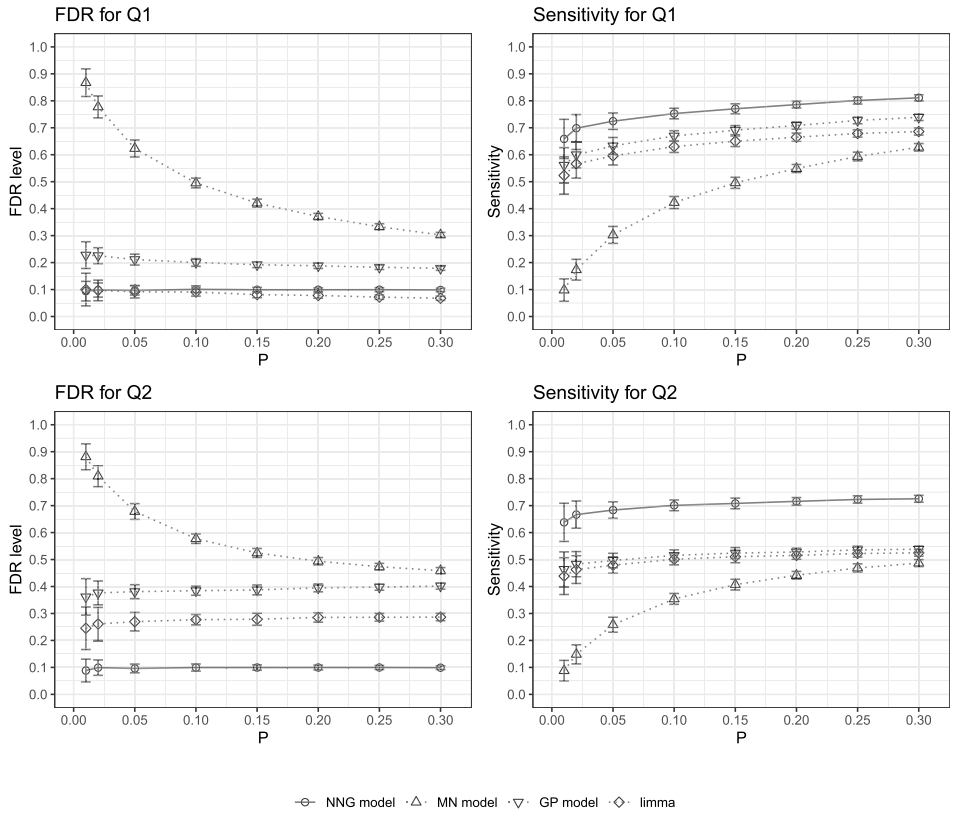

FIG. 2. *Simulation results for different methods. Settings of parameters are*: $\nu_0 = 0$, $\kappa_0 = 0.1$, $\alpha_0 = 1$, $\beta_0 = 10$; *and* P *varies from* 0.01 *to* 0.3. *Mean and standard deviation are plotted for* 100 *simulations on each settings. "NNG model" stands for our Normal Normal-Gamma model, "MN model" stands for the multivariate normal model, "GP model" stands for Gaussian process model, "limma" is the linear model for microarray data* (*MN, GP and limma use the frequentist method to identify change-point positions*).

When applied to the data generated from a Gamma-Gamma model, our NNG model shows similar superiority in comparison with the competing methods which highlights the robustness of our procedures to model misspecification (Figure A1 in the Supplementary Material (Tian, Cheng and Wei (2021a))). Our simulation results also confirm that the proposed testing procedures are able to control the FDR at the nominal level for short-range dependent data (Figure A2 in the Supplementary Material (Tian, Cheng and Wei (2021a))). The running time of our NNG model on simulated datasets of different numbers of genes is summarized in Figure A3 in the Supplementary Material (Tian, Cheng and Wei (2021a)). The result suggests that the running time of NNG model scales linearly with the numbers of genes.

