A Birth and Death Model for RNA-Seq Data Analysis

by

Zhen Yan

(Under the direction of Liu, Liang)

Abstract

The important problem in analyzing the paired RNA-seq data is that current methods cannot account for the dependence between the paired groups. Moreover, the Poisson or Negative Binomial distributions used in most methods may not fit the data with biological replicates as assumed. Thus, a novel idea of RNA-seq study is in need to help capture new data characteristics. By modeling the change between two groups with the homogenous birth and death process, we propose a new method (a birth and death model) to detect differentially expressed genes. To make our model more appropriate for the real data, we generalize the original model to a mixed birth and death model. The usefulness and advantages of our approaches are demonstrated on simulations and three real datasets.

Index words: RNA-seq, Birth and death process, Differential expression detection
A Birth and Death Model for RNA-Seq Data Analysis

by

Zhen Yan

B.S., Wuhan University of Technology, 2011
M.S., University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016
A Birth and Death Model for RNA-Seq Data Analysis

by

Zhen Yan

Approved:

Professors: Liu, Liang

Committee: McCormick, William
Schliekelman, Paul
Zhao, Shayin
Zhong, Wenxuan

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016
A Birth and Death Model for RNA-Seq Data Analysis

Zhen Yan

March 29, 2016
Acknowledgments

I would never have been able to finish this work without encouragement from my parents. First of all, I would like to thank my advisor Liang Liu, for his inspiring guidance to help me understand the beauty of research, and his invaluable comments during the whole work with this dissertation. I am also very grateful to everyone that have read parts of the manuscript, especially Chuan Wang, Paul Schliekelman, Shayin Zhao, Wenxuan Zhong, and William McCormick. At last, I would like to thank the Department of Statistics at the University of Georgia for providing a five-year cultivation, and letting me see a different world.
Contents

1 Introduction
   1.1 Data Generation .................................................. 2
   1.2 Normalization ..................................................... 3
   1.3 Differential Expression (DE) Test ............................. 4
      1.3.1 Nonparametric Approaches .................................. 5
      1.3.2 Parametric Approaches .................................... 7

2 Birth and Death Model ............................................... 12
   2.1 The BD Model ..................................................... 15
   2.2 BD Parameters Estimation ....................................... 20
      2.2.1 Restrictions for λt and µt .................................. 20
      2.2.2 Probability for δ + β > 1 .................................. 23
      2.2.3 Probability for 2α > 1 ...................................... 26
      2.2.4 First and Second Derivatives ............................... 29
      2.2.5 Newton-Raphson Method ..................................... 35
List of Figures

2.1 An example of the paired RNA-seq data. .................................................. 13
2.2 A birth and death model. ................................................................. 16
3.1 A mixture of several BD processes. ......................................................... 41
4.2.1 ROC (0.05) from Simulation 1 .......................................................... 56
4.2.2 ROC (0.015) from Simulation 1 ......................................................... 57
4.2.3 Power plots from Simulation 1 .......................................................... 58
4.2.4 Estimated FDRs from Simulation 1 ................................................... 59
4.2.5 ROC from Simulation 2 ................................................................. 61
4.2.6 Power plots from Simulation 2 .......................................................... 62
4.2.7 Estimated FDRs from Simulation 2 ................................................... 63
4.2.8 ROC from Simulation 2 (separate) ................................................... 65
4.2.9 Power plots from Simulation 2 (separate) .......................................... 66
4.2.10 Estimated FDRs from Simulation 2 (separate) .................................. 67
4.2.11 ROC (0.05) from Simulation 3 .......................................................... 70
4.2.12 ROC (0.015) from Simulation 3 ......................................................... 71
4.2.13 Power plots from Simulation 3 .......................................................... 72
4.2.14 Estimated FDRs from Simulation 3 ................................................... 73
4.2.15 Histograms of simulated parameters ............................................... 75
5.1.1 Cumulative distributions of representative genes . . . . . . . . . . . . . . . . 90
5.2.1 Venn diagrams of BDPM, DESeq2, edgeR and TSPM . . . . . . . . . . . . . 93
5.2.2 Venn diagrams of MixBD, DESeq2, edgeR and TSPM . . . . . . . . . . . . . 93
5.2.3 Venn diagrams of five methods . . . . . . . . . . . . . . . . . . . . . . . . . . 94
List of Tables

2.2.1 MLEs for $\delta = 0.3$ and $\beta = 0.4$ ................................. 36
2.2.2 MLEs for $\delta = 0.4$ and $\beta = 0.8$ ................................. 36
2.2.3 MLEs for $\delta = 0.8$ and $\beta = 0.6$ ................................. 37
2.2.4 MLEs for $\delta = 0.9$ and $\beta = 0.2$ ................................. 37
2.2.5 MLEs for $\alpha = 0.2$ ..................................................... 37
2.2.6 MLEs for $\alpha = 0.45$ ..................................................... 37
2.2.7 MLEs for $\alpha = 0.7$ ..................................................... 38
3.2.1 MLEs for $\theta_\lambda = 0.1$ ................................................. 46
3.2.2 MLEs for $\theta_\lambda = 0.4$ ................................................. 47
3.2.3 MLEs for $\theta_\lambda = 0.7$ ................................................. 47
3.2.4 MLEs for $\theta_\lambda = 0.9$ ................................................. 47
3.2.5 MLEs for $\theta_\lambda = 0.2$ and $\theta_\mu = 1.5$ ......................... 47
3.2.6 MLEs for $\theta_\lambda = 0.9$ and $\theta_\mu = 1$ ......................... 48
3.2.7 MLEs for $\theta_\lambda = 0.6$ and $\theta_\mu = 0.2$ ......................... 48
4.2.1 TPR and V/R from Simulation 1 ......................................... 55
4.2.2 TPR and V/R from Simulation 2 ......................................... 60
4.2.3 TPR and V/R from Simulation 2 (overdispersed) ................. 64
4.2.4 TPR and V/R from Simulation 2 (standard) ......................... 64
4.2.5 TPR and V/R from Simulation 3 ......................................... 69
Chapter 1

Introduction

RNA Sequencing (RNA-seq) is a powerful next-generation sequencing (NGS) technology in transcriptome studies, which has been commonly used for measuring expression levels of genes [1]. Unlike microarray technologies using the fluorescence intensity measurement, RNA-seq treats the expression of biological entities (e.g. genes) as discrete instead of continuous variables [6]. The RNA-Seq technology has gradually replaced microarrays for the high-throughput studies of gene expression, because RNA-seq: 1) can detect novel genes and isoforms, 2) provides a wide range of expression levels, and hence helps unravel some complex characteristics of the transcriptome and detect inactive genes, 3) has low background noise and high sensitivity [4-6]. In addition to its advantages in quality and samples demanded [2], RNA-seq is able to detect novel gene fusion [7], allele-specific expression [8], alternative splicing [9], and single nucleotide variants [10], etc. Moreover, RNA-seq has been frequently used to understand gene expression profiling under different conditions. Meanwhile, there is of great interest to develop cutting-edge statistical methodologies for RNA-seq data analysis [1]. Improving existing methods or proposing novel approaches for RNA-seq data analysis will greatly expand our horizons in understanding genetic diseases, personalized medicine, and cancer genetics.
1.1 Data Generation

In a typical RNA-seq experiment, with reverse transcription and fragmentation, a sample of mRNA is converted to a library of complementary DNA (cDNA) fragments. Then, sequencing adaptors are attached to one or both ends of each DNA fragment, which allows the fragment to be picked out for later sequencing. One of the five high throughput sequencing platforms (Illumina’s Genome Analyzer, Roche’s 454 Life Sciences, Applied Biosystems’ SOLiD, Helicos BioSciences’ HeliScope and Pacific Biosciences’ SMRT) amplifies and sequences in parallel the attached fragments, producing many short sequences-termed reads. The lengths of reads are typically 30-400 base pairs (bp), varying across sequencing platforms. Single-end reads are sequenced from one end (single-end sequencing), while pair-end reads are sequenced from both ends (pair-end sequencing) [3, 11, 12].

In the sequencing process, cDNA fragments are amplified into clusters of double-stranded DNA after being attached to the surface of a lane, where the library of cDNA has been added into one of the eight lanes of a flow-cell. Then each cluster is sequenced with a certain number of cycles for each flow-cell. Within a cycle, the signals of four fluorescently labeled nucleotides are recorded for each cluster. Thus, the length and number of reads depends on the numbers of cycles and clusters [2].

The reads generated from the previous steps are preprocessed to remove low quality sequences. Then the reads are mapped to a reference genome or aligned with assembled transcriptome, by finding the location where each read can be best matched to the reference with allowance of errors. If a read can be mapped equally well to several locations, different aligners/models may match that read in different ways [1]. The abundance of a gene in the sample is measured by the number of reads mapped to that gene, resulting in the expression for a gene. After that, gene expressions are normalized in terms of the library
size and other factors. The normalized expression values are used for detecting differentially expressed genes [3, 5, 6, 12].

1.2 Normalization

Normalization is a prerequisite for downstream analyses of RNA-seq data. For instance, differential expression analysis for each gene across samples requires between-library normalization, since different lanes may have different sequencing depth (i.e., total number of reads), which is also referred to as library size, and a larger library is expected to contain more reads for all genes. Similarly, within-library normalization is necessary, because longer genes tend to have more mapped reads and thus have higher expression levels. In addition, if we do not compare the expressions among different genes, the effect of gene length can be ignored [1]. In other words, for differential expression detection of the same gene, between-library normalization is enough.

A number of methods have been proposed for normalizing the number of reads matching a particular gene. Library Size Scaling is designed for between-library normalization, in which the scaling factor is calculated as Total Count (TC), Upper Quartile (UQ), Median (Med), DESeq, and Trimmed Mean of M-values (TMM) [5]. The first three scaling methods (TC, UQ, and Med) differ in selecting the divider: the read counts for genes are divided by the library size for TC, 0.75 quantile of the counts for UQ, and the median of non-zero counts in the library for Med separately, and then multiplied by the average library size across samples [2, 5]. The DESeq and TMM methods implemented in the Bioconductor packages DESeq and edgeR assume that most genes are not differentially expressed. In DESeq, to account for a few highly and differentially expressed genes’ influence on the total count, the scaling factor of a given sample is calculated as the median of the ratio in the sample, where the ratio is each gene’s read count over its geometric mean across all samples [5, 14]. In TMM, to
account for the proportionality property of the RNA-seq data (e.g., if a large number of genes
are highly expressed in the treatment sample, the reads mapped to the remaining genes in
that sample should decrease), for each sample, it calculates each gene’s log expression ratio
between this test sample and the reference sample, then gets the weighted mean of the log
ratios as the TMM factor, and provides the scaling factor based on the TMM factor [5, 13].
Both the DESeq and TMM methods can account for the feature that the majority of the
reads is occupied by a few highly-expressed genes [2]. In addition, these two methods preserve
the raw data for later analyses [1]. Quantile (Q) is another between-library normalization
method implemented in the limma Bioconductor package by matching the quantiles of the
distributions of gene counts across lanes [15, 16]. With some non-linear transformation
(e.g. logarithm transforms) in the quantile normalization, a t-test is usually applied to the
transformed RNA-seq data for differential expression (DE) detection. However, the quantile
normalization may bring an unsatisfied testing result, since the transformed data is not well
approximated by a normal distribution [1]. Reads per kilobase per million mapped reads
(RPKM) is a method for both between- and within-library normalization by scaling gene
counts with respect to both library size and gene length [17]. Although RPKM may bias the
variance of read counts, especially for genes with low counts [18], RPKM is still one of the
most widely used normalization methods for RNA-seq data [5].