In summary, under extensive simulation scenarios we compare the sensitivity and FDR of our NNG model with several baseline methods and observe the superiority in sensitivity and the robustness of FDR control to detect and identify change points for both Q1 and Q2.

4.3. *Real data application*. To determine the performance on the real data, we apply our NNG model to a public microarray time-course data that studies the systemic inflammation in human (Calvano et al. (2005)). In the study, eight healthy humans were studied, among them, four were selected as cases randomly, then the remaining four became controls. Gene-expression levels were determined by Affymetrix U133A chips immediately before (0 h) and at two, four, six, nine and 24 hours after the intravenous injection of bacterial endotoxin to four cases and placebo to four controls. The goal of this time-course experiment is to identify functional networks responsible for the systemic inflammation activation. Based on the

nature, we can label the inflammatory response process to be the "early" stage and the "late" stage. In the early stage of activation of innate immunity, many inflammatory factors, including cytokines and chemokines, are activated in response to the endotoxin. The activation of proinflammatory factors subsequently triggers the activation of innate immune response genes. Next, in the late period, activities of many negative feedback regulation factors increase and will recover the whole system to normal expression levels finally. We use our NNG model to detect differentially expressed genes and identify when the changes happened.

Some data processing was done before applying our NNG model. For Affymetrix GeneChip one channel array, data processing involves background adjustment, normalization and summarization (Gautier et al. (2004)). We used RMA (Irizarry et al. (2003)) to preprocess raw Affymetrix array data and obtained normalized gene-expression levels. Two conditions of data (cases and controls; note that cases and controls are not paired) were collected in this inflammation time-course experiment. To apply our NNG model, we subtract averaged expression levels of controls from averaged expression levels of cases at each time point to remove baseline fluctuations of gene expressions, and only detect fluctuations specifically related to the stimulus. As a result, we obtained $N = 22{,}283$ probes and $T =$ six time-point sequences that can be used to estimate parameters in NNG model.

After estimating hyperparameters, we answer both Q1 and Q2 for this real dataset on a selected nominal FDR level of 0.1. For Q1, we use limma (Smyth (2005)) with a regression spline to fit the temporal trend as the baseline method. The choice for effective degrees of freedom of the cubic spline is five (suggested by the limma user guide). The limma spline method compares differences in the curves between cases and controls, and uses the same FDR level of 0.1. Under this FDR level, our NNG model detected 2767 probes, and limma spline detected only 917 probes. The comparison of the two methods is summarized in the Venn diagram (Figure 3). As we can see, our NNG model has detected a majority of probes found by the limma spline method, and NNG is more sensitive (detected more probes under the same FDR level). Pathway enrichment analysis is applied to evaluate the rationality of genes represented by the probes detected by our NNG model (GO terms are used; download from MSigDB Collections; the hypergeometric test is applied for the pathway enrichment analysis). We present enriched pathways of all 2767 probes detected to be differentially expressed across time points by our NNG model and the unique 2174 probes detected by NNG model only (Table 2). The result shows that the probes detected by our NNG model are

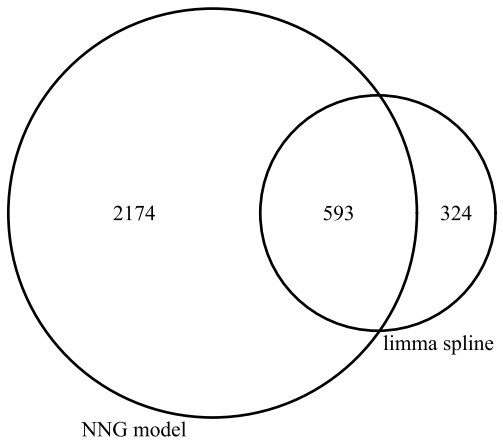

FIG. 3. *Venn diagram of the differentially expressed probes detected by the NNG model* (*our Normal Normal-Gamma model*) *and limma spline at the FDR level of* 0.1 *on the human immune response microarray time–course data.*

TABLE 2

Pathway enrichment analysis of probes detected to be differentially expressed by NNG model from the human immune response microarray time-course data. Top five immunity-related pathways, multiple test adjusted p-values and number of detected genes/total number of genes in a pathway (Hits/Total) are shown

GO term	Adjusted p-value	Hits/Total
(a) GO terms enriched in all 2767 probes detected to be differentially expressed by the NNG model		
GO immune effector process	3.13E−40	309/1010
GO cell activation involved in immune response	1.06E−38	218/613
GO myeloid leukocyte mediated immunity	4.58E−38	187/488
GO leukocyte mediated immunity	3.21E−36	230/687
GO defense response	9.11E−31	350/1336
(b) GO terms enriched in 2174 probes uniquely detected to be differentially expressed by the NNG model		
GO immune effector process	2.99E−29	215/1010
GO myeloid leukocyte activation	7.69E−29	147/566
GO myeloid leukocyte mediated immunity	3.46E−27	131/488
GO leukocyte mediated immunity	1.96E−25	159/687
GO defense response	3.53E−20	235/1366

mainly from immune system related genes which is consistent with the fact that this data was collected to reflect inflammation response genes.

At FDR level of 0.1, we identified change-point positions of 409 probes. Expression profiles and change-point positions of these probes are plotted in Figure 4, and we can confirm the positions of change points identified by the NNG model are indeed at the correct position when changes happened. Gene expression levels separated by these change points are dramatically different. Combining the results on simulated and real data, we can conclude that our NNG model is a compelling and solid method to analyze gene-expression time-course data.

5. Discussion. In summary, we proposed an empirical Bayes change-point model to identify genes with dynamic temporal expression patterns. Theoretically, we show that the performance of our proposed procedure can be asymptotically consistent with the oracle which guarantees that FDR is controlled at a nominal level while minimizing FNR. Simulation and real data studies illustrate that our model is a powerful, accurate and efficient method. To the best of our knowledge, it is the first statistical method for one-conditioned transcriptome time-course data that can detect temporally, dynamically expressed genes and identify when these changes happen on few observations simultaneously. In this change-point model we assume one gene can have three patterns: (1) no change; (2) one change point, expression levels change at one time point and remain that level to the end of experiment; (3) two change points, expression levels change at one time point and restore to original levels later. When considering a scenario to detect genes that are activated by an external stimulus, it is reasonable to assume gene expression levels should restore to normal levels if we can conduct this experiment long enough to capture gene expression levels. The universal existence of the negative feedback regulation mechanism to maintain homeostasis in organisms makes this assumption very reasonable.

With the development of transcriptome study technologies, RNA-seq (Wang, Gerstein and Snyder (2009)) is more and more widely used and replaces microarrays in many research fields. The assumption of our model is that gene-expression levels can be characterized by the normal distribution. For RNA-seq data, we can use proper technique to normalize and adjust RNA-Seq read counts data, so they can be approximated by a normal distribution,