1.3 Differential Expression (DE) Test

The major goal of differential expression analysis for RNA-seq data is to find genes whose
expression levels are significantly different between normal and treatment samples. A variety
of models have been developed for detecting differentially expressed genes across different
conditions. In general, the detecting methods fall into two categories: parametric and non-
parametric approaches, which will be discussed in detail in the following subsections [4].
1.3.1 Nonparametric Approaches

**Fisher’s exact test** [25, 26]

Let $N_{iA}$ and $N_{iB}$ denote the read counts for gene $i$ under conditions $A$ and $B$ respectively, while $N_{(i-)A}$ and $N_{(i-)B}$ stand for the total counts of the remaining genes correspondingly. Then the data forms a $2 \times 2$ contingency table:

<table>
<thead>
<tr>
<th></th>
<th>condition A</th>
<th>condition B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene $i$</td>
<td>$N_{iA}$</td>
<td>$N_{iB}$</td>
<td>$N_{iA} + N_{iB}$</td>
</tr>
<tr>
<td>Remaining genes</td>
<td>$N_{(i-)A}$</td>
<td>$N_{(i-)B}$</td>
<td>$N_{(i-)A} + N_{(i-)B}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{iA} + N_{(i-)A}$</td>
<td>$N_{iB} + N_{(i-)B}$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Then we get the following null hypothesis and alternative hypothesis:

$H_0$: between condition A and B, there is no significant difference in the expression of gene $i$,

$H_A$: gene $i$ expresses differently between condition A and B.

The probability of observing more than $N_{iA}$ reads in condition A under $H_0$ can be calculated from the table, and the two-sided $p$-value of the Fisher’s exact test for gene $i$ is given by

$$p_i = 2Pr(\text{count} \geq N_{iA}) = 2 \sum_{j=N_{iA}}^{N_{iA}+N_{iB}} \binom{N_{iA} + N_{iB}}{j} \left( \frac{N}{j + N_{(i-)A}} \right) \left( \frac{N_{(i-)A} + N_{(i-)B}}{N_{iA} + N_{(i-)A}} \right).$$  (1.3.1)

The problem of this method is that it only works with unreplicated data, i.e. there can be only one sample in each condition. Because of this, Fisher’s exact test cannot provide enough information of a gene, such as the expression variation of a gene across different replicates.
In order to reduce the effect of sequencing depth on DE test, especially for parametric approaches, and thus control the false discovery rate (FDR), a novel nonparametric method - NOISeq, implemented in the NOISeq package, is developed to model the noise distribution of the data, with or without replicates, being more data-adaptive.

Let $N_{ijA}$ and $N_{ijB}$ denote the normalized read counts for gene $i$ in sample $j$ under conditions $A$ and $B$ separately. Then, for $k = A$ or $B$, $N_{i,k} = \sum_j N_{ijk}$ for technical replicates, while $N_{i,k}$ is the mean or median of all replicates $N_{ijk}$ for biological replicates. Without replicates, $N_{i,k}$ is actually $N_{ijk}$, where $j = 1$. The DE statistics for gene $i$ are log-ratio: $M_i = \log_2 \frac{N_{i,A}}{N_{i,B}}$, and absolute expression difference: $D_i = |N_{i,A} - N_{i,B}|$. To obtain the noise distribution of $M$ and $D$, gene counts in the same condition are contrasted and values of each replicates pair are gathered. If random variables $M^*$ and $D^*$ describe the obtained noise distribution, NOISeq detects differential expression gene $i$ when $|M_i|$ and $D_i$ are higher than the noise values: $|M^*|$ and $D^*$. In the package NOISeq, the probability distribution of $M^*$ and $D^*$ is estimated in an empirical way by algorithm NOISeq-real for replicated data, while a simulation based on the multinomial distribution assumption from NOISeq-sim is applied for unreplicated data, where users can decide the number of replicates.

**Mann-Whitney U test** [21]

Mann-Whitney U test included in the SAMseq package performs consistent and better when the Poisson and Negative Binomial distributions assumption do not fit due to the existence of outliers. In addition, this test can be used for multiple-condition comparisons with wider types of data, such as survival outcomes.

Let $N_{ijk}$ denote the read count for gene $i$ of replicate $j$ under condition $k$. Suppose condition $k$ contains $n_k$ samples, $k = A, B$ and $n_A + n_B = n$. Let $R_{ij}(N)$ be the rank of
$N_{ijA}$ in $N_{i1A},...,N_{inA},N_{i1B},...,N_{inB}$. The two-sample Mann-Whitney statistic (Wilcoxon statistic) in this case is expressed as:

$$T_i = \sum_j R_{ij}(N) - \frac{n_A(n + 1)}{2}.$$  \hspace{1cm} (1.3.2)

For between-library normalization, SAMseq uses a resampling strategy based on a novel method 'Poisson sampling' to get the normalized counts $N'_{ijk}$ relative to $N_{ijk}$, and then replaces $T_i$ with $T'_i$ using $N'_{ijk}$. With a preset cutoff $C$, gene $i$ with $|T'_i| > C$ is detected as differentially expressed, where FDR is estimated by the permutation plug-in method.

Some drawbacks with this method are the requirement of moderate sample size for the resampling process and the complexity of estimating FDR from permutation.

### 1.3.2 Parametric Approaches

Parametric approaches can be more powerful than nonparametric tests, if the distributional assumption is correct. Since the expression levels in RNA-seq data are discrete random variables, they are often assumed to have a Poisson or Negative Binomial distribution. Some RNA-seq studies assume that the expression levels, after logarithm transforms, follows the normal distribution.

**Normal Assumption**

In order to use the well-developed statistical testing methods for microarray data, logarithm transformation is employed to RNA-seq data to make the discrete counts closer to Normal distribution. Because the logarithm transformation of low-coverage genes with read counts equal or close to 0 results in large errors, the counts are adjusted by adding a value: $\log_2(N_{ijk} + \text{cutoff})$, where cutoff is usually 1.
After the log-transformation, a simple paired-\(t\) test or a more popular moderated-\(t\) test, based on the linear models, implemented in the \textit{limma} package [29] can be used for finding DE genes. Another method is to use 2-way ANOVA or 1-way ANOVA for the log-transformed values, where one factor is condition and another possible one is replicate ID if each replicate is available in both conditions, e.g. patients with tumor and normal gene expressions. With the ANOVA model, genes are defined as DE if they satisfy \(|\text{fold change}| > 1.5\) and \(\text{FDR} < 0.05\), where FDR is calculated from ANOVA \(p\)-values and fold change is the ratio of one condition value to the other and estimated from least-squares mean in the inequation [30].

A log-transformation is still hard to fit the Normal distribution and performs poorly for count data except situation of large mean counts and small disperson. Poisson and NB models are more appropriate to deal with counts. [28]

\textbf{Poisson Assumption}

Considering the generation of RNA-seq data, if reads in a library are independently picked and mapped to genes, the gene read counts should mathematically follow a multinomial distribution and can be approximated by the Poisson distribution [31], since the total number of reads is very large. For each gene, the mapped reads can be simplified as a binomial distribution with a small mapping probability, which is easier for Poisson approximation. The goodness-of-fit test suggests that for the same sample sequenced at different lanes, only a few genes do not follow the Poisson distribution assumption, which means the lane effect can be mostly measured by Poisson models [20]. A MA-plot-based method included in the \textit{DEGseq} package [33], an approximated most average powerful (AMAP) test from the \textit{AMAP.Seq} package [19, 32], and a cloud-computing pipeline - \textit{Myrna} [35] are built up on the standard Poisson assumption.

However, the property expecting equal mean and variance of the Poisson distribution is violated by the greater variability among biological replicates from the RNA-seq data, which is
called overdispersion statistically. There are some methods able to handle the overdispersion by allowing unequal mean and variance with a scaling factor, such as a two-parameter generalized Poisson (GP) model focusing on the position-level read counts in the GPSeq package [34] and a two-stage Poisson model (TSPM) [36] separating genes with overdispersion or not, accounting for the overdispersion with Poisson distribution. Among which, TSPM is widely used and its fundamental idea is showed subsequently.

Let $N_{ijk}$ denote the read count for gene $i$ of replicate $j$ under condition $k$, then $N_{ijk} \sim \text{Poisson}(\lambda_{ijk})$. If overdispersion exists, $E(N_{ijk}) = \lambda_{ijk}$ and $\text{Var}(N_{ijk}) = \lambda_{ijk}\phi_i$, where $\phi_i$ is a variance inflation factor. The quasi-likelihood (QL) approach estimates $\phi_i$ as $\hat{\phi}_i = X_i^2 / (n_{\text{replicate}} - n_{\text{condition}})$, where $X_i^2 = \sum_{j,k} (N_{ijk} - \hat{\lambda}_{ijk})^2 / \hat{\lambda}_{ijk}$ is a maximum likelihood estimate (MLE), and $n_{\text{replicate}}$ and $n_{\text{condition}}$ are the numbers of total replicates and conditions separately. Notice that $X_i^2$ is a Pearson test statistic asymptotically converging to a $\chi^2$ distribution when the expected counts are infinite.

The first stage in TSPM method is to test if $\phi_i = 1$ using the adjusted score test and then make two lists for genes with overdispersion and without separately. In the second stage, for the overdispersion list, DE genes are detected by a likelihood ratio test (LRT) statistic:

$$D_{i}^{2*} = -2 \frac{\log L_{\hat{\lambda}_{i0}} - \log L_{\hat{\lambda}_{iA}}}{\hat{\phi}_i}$$  \hspace{1cm} (1.3.3)

where $\hat{\lambda}_{i0}$ is the vector of MLE of Poisson means for gene $i$ under the null hypothesis of no DE and $\hat{\lambda}_{iA}$ is for the alternative one. Similarly, for the genes without overdispersion evidence, since $\phi_i = 1$, the standard LRT statistic is used:

$$D_{i}^{2} = -2[\log L_{\hat{\lambda}_{i0}} - \log L_{\hat{\lambda}_{iA}}].$$  \hspace{1cm} (1.3.4)
\( p \)-values of the tests can be obtained by comparing \( D_i^{2*} \) and \( D_i^2 \) to the \( F_{1,(n_{\text{replicate}}-n_{\text{condition}})} \) and the \( \chi_1^2 \) distributions respectively. \([36]\)

It is worth noting that the asymptotical distributions of the test statistics would fail for genes with small counts. Therefore, TSPM only analyzes genes with total counts from all replicates no less than 10. By this, many low-expressed genes with expression values: 0 or 1 are ignored and may still have DE, which is a drawback of TSPM.

**NB Assumption**

NB distribution is a natural extension of the Poisson distribution when overdispersion appears, which requires the estimate of an additional dispersion parameter. If \( Y \sim \text{NB}(\mu, \sigma^2) \) with mean \( \mu \) and variance \( \sigma^2 \), then \( \Pr(Y = y) = \left(\frac{y + r - 1}{r - 1}\right)p^r(1 - p)^y \), where \( p = \frac{\mu}{\sigma^2} \), \( r = \frac{\mu^2}{\sigma^2 - \mu} \), and \( \sigma^2 = \mu + \mu^2\phi \) with dispersion parameter \( \phi \). It can be shown that the distribution of \( Y \) converges to \( \text{Poisson}(\mu) \) as \( \phi \to 0 \). In the \textit{edgeR} package \([37]\), \( \phi \) is estimated commonly across all genes and only \( \mu \) needs to be estimated for each gene, which accounts for the limitation of too few replicates to estimate both \( \mu \) and \( \sigma^2 \). However, the common dispersion assumption may not be appropriate for practical data. An effective algorithm allowing for more general and data-driven relationships is designed in the \textit{DESeq} package \([24]\), which is discussed subsequently.