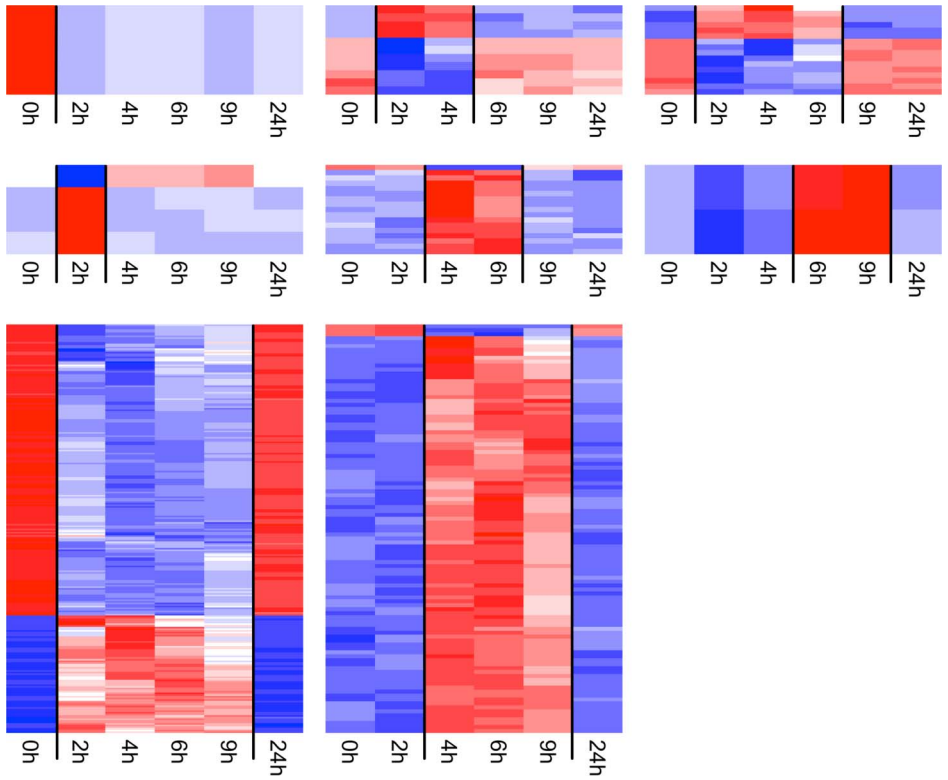

FIG. 4. *Expression profiles and change-point positions of total* 409 *probes identified by the Normal Normal-Gamma model at FDR level of* 0.1. *Each heatmap represents the genes of the same change-point pattern. Rows are probes and columns are time points. The number of rows represents the number of probes having the change-point pattern as identified by our method. Expression levels above average are shown in red, below average are shown in blue, and equivalent to the average are in white. Change-point positions identified by NNG model are shown as vertical black lines.*

such as the "voom" method (Law et al. (2014)) which can generate log scaled counts per million (CPM) values that can be used as input of our NNG model.

Our model considers only two types of change points: expression levels change at a single time point, or expression levels change at one earlier time point and restore at one later time point. Because of the relatively small numbers of time points measured in real experiments (three to 20 usually), it is sufficient to describe the temporal pattern in gene expression levels. However, our change-point model is a flexible framework that can be extended easily, just by following the method for multiple change points (Barry and Hartigan (1992)). But it should be noted multiple change points will introduce a factorial growing complexity of computation and the caveat of over-fitting. The extending to multiple change points and sequences is a future extension of our empirical Bayes change-point method.

Availability. The source code is available at https://github.com/ttgump/EBtimecourse.

Acknowledgments. We thank Dr. Christopher J. Cardinale and Arsh Banerjee for proofreading and editing the manuscript which improved the clarity of the paper. We thank Dr. Jie Zhang for valuable suggestions for experiments. We thank the two anonymous reviewers who provided helpful comments on this manuscript. The research was partially supported by the Natural Science Foundation of China (NSFC) (No. 71771163) and by the National Center for Advancing Translational Sciences (NCATS), a component of the National Institute of Health (NIH) under award number UL1TR003017.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/20-AOAS1403SUPPA; .pdf). We provide proofs for the condition of explicit short-ranged dependency, additional simulation results, and running time of the NNG model.

Source code for the NNG model (DOI: 10.1214/20-AOAS1403SUPPB; .zip). R source code for the NNG model described in this paper.

REFERENCES

ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. et al. (2016). TensorFlow: A system for large-scale machine learning. In 12*th USENIX Symposium on Operating Systems Design and Implementation* (*OSDI* 16) 265–283. USENIX Association, Savannah, GA, USA.

ALLAIRE, J. and TANG, Y. (2020). tensorflow: R Interface to 'TensorFlow'. R package version 2.0.0.

ARBEITMAN, M. N., FURLONG, E. E., IMAM, F., JOHNSON, E., NULL, B. H., BAKER, B. S., KRASNOW, M. A., SCOTT, M. P., DAVIS, R. W. et al. (2002). Gene expression during the life cycle of Drosophila melanogaster. *Science* **297** 2270–2275.