Let \( N_{ijk} \) denote the read count for gene \( i \) of replicate \( j \) under condition \( k \), then \( N_{ijk} \sim \text{NB}(\mu_{ijk}, \sigma_{ijk}^2) \). The mean and the variance are expressed as: \( \mu_{ijk} = N_{jk} \pi_{i,k(j)} \) and \( \sigma_{ijk}^2 = \mu_{ijk} + N_{jk}^2 \nu_k(\pi_{i,k(j)}) \), where \( N_{jk} \) is the size factor for replicate \( j \) under condition \( k \), \( \pi_{i,k(j)} \) is a condition-dependent per-gene value with \( \sum_i \pi_{i,k(j)} = 1 \), and \( \nu_k(\pi_{i,k(j)}) \) is the per-gene raw variance parameter - a smooth function of \( \pi_{i,k(j)} \). The estimate of the size factor is the median of the ratios: \( \hat{N}_{jk} = \text{median}_{i} \left( \frac{N_{ijk}}{(\Pi_{j=1}^{n_k} N_{ijk})^{1/n_k}} \right) \), where \( n_k \) is the number of replicates in condition \( k \). Considering all the replicates \( j \) in the same condition \( k \) share the same per-gene value, \( \pi_{i,k(j)} \) can be estimated by the average counts from replicates relative to condition \( k \):
\( \tilde{\pi}_{i,k(j)} = \tilde{\pi}_{ik} = \frac{1}{n_k} \sum_{j:k(j)=k} \frac{N_{ik}}{N_{jk}} \). Correspondingly, an unbiased estimator for the raw variance is based on the function: \( \hat{\nu}_k(\tilde{\pi}_{ik}) = \omega_k(\tilde{\pi}_{ik}) - z_{ik} \), where \( z_{ik} = \frac{\tilde{\pi}_{ik}}{n_k} \sum_{j:k(j)=k} \frac{1}{N_{jk}} \). \( \omega_k \) is a smooth function from local regression on the graph \( (\tilde{\pi}_{ik}, w_{ik}) \) with \( w_{ik} = \frac{1}{n_k-1} \sum_{j:k(j)=k} (\frac{N_{ik}}{N_{jk}} - \tilde{\pi}_{ik})^2 \).

For DE testing, the null hypothesis is assumed as \( \pi_{iA} = \pi_{iB} \) with two conditions A and B. The total counts in each condition for gene \( i \) are: \( N_{iA} = \sum_j N_{ijA} \) and \( N_{iB} \) similarly. Then, the \( p \)-value of observing a pair of summed counts \((n_{iA}, n_{iB})\) is the sum of possibilities not larger than \( \Pr(n_{iA}, n_{iB}) \):

\[
\phantom{\text{Equation 1.3.5}} p_i = \frac{\sum_{(a+b)=(n_{iA}+n_{iB})} \Pr(a,b)}{\sum_{(a+b)=(n_{iA}+n_{iB})} \Pr(a,b)}
\]

where \( \Pr(a,b) \) is the probability of the event \( N_{iA} = a \) and \( N_{iB} = b \), \( N_{ik} \) follows the NB distribution, in which the summed mean and variance can be estimated from the mentioned process. Notice that both \textit{DESeq} [39] and \textit{edgeR} take exact or approximating exact tests and calculate corresponding \( p \)-values. [24]

Other methods assuming the NB distribution, such as ranking genes by the posterior probabilities for each DE gene model from an empirical Bayes approach in the \textit{baySeq} package [38] and Cuffdiff 2 modeling counts uncertainty and overdispersion in this count through the Beta and the NB distributions respectively in the updated \textit{cufflinks} software [39], are also available.
Chapter 2

Birth and Death Model

Poisson and Negative Binomial (NB) models have been widely used for differential expression (DE) analysis of RNA-seq data. In RNA-Seq data analysis, the generated reads are independently mapped to the genes in the reference genome. Let $p$ be the probability that a randomly selected read is mapped to a particular gene. Let $n$ be the total number of reads. Thus, the number of reads being mapped to that gene follows the Binomial distribution with parameters $n$ and $p$. Since a Binomial distribution can be approximated by a Poisson distribution when $n$ is large and $p$ is small, it is assumed that the count of reads being mapped to a gene follows the Poisson distribution with parameter $\lambda = np$. In the Poisson distribution, the variance is equal to the mean, which may be seriously violated in real data analysis. For example, real data analyses indicate that the variance is often much larger than the mean of read counts, i.e., over-dispersion. Thus, the NB distribution is employed to correct the problem of over-dispersion. It has been suggested with the kidney data that RNA-seq data with technical replicates can be adequately explained by the Poisson distribution [20]. However, there is limited research to test goodness of fit of the over-dispersed Poisson and NB models for real RNA-seq data with biological replicates. Since the Poisson distribution fits well to the kidney data with technical replicates, it is arbitrarily assumed
that over-dispersed Poisson and NB distributions can properly account for the extra variance introduced by adding biological replicates to the RNA-Seq data [14]. However, our simulation and real data analyses indicate that the over-dispersed Poisson and NB distributions may not fit well to RNA-seq data.

In this dissertation, we are interested in developing statistical tools for analyzing the paired RNA-seq data, where the two groups are correlated. As showed in Figure 2.1, the dataset consists of $N$ individuals, which are selected at random from the population. Two samples (normal sample and treatment sample) are taken from each individual. Let $X_1, X_2, \ldots, X_N$ be the gene expressions for the normal group, and $Y_1, Y_2, \ldots, Y_N$ are for the treatment group. Since the expression of a gene from each individual is measured twice (normal and treatment samples), we call this type of data paired RNA-seq data. There are a large number of paired RNA-seq data, e.g. the gene expression of a plant’s stem tissue and root tissue. In cancer studies, the normal and tumor tissues are taken from each cancer patient and the expression levels of the normal and tumor tissues are measured using RNA-seq techniques. In Figure 2.1, each patient is measured twice, one is the gene expression of the normal tissue ($X_i$), and the other is the expression of the tumor tissue ($Y_i$). Thus, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ can be considered as paired RNA-seq data. In this study, gene expression is defined as the sum of the reads mapped to the isoforms of the given gene.

\[
\begin{align*}
\text{Group 1:} & \quad (X_1, X_2, \ldots, X_N) \\
\text{Group 2:} & \quad (Y_1, Y_2, \ldots, Y_N) \\
\downarrow & \quad \downarrow & \quad \downarrow \\
\text{Individual 1, Individual 2, \ldots, Individual N}
\end{align*}
\]

Figure 2.1: An example of the paired RNA-seq data.

The traditional Poisson and NB models assume that the gene counts in the treatment group are independent of those in the normal group, and compare the expected counts between the
two groups to decide if a gene is differentially expressed. For example, \(X_1, X_2, \ldots, X_N \sim \text{i.i.d.} \text{NB}(\mu_X, \sigma_X^2)\), and \(Y_1, Y_2, \ldots, Y_N \sim \text{i.i.d.} \text{NB}(\mu_Y, \sigma_Y^2)\), and \(X\) and \(Y\) are independent. Then, if the means (\(\mu_X\) and \(\mu_Y\)) are significantly different, the given gene is considered differentially expressed between the normal and tumor tissues. However, the NB model does not account for the dependence between the two groups.

To account for the dependence of the paired RNA-seq data, we propose that the individual change between the two groups follows a homogeneous birth and death (BD) process. That is, given a gene from individual \(i\), we assume \(Y_i|X_i \sim \text{BD}(\lambda, \mu)\), where \(\lambda\) is the constant birth rate, \(\mu\) is the constant death rate, and individuals share common \(\lambda\) and \(\mu\). The conditional probability distribution of \(Y_i\), given \(X_i, \lambda\) and \(\mu\) \(p(Y_i|X_i, \lambda, \mu)\), can be derived from the homogeneous BD process. Since the expression levels are assumed independent across individuals, the joint likelihood function of \(N\) individuals is equal to the product of individual likelihoods: 
\[
p(Y_1, Y_2, \ldots, Y_N|X_1, X_2, \ldots, X_N, \lambda, \mu) = \prod_{i=1}^{N} p(Y_i|X_i, \lambda, \mu).
\]
With the joint likelihood function, we can find the maximum likelihood estimates (MLE) of \(\lambda\) and \(\mu\), and hence differentially expressed genes can be detected through likelihood ratio test (LRT).

When the death rate is 0, a BD process reduces to a pure birth process. In addition, the probability distribution of gene expression derived from the pure birth process is a NB distribution \([46]\), indicating that the BD process is a generalization of the NB distribution. Although we propose to use the BD model to analyze RNA-Seq data, the BD model can be applied to DNA copy number data (array Comparative Genomic Hybridization data).
2.1 The BD Model

We assume that the birth and death rates are constants, i.e., a homogeneous BD process. Let \( X_t \) denote the number of copies at time \( t \). Given an extremely short period of time \( \Delta t \), starting from \( X_t \), the transition probabilities are given by [46]:

\[
p(X_{t+\Delta t} = X_t + 1 | X_t) = \lambda X_t \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t - 1 | X_t) = \mu X_t \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t | X_t) = 1 - (\lambda + \mu) X_t \Delta t + o(\Delta t)
\]

Chance of more than one of these events = \( o(\Delta t) \)

Let \( \theta_{ij} \) denote the true gene expression from the normal tissue for gene \( i \) from patient \( j \). Similarly, let \( \tau_{ij} \) denote the corresponding true gene expression from the tumor tissue. Then, we assume changing from \( \theta_{ij} \) to \( \tau_{ij} \) in time \( t \) is a BD process, and patients share common BD parameters for each gene. That is,

\[
\tau_{ij} | \theta_{ij} \sim BD(\theta_{ij}, \lambda_i, \mu_i)
\]

Then, the probability of \( \tau_{ij} \) given \( \theta_{ij} \) is derived from the BD process, i.e., [46]:

\[
P(\tau_{ij} | \theta_{ij}) = \begin{cases} 
\sum_{k=0}^{\min(\theta_{ij}, \tau_{ij})} \binom{\theta_{ij}}{k} \binom{\tau_{ij}}{k} \alpha_i^{\theta_{ij} + \tau_{ij} - 2k} (1 - 2\alpha_i)^k, \lambda_i = \mu_i \\
\sum_{k=0}^{\min(\theta_{ij}, \tau_{ij})} \binom{\theta_{ij}}{k} \binom{\tau_{ij}}{k} \delta_i^{\theta_{ij} - k} \beta_i^{\tau_{ij} - k} (1 - \delta_i - \beta_i)^k, \lambda_i \neq \mu_i
\end{cases}
\]

(2.1.1)

where \( \alpha_i = \frac{\mu_i}{1+\mu_i} \), \( \delta_i = \frac{\mu_i (e^{(\lambda_i-\mu_i)t} - 1)}{\lambda_i e^{(\lambda_i-\mu_i)t} - 1} \), \( \beta_i = \frac{\lambda_i (e^{(\lambda_i-\mu_i)t} - 1)}{\lambda_i e^{(\lambda_i-\mu_i)t} - \mu_i} \).

In real data analysis, \( \theta_{ij} \) and \( \tau_{ij} \) are unobservable. The observed expressions of normal and tumor tissues deviate from the true expressions \( \theta_{ij} \) and \( \tau_{ij} \) due to mapping uncertainty. Let
$X_{ijm}$ and $Y_{ijm}$ denote the observed count from the normal and tumor tissues respectively, for the $m$-th replicate of patient $j$ in terms of gene $i$, where $i = 1, 2, \ldots, G$, $j = 1, 2, \ldots, N$, $m = 1, 2, \ldots, M_j$. It is assumed that (Figure 2.2):

$$X_{ijm} \sim \text{Poisson}(\theta_{ij})$$
$$Y_{ijm} \sim \text{Poisson}(\tau_{ij})$$
$$\theta_{ij} \sim \text{Poisson}(a_i)$$

where $a_i$ is the average true expression for gene $i$ across patients. Different from the traditional models, we do not assume a common true gene expression across biological replicates. That is, different patients may have different true gene expressions. Next, we want to derive the probability $Y_{ijm}|X_{ijm}$.

![Figure 2.2: A birth and death model.](image)

We can get $P(Y_{ijm}|\theta_{ij}) = \sum_{\tau_{ij}} P(Y_{ijm}|\tau_{ij}, \theta_{ij}) P(\tau_{ij}|\theta_{ij}) = \sum_{\tau_{ij}} P(Y_{ijm}|\tau_{ij}) P(\tau_{ij}|\theta_{ij})$, since $Y_{ijm}$ is entirely decided by $\tau_{ij}$. Then we get $P(\theta_{ij}|X_{ijm}) = P(X_{ijm}|\theta_{ij}) P(\theta_{ij})/ \sum_{\theta_{ij}} P(X_{ijm}|\theta_{ij}) P(\theta_{ij})$, \[16\]
where

\[
\sum_{\theta_{ij}} P(X_{ijm}|\theta_{ij})P(\theta_{ij}) = \sum_{\theta_{ij}} \frac{\theta_{ij}^{X_{ijm}} e^{-\theta_{ij}} a_i e^{-a_i}}{X_{ijm}! \theta_{ij}^i} = \frac{e^{-a_i \frac{e^{a_i}}{e}}}{X_{ijm}!} \sum_{\theta_{ij}} \theta_{ij}^{X_{ijm}} (\frac{a_i}{e}) \theta_{ij} e^{-(a_i/e)} \theta_{ij}!.
\]

Let \( \theta_{ij}^* = \theta_{ij} \) and \( \theta_{ij}^* \sim \text{Poisson}(a_i/e) \) with moment generating function \( M_{\theta_{ij}^*}(t^*) = E(e^{t^* \theta_{ij}^*}) = e^{\frac{a_i}{e} (e^{t^*} - 1)} \). Then

\[
\sum_{\theta_{ij}} \theta_{ij}^{X_{ijm}} (\frac{a_i}{e}) \theta_{ij} e^{-(a_i/e)} \theta_{ij}! = \sum_{\theta_{ij}} \theta_{ij}^{X_{ijm}} = E(\theta_{ij}^{X_{ijm}}) = M_{\theta_{ij}^*}(X_{ijm})(0) = \frac{\partial X_{ijm} \exp(a_i \frac{e^{t^*}}{e})}{\partial t^* X_{ijm}} |_{t^*=0}.
\]

Then, we get \( P(Y_{ijm}|X_{ijm}) = \sum_{\theta_{ij}} P(Y_{ijm}, \theta_{ij}|X_{ijm}) = \sum_{\theta_{ij}} P(\theta_{ij}|X_{ijm})P(Y_{ijm}|\theta_{ij}) \), since \( P(Y_{ijm}|\theta_{ij}, X_{ijm}) = P(Y_{ijm}|\theta_{ij}) \). Then, with equation (2.1.1) we get the probability function:

\[
P(Y_{ijm}|X_{ijm}) = \left\{ \begin{array}{ll}
\sum_{\theta_{ij}} \sum_{\tau_{ij}} e^{X_{ijm}/\tau_{ij} (a_i/e) \theta_{ij} \tau_{ij}} \frac{\theta_{ij}^{X_{ijm}}min(\theta_{ij}, \tau_{ij})}{\theta_{ij}^{\theta_{ij}} \tau_{ij}^{\tau_{ij} - 1}} \theta_{ij}^{\theta_{ij} + \tau_{ij} - k - 1} \theta_{ij} e^{-(a_i/e)} \theta_{ij}! (1 - 2 \alpha_i) \beta_{ij}^{-k} \mu_i, & \text{if } \lambda_i = \mu_i \\
\sum_{\theta_{ij}} \sum_{\tau_{ij}} e^{X_{ijm}/\tau_{ij} (a_i/e) \theta_{ij} \tau_{ij}} \frac{\theta_{ij}^{X_{ijm}}min(\theta_{ij}, \tau_{ij})}{\theta_{ij}^{\theta_{ij}} \tau_{ij}^{\tau_{ij} - 1}} \theta_{ij}^{\theta_{ij} + \tau_{ij} - k - 1} \theta_{ij} e^{-(a_i/e)} \theta_{ij}! (1 - 2 \alpha_i) \beta_{ij}^{-k} \mu_i, & \text{otherwise}
\end{array} \right.
\]

(2.1.2)

where \( \alpha_i, \delta_i \) and \( \beta_i \) are parameters for the BD process.

Another way worth of trying is to use the Bayesian method. In this section, we only explore the Bayesian estimation of BD parameters for one gene, which can be similarly applied to all the genes. Assume there are \( j = 1, 2, \ldots, J \) patients with individual expected counts, and \( k = 1, 2, \ldots, K_j \) technical replicates for the corresponding individual (another possible denotation is that \( j \) represents different groups, such as cancer stages, and \( k \) denotes patients
within each group), we have the following conditional distributions:

\[ X_{jk} | \theta_j \sim \text{Poisson}(\theta_j) \]
\[ Y_{jk} | \tau_j \sim \text{Poisson}(\tau_j) \]
\[ \tau_j | \theta_j, \alpha \sim \text{BD}(\theta_j, \alpha) \text{ or } \tau_j | \theta_j, \delta, \beta \sim \text{BD}(\theta_j, \delta, \beta), \]

where BD (\( \theta_j, \alpha \)) is the birth and death process when birth rate = death rate, and BD (\( \theta_j, \delta, \beta \)) is for unequal rates.

Using a conjugate prior for the parameter of Poisson distribution, we have a Gamma distribution as the prior distribution for \( \theta_j \):

\[ \theta_j \sim \text{Gamma}(a_j, b_j), \]

where \( a_j > 0 \) and \( b_j > 0 \) are parameters of shape and rate respectively, with mean = \( a_j / b_j \) and variance = \( a_j / b_j^2 \). Then the posterior distribution of \( \theta_j \) is:

\[ \theta_j | X_{j1}, \ldots, X_{jk} \sim \text{Gamma}(A_j, B_j), \]

where \( A_j = a_j + \sum_{k=1}^{K_j} X_{jk} \) and \( B_j = b_j + K_j \). In the Bayesian method, \( K_j \) can be 1, which means there is no technical replicates and expected counts are assumed to vary across patients. The mean of this posterior distribution (the Bayesian estimate of \( \theta_j \)) is \( A_j / B_j \).

Using a conjugate prior makes the Bayesian estimates heavily depend on the parameters in the prior distribution, which means an inappropriate prior distribution may make the posterior distribution unreliable. To reduce the effect of the choice of prior parameters, we consider a hierarchial Bayesian model via assigning an Inverse-Gamma distribution to \( b_j \) as the hyperprior [47]:

\[ b_j \sim \text{Inverse-Gamma}(c_j, d_j). \]
Then the posterior distributions can be written as:

\[ b_j | \theta_j \sim \text{Inverse-Gamma}(a_j + c_j, \theta_j + 1/d_j) \]

\[ \theta_j | b_j, X_{j1}, \ldots, X_{jK_j} \sim \text{Gamma}(a_j + \sum_{k=1}^{K_j} X_{jk}, b_j + K_j). \]

Similarly, we can get the posterior for \( \tau_j \) under a hierarchical Bayesian model.

Assume the \( J \) patients are independent and let \( \tau = (\tau_1, \ldots, \tau_J) \) and \( \theta = (\theta_1, \ldots, \theta_J) \). Thus, \( f(\tau|\theta, \alpha) = \Pi_{j=1}^J f(\tau_j|\theta_j, \alpha) \) and \( f(\tau|\theta, \delta, \beta) = \Pi_{j=1}^J f(\tau_j|\theta_j, \delta, \beta) \). Then at each iteration, with the updated values of \( \tau \) and \( \theta \), we have the posterior probability distributions \( f(\alpha|\tau, \theta) \) of \( \alpha \), or \( f(\delta, \beta|\tau, \theta) \) of \( \delta \) and \( \beta \):

\[
\begin{align*}
    f(\alpha|\tau, \theta) &\propto f(\alpha, \tau, \theta) \propto f(\alpha, \tau|\theta) = f(\tau|\theta, \alpha)f(\alpha) \\
    f(\delta, \beta|\tau, \theta) &\propto f(\delta, \beta, \tau, \theta) \propto f(\delta, \beta, \tau|\theta) = f(\tau|\theta, \delta, \beta)f(\delta)f(\beta),
\end{align*}
\]

where \( f(\alpha) \), \( f(\delta) \), and \( f(\beta) \) are the prior probability distributions, and we can use the non-informative prior \( \text{Uniform}(0, 1) \), which means \( f(\alpha) = f(\delta) = f(\beta) = 1 \). The Metropolis-Hastings algorithm [48] under the Markov Chain Monte Carlo (MCMC) method is employed to estimate \( f(\alpha|\tau, \theta) \) and \( f(\delta, \beta|\tau, \theta) \). At each iteration, \( \theta \) and \( \tau \) are updated from the hierarchical Bayesian model, and starting with arbitrary values of \( \alpha \), \( \delta \), and \( \beta \), parameters are updated with an acceptance probability defined by a Hastings ratio \( H \) in the Metropolis-Hastings algorithm. In other words, at each iteration, the new value \( \alpha' \) is accepted with the probability:

\[
H_\alpha = \min \left\{ \frac{f(\tau|\theta, \alpha')f(\alpha')}{f(\tau|\theta, \alpha)f(\alpha)}, 1 \right\},
\]

and the new value \( \delta' \) (or \( \beta' \)) is accepted with the probability:

\[
H_{\delta, \beta} = \min \left\{ \frac{f(\tau|\theta, \delta', \beta)f(\delta')f(\beta)}{f(\tau|\theta, \delta, \beta)f(\delta)f(\beta)}, 1 \right\}.
\]
In this dissertation, we do not explore the algorithm for the Bayesian model. We plan to get parameter estimates based on Equation 2.1.2. In the future, we may want to get some results for the Bayesian method.

2.2 BD Parameters Estimation

To get the MLEs of BD parameters: \((\hat{\lambda}_i, \hat{\mu}_i) = \arg \max_{(\lambda_i, \mu_i)} \prod_j \prod_m P(Y_{ijm}|X_{ijm})\), we have to calculate integrals over \(\theta_{ij}\) and \(\tau_{ij}\) in Equation 2.1.2, which is computationally intractable. Thus, instead, we can use estimates of \(\theta_{ij}\) and \(\tau_{ij}\) to simplify the computation, and get \((\hat{\lambda}_i, \hat{\mu}_i) = \arg \max_{(\lambda_i, \mu_i)} \prod_j P(\hat{\tau}_{ij}|\hat{\theta}_{ij})\), where \(\hat{\theta}_{ij} = \sum_{m=1}^{M_j} X_{ijm}/M_j\) and \(\hat{\tau}_{ij} = \sum_{m=1}^{M_j} Y_{ijm}/M_j\).

Next, we want to use appropriate algorithms to get the MLEs of BD parameters. Let \(\lambda\) and \(\mu\) denote birth and death rates separately, and \(p\) and \(r\) represent the previous and latter states respectively. We need to consider two cases: unequal rates \((\lambda \neq \mu)\) and equal rates \((\lambda = \mu)\). For each case, we can obtain the probability expression and the corresponding likelihood function, and then get the maximum likelihood estimate (MLE) of the BD parameters. Among some algorithms we tried, the Newton Raphson method provides a good balance between accuracy and speed for calculating the MLEs. Thus we get the derivative functions for applying the Newton Raphson algorithm.

2.2.1 Restrictions for \(\lambda t\) and \(\mu t\)

Considering unequal birth and death rates, the probability of reaching the latter state \(r\) from previous state \(p\) during time \(t\) is [46]:

\[
P = \Pr\{r|p, \lambda, \mu, t\} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \left(\frac{p + r - k - 1}{p - 1}\right) \delta^{p-k} \beta^{r-k} (1 - \delta - \beta)^k, \quad (2.2.1)
\]
where $\delta = \frac{\mu(e^{(\lambda-\mu)t-1})}{\lambda e^{(\lambda-\mu)t-\mu}}$, $\beta = \frac{\lambda(e^{(\lambda-\mu)t-1})}{\lambda e^{(\lambda-\mu)t-\mu}}$, with specific situation $p = 1$:

\[
\begin{align*}
p_r(\delta, \beta) &= \Pr\{r|p, \lambda, \mu, t\} = (1-\delta)(1-\beta)^{r-1}, \ r \geq 1 \\
p_0(\delta, \beta) &= \Pr\{0|p, \lambda, \mu, t\} = \delta.
\end{align*}
\] (2.2.2)

To make the probability nonnegative and less than 1, it is easy to get the restrictions: $0 \leq \delta \leq 1$ and $0 \leq \beta \leq 1$, which can be satisfied by the expression of $\delta$ and $\beta$. If $\lambda > \mu$, $0 < \mu(e^{(\lambda-\mu)t-1}) < \lambda e^{(\lambda-\mu)t-1}$ and $0 < \lambda(e^{(\lambda-\mu)t-1}) < \lambda e^{(\lambda-\mu)t-1} - \mu$, then $0 < \delta < 1$ and $0 < \beta < 1$. Similarly, if $\lambda < \mu$, $\lambda e^{(\lambda-\mu)t-\mu} < \mu(e^{(\lambda-\mu)t-1}) < \lambda e^{(\lambda-\mu)t-1}$, then $0 < \delta < 1$ and $0 < \beta < 1$ as well.