BAR-JOSEPH, Z., GITTER, A. and SIMON, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13** 552–564. https://doi.org/10.1038/nrg3244

BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. MR1150343 https://doi.org/10.1214/aos/1176348521

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

CALVANO, S. E., XIAO, W., RICHARDS, D. R., FELCIANO, R. M., BAKER, H. V., CHO, R. J., CHEN, R. O., BROWNSTEIN, B. H., COBB, J. P. et al. (2005). A network-based analysis of systemic inflammation in humans. *Nature* **437** 1032–1037.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. and SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1962778

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

GAUTIER, L., COPE, L., BOLSTAD, B. M. and IRIZARRY, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20** 307–315.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

KALAITZIS, A. A. and LAWRENCE, N. D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinform.* **12** 180. https://doi.org/10.1186/1471-2105-12-180

KENDZIORSKI, C., NEWTON, M., LAN, H. and GOULD, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22** 3899–3914.

KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41** 495–502. MR0254999 https://doi.org/10.1214/aoms/1177697089

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** 1.

LÖNNSTEDT, I., GRANT, S., BEGLEY, G. and SPEED, T. (2005). Microarray analysis of two interacting treatments: A linear model and trends in expression over time.

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. https://doi.org/10.1186/s13059-014-0550-8

MURPHY, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *Def.* **1** 16.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. and TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8** 37–52. https://doi.org/10.1089/106652701300099074

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SMYTH, G. K. (2005). limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420. Springer, New York, NY.

SUN, W. and WEI, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.* **106** 73–88. MR2816703 https://doi.org/10.1198/jasa.2011.ap09587

TAI, Y. C. and SPEED, T. P. (2005). *DNA Microarrays. Chapter* 20: *Statistical Analysis of Microarray Time Course Data*. CRC Press/CRC, New York.

TAI, Y. C. and SPEED, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* **34** 2387–2412. MR2291504 https://doi.org/10.1214/009053606000000759

TIAN, T., CHENG, R. and WEI, Z. (2021a). Supplement to "An empirical bayes change-point model for transcriptome time course data." https://doi.org/10.1214/20-AOAS1403SUPPA

TIAN, T., CHENG, R. and WEI, Z. (2021b). Source code to "An empirical bayes change-point model for transcriptome time course data." https://doi.org/10.1214/20-AOAS1403SUPPB

TIAN, B., NOWAK, D. E. and BRASIER, A. R. (2005). A TNF-induced gene expression program under oscillatory NF-κB control. *BMC Genomics* **6** 1.

WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. https://doi.org/10.1038/nrg2484

WANG, W., WEI, Z. and LI, H. (2014). A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics* **30** 2162–2170.

XIE, J., CAI, T. T., MARIS, J. and LI, H. (2011). Optimal false discovery rate control for dependent data. *Stat. Interface* **4** 417–430. MR2868825 https://doi.org/10.4310/SII.2011.v4.n4.a1

XUAN, X. and MURPHY, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the* 24*th International Conference on Machine Learning* 1055–1062. ACM, New York.

YANG, J., PENFOLD, C. A., GRANT, M. R. and RATTRAY, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics* **32** 2956–2964.

YUAN, M. and KENDZIORSKI, C. (2006). Hidden Markov models for microarray time course data in multiple biological conditions. *J. Amer. Statist. Assoc.* **101** 1323–1332. MR2307565 https://doi.org/10.1198/016214505000000394

ZHANG, J. and WEI, Z. (2016). An empirical Bayes change-point model for identifying 3′ and 5′ alternative splicing by next-generation RNA sequencing. *Bioinformatics* **32** 1823–1831.

ZHAO, L. P., PRENTICE, R. and BREEDEN, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA* **98** 5631–5636.

ZHAO, Z., WANG, W. and WEI, Z. (2013). An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Ann. Appl. Stat.* **7** 2229–2248. MR3161720 https://doi.org/10.1214/13-AOAS660