For the RNA-seq data, even after normalization, the read counts can still be greater than 1000, which may make the binomial coefficient terms and the power terms in equation 2.2.1 be infinite and 0 respectively, and thus the multiplication $P$ becomes not computable. To solve this problem, we use log transformations for both binomial coefficient and power terms, and use the exponential function to transform the value back, when calculating $P$. To simplify our algorithms and the computing process, we first try to request all the individual term in equation 2.2.1 to be positive. Therefore, we add one more restriction in our Birth and Death process for RNA-seq data: $0 < 1 - \delta - \beta$.

To make $\delta + \beta < 1$ is equal to satisfy:

\[
\frac{\mu(e^{(\lambda-\mu)t-1}) + \lambda(e^{(\lambda-\mu)t-1})}{\lambda e^{(\lambda-\mu)t-\mu}} < 1.
\]
Then,

\[
\mu e^{(\lambda - \mu)t} - \mu + \lambda e^{(\lambda - \mu)t} - \lambda < \lambda e^{(\lambda - \mu)t} - \mu, \text{ if } \lambda > \mu,
\]

\[
\mu e^{(\lambda - \mu)t} - \mu + \lambda e^{(\lambda - \mu)t} - \lambda > \lambda e^{(\lambda - \mu)t} - \mu, \text{ if } \lambda < \mu.
\]

Eliminating the common terms in both sides of the inequation, moving the negative term in the right side to the left side, and multiplying both sides with time \(t\), we get:

\[
\mu t e^{(\lambda - \mu)t} < \lambda t, \text{ if } \lambda > \mu,
\]

\[
\mu t e^{(\lambda - \mu)t} > \lambda t, \text{ if } \lambda < \mu.
\]

(2.2.3)

Since both birth and death rates should be nonnegative, i.e. \(\lambda t \geq 0, \mu t \geq 0\), applying the monotone increasing function - natural logarithm function to the inequations (2.2.3) we get:

\[
\log(\lambda t) - \lambda t > \log(\mu t) - \mu t, \text{ if } \lambda > \mu,
\]

\[
\log(\lambda t) - \lambda t < \log(\mu t) - \mu t, \text{ if } \lambda < \mu.
\]

(2.2.4)

Suppose for \(x \geq 0\), we define a concave function \(h(x) = \log(x) - x\). Then, making the first derivative equal to 0:

\[
h'(x^*) = \frac{1}{x^*} - 1 = 0,
\]

we get the extreme point of the concave function: \(x^* = 1\), where \(h'(x^*) > 0\) for \(x^* < 1\), while \(h'(x^*) < 0\) for \(x^* > 1\). In other words, \(h(x)\) is monotone increasing within \((0, 1)\) and monotone decreasing within \((1, +\infty)\). Therefore, to let the inequations (2.2.4) hold, the sufficient condition is: \(0 \leq \mu t \leq \lambda t \leq 1\) or \(1 \leq \lambda t \leq \mu t\), . However, including the restrictions \(0 \leq \mu t \leq \lambda t \leq 1\) or \(1 \leq \lambda t \leq \mu t\) in the model is not realistic, which may exclude many real
cases. Thus, we should be careful about using the restrictions in the algorithm for parameter estimating.

2.2.2 Probability for $\delta + \beta > 1$

With the additional restrictions on $\lambda t$ and $\mu t$ ($\delta + \beta < 1$), the computation is simplified and the estimating speed is acceptable. However, applying our algorithm to the real data, the estimating and fitting results seem not very reasonable. Because the mentioned restrictions actually limit the variance to a certain range, while the true variance of the RNA-seq data is usually considered comparatively large. Thus, we remove the parameter restrictions and let $\delta + \beta$ can be larger than 1. However, this brings us new problems. For $\delta + \beta > 1$, the substraction coming from the negative term in equation 2.2.1 is strongly affected by the rounding error of software, such as R, which makes the calculated probability from the program unreliable. So, we need to derive a new BD process probability expression to assure all the individual terms positive when $\delta + \beta > 1$.

For $p = 1$, we define probability generating function: $g_{x}(s, \delta, \beta) = E(s^{x}) = \sum_{i=0}^{\infty} s^{i}p_{i}(\delta, \beta)$, where $p_{i}(\delta, \beta)$ can be found in equations 2.2.2. Then supposing $0 \leq s \leq 1$, recalling $0 \leq \beta \leq$
1, hence $0 \leq s \beta \leq 1$ and $\sum_{i=0}^{\infty} (s \beta)^i = 1/(1 - s \beta)$, we have probability generating function:

\[
g_x(s, \delta, \beta) = \delta s^0 + \sum_{i=1}^{\infty} s^i (1 - \delta) (1 - \beta)^{i-1}
= \delta + s(1 - \delta) (1 - \beta) \sum_{i=0}^{\infty} (s \beta)^i
= \delta + \frac{s(1 - \delta) (1 - \beta)}{1 - s \beta}
= \frac{\delta + s(1 - \delta - \beta)}{1 - s \beta}
= \frac{\delta \beta + \beta s(1 - \delta - \beta)}{\beta (1 - s \beta)}
= \frac{(1 - \delta - \beta + \delta \beta) + (\delta + \beta - 1)(1 - \beta s)}{\beta (1 - s \beta)}
= \frac{(1 - \alpha)(1 - \beta)}{\beta (1 - \beta s)} + \frac{\alpha + \beta - 1}{\beta}.
\]

For $p > 1$, it is actually $p$ independent and identical BD processes moving from state 1 to corresponding state $x_i$, with the sum: $y = \sum_{i=1}^{p} x_i$. From the property of probability generating function:

\[
E(s^y) = E(\sum_{i=1}^{p} x_i)
= E(s^{x_1}) E(s^{x_2}) \cdots E(s^{x_p})
= \prod_{i=1}^{p} g_{x_i}(s, \delta, \beta),
\]
using the Binomial theorem, we have probability generating function of $y$:

$$g_y(s, \delta, \beta) = [g_x(s, \delta, \beta)]^p = \left[\frac{(1 - \delta)(1 - \beta)}{\beta(1 - \beta s)} + \frac{\delta + \beta - 1}{\beta}\right]^p$$

$$= \sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array}\right) \left[\frac{(1 - \delta)(1 - \beta)}{\beta(1 - \beta s)}\right]^{p-k} \left[\frac{\delta + \beta - 1}{\beta}\right]^k$$

$$= \sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array}\right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k (1 - \beta s)^{-(p-k)}.$$

Let $w(s, k) = (1 - \beta s)^{-(p-k)}$, then $g_y(s, \delta, \beta) = \sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array}\right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k w(s, k)$. In addition, for $r \geq 1$, the derivative of $w(s, k)$ can be expressed as:

$$\frac{\partial^r w(s, k)}{\partial s^r} = (p - k)(p - k + 1) \cdots (p - k + r - 1)\beta^r(1 - \beta s)^{-(p-k)-r}$$

$$= \beta^r \prod_{i=0}^{r-1} (p - k + i)(1 - \beta s)^{-(p-k)-r}.$$

Then, when $s = 0$,

$$\frac{\partial^r w(s, k)}{\partial s^r} |_{s=0} = \beta^r \prod_{i=0}^{r-1} (p - k + i).$$

Correspondingly,

$$\frac{\partial^r g_y(s, \delta, \beta)}{\partial s^r} |_{s=0} = \sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array}\right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k \frac{\partial^r w(s, k)}{\partial s^r} |_{s=0}$$

$$= \sum_{k=0}^{p} \left(\begin{array}{c} p \\ k \end{array}\right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k \beta^r \prod_{i=0}^{r-1} (p - k + i).$$

From the property of probability generating function, we also have $\frac{\partial^r g_y(s, \delta, \beta)}{\partial s^r} |_{s=0} = (r!) p_r(\delta, \beta)$,
where $p_r(\delta, \beta)$ is the probability of reaching state $r \geq 1$ when $p > 1$. Then

$$(r!)p_r(\delta, \beta) = \sum_{k=0}^{p} \binom{p}{k} (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k \beta^r \prod_{i=0}^{r-1} (p - k + i).$$

Thus,

$$p_r(\delta, \beta) = \sum_{j=0}^{p} \binom{p}{j} (1 - \delta)^{p-j}(1 - \beta)^{p-j} \beta^{-p}(\delta + \beta - 1)^j \beta^r \prod_{i=0}^{r-1} (p - k + i)/(r!)$$

$$= \sum_{k=0}^{p} \binom{p}{k} (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k \frac{(p + r - k - 1)!}{((p - k - 1)!(r!)}$$

$$= \sum_{k=0}^{p} \binom{p}{k} \left( p + r - k - 1 \right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k.$$

Briefly, $g_y(0, \delta, \beta) = [g_x(0, \delta, \beta)]^p = \delta^p$, then $p_0(\delta, \beta) = \delta^p$, where $p_0(\delta, \beta)$ is the probability of reaching state $0$ when $p > 1$.

Therefore, the probability of reaching the state $r$ from state $p$ can also be expressed as:

$$P = \begin{cases} 
\Pr\{0|p, \lambda, \mu, t\} = \delta^p \\
\Pr\{r|p, \lambda, \mu, t\} = \sum_{k=0}^{p} \binom{p}{k} \left( p + r - k - 1 \right) (1 - \delta)^{p-k}(1 - \beta)^{p-k} \beta^{-p}(\delta + \beta - 1)^k, \quad r \geq 1 
\end{cases}$$

which is equivalent to expression 2.2.1. To minimize the influence from the system rounding error, we use expressions 2.2.1 and 2.2.5 respectively for conditions $\delta + \beta < 1$ and $\delta + \beta > 1$.

### 2.2.3 Probability for $2\alpha > 1$

Similarly, we need to get the two-condition probability expression for the equal birth and death rates case. When $\lambda = \mu$, the probability of reaching the latter state $r$ from the previous
state $p$ during time $t$ is [46]:

$$
P = \Pr \{ r | p, \mu, t \} = \sum_{k=0}^{\text{min}(p,r)} \binom{p}{k} \left( \frac{p + r - k - 1}{p - 1} \right) \alpha^{p+r-2k}(1 - 2\alpha)^k, \tag{2.2.6}$$

where $\alpha = \frac{\mu t}{1 + \mu t}$ and $\mu t = \frac{\alpha}{1 - \alpha}$, with specific situation $p = 1$:

$$
p_r(\alpha) = \Pr \{ r | p, \mu, t \} = \frac{\mu t r^{-1}}{(1 + \mu t)^{r+1}} = \alpha^{r-1}(1 - \alpha)^2, \ r \geq 1
$$

$$
p_0(\alpha) = \Pr \{ 0 | p, \mu, t \} = \frac{\mu t}{1 + \mu t} = \alpha. \tag{2.2.7}$$

Similarly, when $p = 1$ and $0 \leq s \leq 1$, with expressions [2.2.7] the probability generating function is:

$$
g_x(s, \alpha) = \sum_{i=0}^{\infty} s^i p_i(\alpha) = \alpha s^0 + \sum_{i=1}^{\infty} s^i(1 - \alpha)^2 \alpha^{i-1}
$$

$$
= \alpha + s(1 - \alpha)^2 \sum_{i=0}^{\infty} (\alpha s)^i
$$

$$
= \alpha + \frac{s(1 - \alpha)^2}{1 - \alpha s}
$$

$$
= \frac{\alpha(1 - \alpha s) + s(1 - 2\alpha + \alpha^2)}{1 - \alpha s}
$$

$$
= \frac{\alpha + s(1 - 2\alpha)}{1 - \alpha s}
$$

$$
= \frac{\alpha^2 + \alpha s(1 - 2\alpha) + (2\alpha - 1) - (2\alpha - 1)}{\alpha(1 - \alpha s)}
$$

$$
= \frac{(2\alpha - 1)(1 - \alpha s) + (1 - \alpha)^2}{\alpha(1 - \alpha s)}
$$

$$
= \frac{(2\alpha - 1)}{\alpha} + \frac{(1 - \alpha)^2}{\alpha(1 - \alpha s)}. \tag{2.2.7}$$
Then, when $p > 1$, with $y = \sum_{i=1}^{p} x_i$, we have probability generating function:

$$g_y(s, \alpha) = \left[ g_x(s, \alpha) \right]^p$$

$$= \left[ \frac{(2\alpha - 1)}{\alpha} + \frac{(1 - \alpha)^2}{\alpha(1 - \alpha s)} \right]^p$$

$$= \sum_{k=0}^{p} \binom{p}{k} \left[ \frac{2\alpha - 1}{\alpha} \right]^{p-k} \left[ \frac{(1 - \alpha)^2}{\alpha(1 - \alpha s)} \right]^k$$

$$= \sum_{k=0}^{p} \binom{p}{k} (2\alpha - 1)^{p-k}(1 - \alpha)^{2k}\alpha^{-p}(1 - \alpha s)^{-k}.$$ 

Let $w(s, k) = (1 - \alpha s)^{-k}$ with the derivative for $r \geq 1$:

$$\frac{\partial^r w(s, k)}{\partial s^r} = k(k + 1) \cdots (k + r - 1)\alpha^k(1 - \alpha s)^{-k-r}$$

$$= \alpha^r \prod_{i=0}^{r-1} (k + i)(1 - \alpha s)^{-k-r}.$$ 

Then, putting the two equivalent expressions of $\frac{\partial^r g_y(s,\delta,\beta)}{\partial s^r} |_{s=0}$ together:

$$(r!)p_r(\alpha) = \sum_{k=0}^{p} \binom{p}{k} (2\alpha - 1)^{p-k}(1 - \alpha)^{2k}\alpha^{-p}\frac{\partial^r w(s, k)}{\partial s^r} |_{s=0},$$

we get the probability of reaching state $r \geq 1$:

$$p_r(\alpha) = \sum_{k=0}^{p} \binom{p}{k} (2\alpha - 1)^{p-k}(1 - \alpha)^{2k}\alpha^{-p}\alpha^r \prod_{i=0}^{r-1} (k + i)/(r!)$$

$$= \sum_{k=0}^{p} \binom{p}{k} \binom{k + r - 1}{r} (2\alpha - 1)^{p-k}(1 - \alpha)^{2k}\alpha^{r-p}.$$

Briefly, $g_y(0, \alpha) = [g_x(0, \alpha)]^p = \alpha^p$, then $p_0(\alpha) = \alpha^p$.

Therefore, the probability of reaching the state $r$ from state $p$, when $\lambda = \mu$, can also be
expressed as:

\[
P = \begin{cases} 
\Pr\{0|p, \mu, t\} = \alpha^p \\
\Pr\{r|p, \mu, t\} = \sum_{k=0}^{p} \binom{p}{k} \binom{k + r - 1}{r} (2\alpha - 1)^{p-k}(1-\alpha)^{2k}\alpha^{r-p}, \ r \geq 1
\end{cases}
\]  

(2.2.8)

which is equivalent to expression 2.2.6. To minimize the effect of the rounding error, we use expressions 2.2.6 and 2.2.8 respectively for conditions \(2\alpha < 1\) and \(2\alpha > 1\).

### 2.2.4 First and Second Derivatives

**Case 1: \(\lambda \neq \mu, \delta + \beta < 1\)**

From equation 2.2.1, we can get the first derivatives in terms of \(\delta\) and \(\beta\) separately:

\[
\frac{\partial P}{\partial \delta} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p + r - k - 1}{p - 1} ((p - k)(1 - \beta) - p\delta)\delta^{p-k-1}\beta^{-k}(1 - \delta - \beta)^{k-1},
\]

\[
\frac{\partial P}{\partial \beta} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p + r - k - 1}{p - 1} ((r - k)(1 - \delta) - r\beta)\delta^{p-k}\beta^{r-k-1}(1 - \delta - \beta)^{k-1}.
\]

Similarly, the second derivatives in terms of \(\delta\) and \(\beta\) are:

\[
\frac{\partial^2 P}{\partial \delta^2} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p + r - k - 1}{p - 1} (-p\delta(1 - \delta - \beta) + ((p - k)(1 - \beta) - p\delta)
\]

\[
(((p - k - 1)(1 - \beta) - (p - 2)\delta))\delta^{p-k-2}\beta^{r-k}(1 - \delta - \beta)^{k-2},
\]
\[
\frac{\partial^2 P}{\partial \beta^2} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p + r - k - 1}{p - 1} (-r \beta (1 - \delta - \beta) + ((r - k)(1 - \delta) - r \beta) \\
((r - k - 1)(1 - \delta) - (r - 2) \beta) \delta^{p-k} \beta^{r-k-2} (1 - \delta - \beta)^{k-2},
\]

\[
\frac{\partial^2 P}{\partial \delta \partial \beta} = \sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p + r - k - 1}{p - 1} ((p - k) \beta (1 - \delta - \beta) + ((p - k)(1 - \beta) - p \delta) \\
((r - k)(1 - \delta) - (r - 1) \beta) \delta^{p-k-1} \beta^{r-k-1} (1 - \delta - \beta)^{k-2}.
\]

As we mentioned before, to avoid rounding error and make all the above equation expressions able to handle large count values (large \(p\) and \(r\)) in the algorithm, we use log transformation and exponential function in computation. To realize this, adding restrictions: \(0 \leq \delta \leq 1\), \(0 \leq \beta \leq 1\) and \(0 \leq \delta + \beta \leq 1\), the probability expression and derivative equations are rewritten as:

\[
P = \sum_{k=0}^{\min(p,r)} \exp\{\log(\binom{p}{k}) + \log(\binom{p + r - k - 1}{p - 1}) + (p - k) \log(\delta) + (r - k) \log(\beta) \\
+ k \log(1 - \delta - \beta)\}. \]

\[
\frac{\partial P}{\partial \delta} = \sum_{k=0}^{\min(p,r)} ((p - k)(1 - \beta) - p \delta) \exp\{\log(\binom{p}{k}) + \log(\binom{p + r - k - 1}{p - 1}) + (p - k - 1) \log(\delta) \\
+ (r - k) \log(\beta) + (k - 1) \log(1 - \delta - \beta)\}, \]

\[
\frac{\partial P}{\partial \beta} = \sum_{k=0}^{\min(p,r)} ((r - k)(1 - \delta) - r \beta) \exp\{\log(\binom{p}{k}) + \log(\binom{p + r - k - 1}{p - 1}) + (p - k) \log(\delta) \\
+ (r - k - 1) \log(\beta) + (k - 1) \log(1 - \delta - \beta)\}. \]
\[
\frac{\partial^2 P}{\partial \delta^2} = \sum_{k=0}^{\min(p,r)} (-p\delta(1-\delta-\beta) + ((p-k)(1-\beta)-p\delta)((p-k-1)(1-\beta)-(p-2)\delta)) \\
\exp\{\log\left(\frac{p}{k}\right) + \log\left(\frac{p+r-k-1}{p-1}\right)\} + (p-k-2)\log(\delta) \\
+ (r-k)\log(\beta) + (k-2)\log(1-\delta-\beta),
\]

\[
\frac{\partial^2 P}{\partial \beta^2} = \sum_{k=0}^{\min(p,r)} (-r\beta(1-\delta-\beta) + ((r-k)(1-\delta)-r\beta)((r-k-1)(1-\delta)-(r-2)\beta)) \\
\exp\{\log\left(\frac{p}{k}\right) + \log\left(\frac{p+r-k-1}{p-1}\right)\} + (p-k)\log(\delta) \\
+ (r-k-2)\log(\beta) + (k-2)\log(1-\delta-\beta),
\]

\[
\frac{\partial^2 P}{\partial \delta \partial \beta} = \sum_{k=0}^{\min(p,r)} (-p-k)\beta(1-\delta-\beta) + ((p-k)(1-\beta)-p\delta)((r-k)(1-\delta)-(r-1)\beta)) \\
\exp\{\log\left(\frac{p}{k}\right) + \log\left(\frac{p+r-k-1}{p-1}\right)\} + (p-k-1)\log(\delta) \\
+ (r-k-1)\log(\beta) + (k-2)\log(1-\delta-\beta).}
\]

Considering \(J\) processes following a common BD process with previous states: \(p = (p_1, \cdots, p_J)\) and corresponding latter states: \(r = (r_1, \cdots, r_J)\), we get the likelihood:

\[
f(r|p, \delta, \beta) = \prod_{j=1}^{J} P_j = \prod_{j=1}^{J} \Pr\{r_j|p_j, \delta, \beta\}.
\]
Define the log likelihood function: \( g = \sum_{j=1}^{J} \log(P_j) \). Then the first order differentials of \( g \) are:

\[
\frac{\partial g}{\partial \delta} = \sum_{j=1}^{J} \left( \frac{1}{P_j} \frac{\partial P_j}{\partial \delta} \right),
\]

\[
\frac{\partial g}{\partial \beta} = \sum_{j=1}^{J} \left( \frac{1}{P_j} \frac{\partial P_j}{\partial \beta} \right).
\]

In addition, the second order differentials are:

\[
\frac{\partial^2 g}{\partial \delta^2} = \sum_{j=1}^{J} \left( -\left( \frac{1}{P_j} \right)^2 \frac{\partial^2 P_j}{\partial \delta^2} + \frac{1}{P_j} \frac{\partial^2 P_j}{\partial \delta \partial \beta} \right),
\]

\[
\frac{\partial^2 g}{\partial \beta^2} = \sum_{j=1}^{J} \left( -\left( \frac{1}{P_j} \right)^2 \frac{\partial^2 P_j}{\partial \beta^2} + \frac{1}{P_j} \frac{\partial^2 P_j}{\partial \delta \partial \beta} \right),
\]

\[
\frac{\partial^2 g}{\partial \delta \partial \beta} = \sum_{j=1}^{J} \left( -\left( \frac{1}{P_j} \right)^2 \frac{\partial P_j}{\partial \delta} \frac{\partial P_j}{\partial \beta} + \frac{1}{P_j} \frac{\partial^2 P_j}{\partial \delta \partial \beta} \right).
\]

In the following first and second derivatives under other conditions, we also need to take logarithm and exponential transformations for computing. But we do not give repeated details in this dissertation. Also, it is unnecessary to repeat the differentials of the log likelihood function in this section for each case. Here, we only show the derivatives under the following conditions.

**Case 2: \( \lambda \neq \mu, \delta + \beta > 1 \)**

From equation 2.2.5, we can get the first derivatives in terms of \( \delta \) and \( \beta \) separately. If \( r \geq 1 \), then

\[
\frac{\partial P}{\partial \delta} = \sum_{k=0}^{p} \binom{p}{k} \binom{p + r - k - 1}{r} (1 - \delta)^{p-k-1}(1 - \beta)^{p-k}\beta^{-r}((k - p)\beta + p(1 - \delta)),
\]
\[
\frac{\partial P}{\partial \beta} = \sum_{k=0}^{p} \binom{p}{k} \left( \frac{p + r - k - 1}{r} \right) (1 - \delta)^{p-k}(1 - \beta)^{p-k-1} \beta^{r-p-1}(\delta + \beta - 1)^{k-1} (-r\beta^2
\]

\[
+ (2r - p)\beta + (k - r)\beta \delta + (p - r)(1 - \delta)).
\]

Otherwise, for \( r = 0 \), we have:

\[
\frac{\partial P}{\partial \delta} = p\delta^{p-1},
\]

\[
\frac{\partial P}{\partial \beta} = 0.
\]

Similarly, the second derivatives in terms of \( \delta \) and \( \beta \) for \( r \geq 1 \) are:

\[
\frac{\partial^2 P}{\partial \delta^2} = \sum_{k=0}^{p} \binom{p}{k} \left( \frac{p + r - k - 1}{r} \right) (1 - \delta)^{p-k-2}(1 - \beta)^{p-k-2} \beta^{r-p-2}(\delta + \beta - 1)^{k-2}
\]

\[
(((k + 1 - p)\beta + (p - 2)(1 - \delta))((k - p)\beta + p(1 - \delta)) - p(1 - \delta)(\delta + \beta - 1)),
\]

\[
\frac{\partial^2 P}{\partial \beta^2} = \sum_{k=0}^{p} \binom{p}{k} \left( \frac{p + r - k - 1}{r} \right) (1 - \delta)^{p-k}(1 - \beta)^{p-k-2} \beta^{r-p-2}(\delta + \beta - 1)^{k-2}(((3 - r)\beta^2
\]

\[
+ (2r - p - 4)\beta + (k + 2 - r)\delta \beta + (p + 1 - r)(1 - \delta))(-r\beta^2 + (2r - p)\beta
\]

\[
+ (k - r)\delta \beta + (p - r)(1 - \delta)) + (1 - \beta)\beta(\delta + \beta - 1)(-2r\beta + (k - r)\delta + (2r - p)).
\]

\[
\frac{\partial^2 P}{\partial \delta \partial \beta} = \sum_{k=0}^{p} \binom{p}{k} \left( \frac{p + r - k - 1}{r} \right) (1 - \delta)^{p-k-1}(1 - \beta)^{p-k-1} \beta^{r-p-1}(\delta + \beta - 1)^{k-2}
\]

\[
(((p - k)\beta(\delta + \beta - 1) + (r - p)(1 - \beta)(\delta + \beta - 1) + (k - 1)(1 - \beta)\beta
\]

\[
(((k - p)\beta + p(1 - \delta)) + (k - p)(1 - \beta)\beta(\delta + \beta - 1)).
\]
In addition, for \( r = 0 \), we have:

\[
\begin{align*}
\frac{\partial^2 P}{\partial \delta^2} &= p(p - 1)\delta^{p-2}, \\
\frac{\partial^2 P}{\partial \beta^2} &= 0, \\
\frac{\partial^2 P}{\partial \delta \partial \beta} &= 0.
\end{align*}
\]

**Case 3:** \( \lambda = \mu, 2\alpha < 1 \)

From equation [2.2.6], we can get the first and second derivatives in terms of \( \alpha \). If \( r \geq 1 \), then

\[
\frac{\partial P}{\partial \alpha} = \sum_{k=0}^{p} \binom{p}{k} \frac{1}{p} \left( p + r - k - 1 \right) (p + r - 2k - 2(p + r - k)\alpha)\alpha^{p+r-2k-1}(1 - 2\alpha)^{k-1},
\]

\[
\frac{\partial^2 P}{\partial \alpha^2} = \sum_{k=0}^{p} \binom{p}{k} \frac{1}{p} \left( p + r - k - 1 \right) (-2(p + r - k)\alpha(1 - 2\alpha) + (p + r - 2k - 2(p + r - k)\alpha)(p + r - 2k - 1 - 2(p + r - k - 2)\alpha)\alpha^{p+r-2k-2}(1 - 2\alpha)^{k-2}.
\]

Similarly, considering \( J \) processes following a common BD process with previous states: \( \mathbf{p} = (p_1, \cdots, p_J) \) and corresponding latter states: \( \mathbf{r} = (r_1, \cdots, r_J) \), for the equal birth and death rates case, we get the log likelihood: \( g = \sum_{j=1}^{J} \log(P_j) = \sum_{j=1}^{J} \log(\Pr(r_j|p_j, \alpha)) \). Then the first and second order differentials of \( g \) are:

\[
\begin{align*}
\frac{\partial g}{\partial \alpha} &= \sum_{j=1}^{J} \frac{1}{P_j} \frac{\partial P_j}{\partial \alpha}, \\
\frac{\partial^2 g}{\partial \alpha^2} &= \sum_{j=1}^{J} \left( -\frac{1}{P_j} \left( \frac{\partial P_j}{\partial \alpha} \right)^2 + \frac{1}{P_j} \frac{\partial^2 P_j}{\partial \alpha^2} \right).
\end{align*}
\]
Case 4: \( \lambda = \mu, 2\alpha > 1 \)

From equation [2.2.8], we can get the first and second derivatives in terms of \( \alpha \). If \( r \geq 1 \), then

\[
\frac{\partial P}{\partial \alpha} = \sum_{k=0}^{p} \binom{p}{k} \binom{k + r - 1}{r} (2\alpha - 1)^{p-k-1}(1 - \alpha)^{2k-1}\alpha^{r-p-1}(-2(k + r)\alpha^2 \\
+ (3r - p)\alpha + (p - r)),
\]

\[
\frac{\partial^2 P}{\partial \alpha^2} = \sum_{k=0}^{p} \binom{p}{k} \binom{k + r - 1}{r} (2\alpha - 1)^{p-k-2}(1 - \alpha)^{2k-2}\alpha^{r-p-2}((2(3 - k - r)\alpha^2 \\
+ (3r - p - 6)\alpha + (p + 1 - r))(-2(k + r)\alpha^2 + (3r - p)\alpha + (p - r)) \\
+ (2\alpha - 1)(1 - \alpha)\alpha(-4(k + r)\alpha + (3r - p))).
\]

Otherwise, for \( r = 0 \), we have:

\[
\frac{\partial P}{\partial \alpha} = p\alpha^{p-1},
\]

\[
\frac{\partial^2 P}{\partial \alpha^2} = p(p - 1)\alpha^{p-2}.
\]

**2.2.5 Newton-Raphson Method**

Newton-Raphson Method is a numerical analysis method to help finding the successfully better approximation of the MLE, which is the root of \( g' = 0 \). Then the Newton-Raphson iterations can be written as:

\[
\alpha_{n+1} = \alpha_n - \frac{g'(\alpha_n)}{g''(\alpha_n)},
\]
\[
\begin{pmatrix}
\delta_{n+1} \\
\beta_{n+1}
\end{pmatrix} = 
\begin{pmatrix}
\delta_n \\
\beta_n
\end{pmatrix} - 
\begin{bmatrix}
\frac{\partial^2 g}{\partial \delta^2} & \frac{\partial^2 g}{\partial \delta \beta} \\
\frac{\partial^2 g}{\partial \beta \delta} & \frac{\partial^2 g}{\partial \beta^2}
\end{bmatrix}^{-1} 
\begin{bmatrix}
\frac{\partial g}{\partial \delta} \\
\frac{\partial g}{\partial \beta}
\end{bmatrix}
\bigg|_{(\delta_n, \beta_n)} \times 
\begin{bmatrix}
\frac{\partial g}{\partial \delta} \\
\frac{\partial g}{\partial \beta}
\end{bmatrix}
\bigg|_{(\delta_n, \beta_n)}^{-1},
\]

where initial values \( \alpha_0, \delta_0 \) and \( \beta_0 \) can be decided according to the data features.

We apply a simulation study to check the performance of MLE using the Newton-Raphson Method. Table 2.2.1-2.2.7 show the average and the corresponding mean square error (MSE) of the estimates over 50 simulations. We can see for all the values of the true parameters, the estimate mean approaches the true value and the MSE decreases when increasing sample size. The averaged estimate is very close to the real parameter for \( N \geq 200 \), and remains reasonable results for \( N = 10 \). Based on the results of simulation study in this section, our method provides satisfying estimates of the parameters in the BD model.

Table 2.1: Estimating results across different sample sizes for \( \delta = 0.3 \) and \( \beta = 0.4 \). Mean(\( \hat{\delta} \)) and Mean(\( \hat{\beta} \)) are the averages of estimated \( \delta \) and \( \beta \) respectively over 50 simulations, and MSE(\( \hat{\delta} \)) and MSE(\( \hat{\beta} \)) are corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean(( \hat{\delta} ))</td>
<td>0.2352</td>
<td>0.2843</td>
<td>0.2888</td>
<td>0.2964</td>
<td>0.2985</td>
</tr>
<tr>
<td>Mean(( \hat{\beta} ))</td>
<td>0.3460</td>
<td>0.3858</td>
<td>0.3893</td>
<td>0.3975</td>
<td>0.3985</td>
</tr>
<tr>
<td>MSE(( \hat{\delta} ))</td>
<td>0.0167</td>
<td>0.0029</td>
<td>0.0011</td>
<td>0.0006</td>
<td>0.0002</td>
</tr>
<tr>
<td>MSE(( \hat{\beta} ))</td>
<td>0.0127</td>
<td>0.0022</td>
<td>0.0009</td>
<td>0.0004</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Table 2.2: Estimating results across different sample sizes for \( \delta = 0.4 \) and \( \beta = 0.8 \). Mean(\( \hat{\delta} \)) and Mean(\( \hat{\beta} \)) are the averages of estimated \( \delta \) and \( \beta \) respectively over 50 simulations, and MSE(\( \hat{\delta} \)) and MSE(\( \hat{\beta} \)) are corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean(( \hat{\delta} ))</td>
<td>0.3397</td>
<td>0.3724</td>
<td>0.3747</td>
<td>0.3887</td>
<td>0.3904</td>
</tr>
<tr>
<td>Mean(( \hat{\beta} ))</td>
<td>0.7764</td>
<td>0.7904</td>
<td>0.7912</td>
<td>0.7964</td>
<td>0.7968</td>
</tr>
<tr>
<td>MSE(( \hat{\delta} ))</td>
<td>0.0229</td>
<td>0.0066</td>
<td>0.0034</td>
<td>0.0015</td>
<td>0.0006</td>
</tr>
<tr>
<td>MSE(( \hat{\beta} ))</td>
<td>0.0032</td>
<td>0.0007</td>
<td>0.0004</td>
<td>0.0002</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Table 2.2.3: Estimating results across different sample sizes for $\delta = 0.8$ and $\beta = 0.6$. Mean($\hat{\delta}$) and Mean($\hat{\beta}$) are the averages of estimated $\delta$ and $\beta$ respectively over 50 simulations, and MSE($\hat{\delta}$) and MSE($\hat{\beta}$) are corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\delta}$)</td>
<td>0.7708</td>
<td>0.7905</td>
<td>0.7924</td>
<td>0.7981</td>
<td>0.7989</td>
</tr>
<tr>
<td>Mean($\hat{\beta}$)</td>
<td>0.5308</td>
<td>0.5791</td>
<td>0.5848</td>
<td>0.5952</td>
<td>0.5976</td>
</tr>
<tr>
<td>MSE($\hat{\delta}$)</td>
<td>0.0062</td>
<td>0.0010</td>
<td>0.0004</td>
<td>0.0002</td>
<td>0.0001</td>
</tr>
<tr>
<td>MSE($\hat{\beta}$)</td>
<td>0.0305</td>
<td>0.0045</td>
<td>0.0013</td>
<td>0.0007</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Table 2.2.4: Estimating results across different sample sizes for $\delta = 0.9$ and $\beta = 0.2$. Mean($\hat{\delta}$) and Mean($\hat{\beta}$) are the averages of estimated $\delta$ and $\beta$ respectively over 50 simulations, and MSE($\hat{\delta}$) and MSE($\hat{\beta}$) are corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\delta}$)</td>
<td>0.8735</td>
<td>0.8982</td>
<td>0.8983</td>
<td>0.8983</td>
<td>0.9000</td>
</tr>
<tr>
<td>Mean($\hat{\beta}$)</td>
<td>0.1692</td>
<td>0.1783</td>
<td>0.1826</td>
<td>0.1897</td>
<td>0.1982</td>
</tr>
<tr>
<td>MSE($\hat{\delta}$)</td>
<td>0.0031</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>MSE($\hat{\beta}$)</td>
<td>0.0291</td>
<td>0.0070</td>
<td>0.0046</td>
<td>0.0025</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Table 2.2.5: Estimating results across different sample sizes for $\alpha = 0.2$. Mean($\hat{\alpha}$) and MSE($\hat{\alpha}$) are the average and the mean square error respectively of the estimated $\alpha$ over 50 simulations.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\alpha}$)</td>
<td>0.1967</td>
<td>0.1948</td>
<td>0.2034</td>
<td>0.2004</td>
<td>0.2008</td>
</tr>
<tr>
<td>MSE($\hat{\alpha}$)</td>
<td>0.0061</td>
<td>0.0008</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Table 2.2.6: Estimating results across different sample sizes for $\alpha = 0.45$. Mean($\hat{\alpha}$) and MSE($\hat{\alpha}$) are the average and the mean square error respectively of the estimated $\alpha$ over 50 simulations.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\alpha}$)</td>
<td>0.4369</td>
<td>0.4398</td>
<td>0.4415</td>
<td>0.4438</td>
<td>0.4480</td>
</tr>
<tr>
<td>MSE($\hat{\alpha}$)</td>
<td>0.0113</td>
<td>0.0022</td>
<td>0.0020</td>
<td>0.0008</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
### 2.3 Detecting Differentially Expressed Genes

Under the birth and death model, differential expression detection of gene $i$ between two groups can be achieved from the following null and alternative hypotheses:

$$H_{0i} : \lambda_i = \mu_i \text{ vs. } H_{Ai} : \lambda_i \neq \mu_i,$$

where $\lambda_i$ and $\mu_i$ are birth and death rates respectively in the BD process of gene $i$. We expect this: compared to the normal tissue, gene $i$ is over-expressed when $\lambda_i > \mu_i$, under-expressed when $\lambda_i < \mu_i$, and not differentially expressed when $\lambda_i = \mu_i$.

Let $\hat{\alpha}_i$ be the estimated BD parameter for gene $i$ when $\lambda_i = \mu_i$, while $\hat{\delta}_i$ and $\hat{\beta}_i$ are estimated under unequal-rate assumption for the BD process of gene $i$. Then $g(\hat{\alpha}_i)$ and $g(\hat{\delta}_i, \hat{\beta}_i)$ are the log likelihoods at the corresponding MLEs. Comparing the log likelihoods at the MLEs obtained under the two hypotheses, we get the likelihood ratio test (LRT) statistic:

$$D_i^2 = -2(g(\hat{\alpha}_i) - g(\hat{\delta}_i, \hat{\beta}_i)),$$

and the LRT $p$-value can be acquired through comparing $D_i^2$ to the $\chi^2_1$ distribution.

In a gene expression study, the DE detection is usually applied to thousands of genes, e.g., the RNA-seq data of humans typically include more than 20,000 genes. For a multiple testing

---

Table 2.2.7: Estimating results across different sample sizes for $\alpha = 0.7$. Mean($\hat{\alpha}$) and MSE($\hat{\alpha}$) are the average and the mean square error respectively of the estimated $\alpha$ over 50 simulations.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=50</th>
<th>N=100</th>
<th>N=200</th>
<th>N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\alpha}$)</td>
<td>0.6713</td>
<td>0.6899</td>
<td>0.7018</td>
<td>0.6968</td>
<td>0.6975</td>
</tr>
<tr>
<td>MSE($\hat{\alpha}$)</td>
<td>0.0126</td>
<td>0.0017</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
problem, the proportion of incorrect rejections can reach an unacceptable level considering thousands of tests, if we do not adjust the $p$-values. Hence, in the BD model, the vector of $p$-values in terms of thousands of DE tests is adjusted according to the Benjamini-Hochberg (BH) procedure [45] to control the false discovery rate (FDR). Then, the adjusted $p$-values are compared to the desired FDR level to decide the DE genes.
Chapter 3

Mixed Birth and Death Model

For the BD model proposed in chapter 2, we use only one BD process to model the change between normal and tumor tissues. However, in the cancer study, even in a sample of normal tissue, it is a mixture of different cell types, which may include some tumor cells. Thus, given a gene and a patient, the observed gene expression may be measured from a mixture of different cells, and there may be different BD processes corresponding to different cell types. Thus, in this chapter, we proposed a mixed BD model to study the combination of different BD processes.

Suppose we have \( m \) different BD processes, with corresponding proportions \( p_1, p_2, \ldots, p_m \). Let \( X^i_t \) denote the total population size for individual \( i \) at time \( t \), where \( i = 1, 2, \ldots, m \), and assume that \( X^i_t \) is capable of increasing with birth rate \( \lambda_i \) and decreasing with death rate \( \mu_i \). For the mixture of \( m \) different BD processes, let \( X_t = p_1X^1_t + p_2X^2_t + \ldots + p_mX^m_t \), and \( \lambda_t \) and \( \mu_t \) denote the birth and death rates at time \( t \). Then, given an extremely short period
of time $\Delta t$, starting from $X_t$, we have the transition probabilities [46]:

\[
p(X_{t+\Delta t} = X_t + 1|X_t) = (p_1\lambda_1 X_t^1 + p_2\lambda_2 X_t^2 + \cdots + p_m\lambda_m X_t^m) \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t - 1|X_t) = (p_1\mu_1 X_t^1 + p_2\mu_2 X_t^2 + \cdots + p_m\mu_m X_t^m) \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t|X_t) = 1 - (p_1(\lambda_1 + \mu_1) X_t^1 + p_2(\lambda_2 + \mu_2) X_t^2 + \cdots + p_m(\lambda_m + \mu_m) X_t^m) \Delta t + o(\Delta t)
\]

Chance of more than one of these events $= o(\Delta t)$

If we let $\lambda_t = \sum^m p_i X_i^i \lambda_i / X_t$ and $\mu_t = \sum^m p_i X_i^i \mu_i / X_t$, we can have:

\[
p(X_{t+\Delta t} = X_t + 1|X_t) = \lambda_t X_t \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t - 1|X_t) = \mu_t X_t \Delta t + o(\Delta t)
\]

\[
p(X_{t+\Delta t} = X_t|X_t) = 1 - (\lambda_t + \mu_t) X_t \Delta t + o(\Delta t)
\]

which is a non-homogeneous BD process, since $\lambda_t$ and $\mu_t$ are functions of $t$, depending on the time point.

For the $m$ different BD processes, let $X_i$ and $Y_i$ denote starting and ending values respectively. We want to get the conditional probability function $p(Y|X)$ in Figure 3.1, where $X = \sum^m p_i X_i$ and $Y = \sum^m p_i Y_i$.

![Figure 3.1](image-url)
Suppose the \( m \) BD processes are independent: \( X_i|Y_i \sim BD(Y_i, \lambda_i, \mu_i) \), where \( i = 1, 2, \ldots, m \), we have the moment generating function for \( Y \) [46]:

\[
M_Y(\theta, t) = \prod_{i=1}^{m} \left( \frac{\mu_i \nu_i(\theta, t) - 1}{\lambda_i \nu_i(\theta, t) - 1} \right)^{X_i}, \text{ where } \nu_i(\theta, t) = \frac{(e^\theta - 1)e^{(\lambda_i - \mu_i)\nu_i t}}{\lambda_i e^\theta - \mu_i} \text{ and } i = 1, 2, \ldots, m
\]

The probability function of \( Y|X \) can be derived from the moment generating function, by differentiating \( M_Y(\theta, t) \) with respect to \( t \). However, it is computationally intensive. Thus, we proposed a mixed BD model in section 3.1 with a prior distribution to measure the different \( \lambda_i \) and \( \mu_i \).

### 3.1 The Mixed BD Model

In the mixed sample, birth (death) rate may vary across different cell types. Thus, we assume birth (\( \lambda \)) and death (\( \mu \)) rates are random variables following appropriate prior distributions. Exponential distribution is an intuitive prior choice, since the Exponential variable is only defined in the positive region, like \( \lambda \) and \( \mu \), and its probability density function is simple. According to the probability density function of Exponential distributions, it means most of the Exponential-prior generated rates are comparatively small. That is, with an Exponential prior, a large proportion of the generated BD processes have small \( \lambda \) and \( \mu \). To start with a simple case, we first study the mixed BD model with Exponential prior distribution.

From the probability expression in a BD process, we know it is the same to rewrite \( \lambda t \) and \( \mu t \) as \( \lambda \) and \( \mu \), since \( \lambda \) and \( \mu \) always come with \( t \), and it is impossible to separate \( \lambda \) (\( \mu \)) from \( \lambda t \) (\( \mu t \)) in estimation. So we let \( \lambda = \lambda t \) and \( \mu = \mu t \), where \( \lambda \sim \text{Exponential}(\theta \lambda) \) and
\( \mu \sim \text{Exponential}(\theta_\mu) \) independently. That is,

\[
\begin{align*}
    f(\lambda) &= \frac{1}{\theta_\lambda} e^{-\frac{\lambda}{\theta_\lambda}}, \quad \theta_\lambda > 0, \\
    f(\mu) &= \frac{1}{\theta_\mu} e^{-\frac{\mu}{\theta_\mu}}, \quad \theta_\mu > 0, \\
    f(\lambda, \mu) &= \frac{1}{\theta_\lambda \theta_\mu} e^{-\frac{\lambda}{\theta_\lambda} - \frac{\mu}{\theta_\mu}}, \quad \theta_\lambda, \theta_\mu > 0.
\end{align*}
\]

For a BD process changing from the first stage \( p \) to the second stage \( r \), where \( p, r \geq 0 \), if the expressions of the two stages are expected to be not different, which is the non-DE case, the expected value of \( r \) should be equal to \( p \), that is [46]:

\[
E(r) = E(E(r|\lambda, \mu)) = E(\exp(\lambda - \mu)) = p,
\]

where

\[
E(E(r|\lambda, \mu)) = \int_0^\infty \int_0^\infty \exp(\lambda - \mu) f(\lambda, \mu) d\lambda d\mu
\]

\[
= p \int_0^\infty \frac{1}{\theta_\lambda} e^{\frac{\lambda}{\theta_\lambda} - 1} d\lambda \int_0^\infty \frac{1}{\theta_\mu} e^{\frac{\mu}{\theta_\mu} - 1} d\mu
\]

\[
= p \frac{1}{1 - \theta_\lambda} \frac{1}{1 + \theta_\mu}.
\]

Considering \( p, r \geq 0 \), we have the limitation: \( \theta_\lambda \in (0, 1) \). Thus, our mixed BD model with Exponential prior has hyperparameters with domain: \( \theta_\lambda \in (0, 1) \) and \( \theta_\mu \in (0, \infty) \). To make \( E(r) = p \), we get:

\[
(1 - \theta_\lambda)(1 - \theta_\mu) = 1.
\]

That is,

\[
\theta_\mu = \frac{\theta_\lambda}{(1 - \theta_\lambda)}.
\]
In sum, under mixed BD model with Exponential prior, the value of the latter stage is expected to be equal to the value of the previous stage, if \( \theta_\mu = \frac{\theta_\lambda}{(1-\theta_\lambda)} \).

### 3.2 Mixed BD Parameters Estimation

With the probability expressions in the BD model, we write the probability distribution function of BD process as:

\[
\Pr\{r|p, \lambda, \mu\} = \begin{cases} 
\sum_{k=0}^{\min(p,r)} \binom{p}{k} \binom{p+r-k-1}{p-1} \delta^{p-k} \beta^{r-k} (1 - \delta - \beta)^k, & \text{if } \delta + \beta < 1 \\
\sum_{k=0}^{p} \binom{p}{k} \binom{p+r-k-1}{r} (1 - \delta)^{p-k} (1 - \beta)^p \beta^{r-p} (\delta + \beta - 1)^k, & \text{if } \delta + \beta > 1 
\end{cases}
\]

(3.2.1)

where \( \delta = \frac{\mu(\lambda - \mu - 1)}{\lambda(\lambda - \mu - 1)} \), \( \beta = \frac{\mu(\lambda - \mu - 1)}{\lambda(\lambda - \mu - 1)} \), \( \lambda \) and \( \mu \) are the numbers of birth and death happening during the BD process.

#### 3.2.1 \( \theta_\mu \neq \theta_\lambda/(1 - \theta_\lambda) \)

Suppose \( \lambda \sim \text{Exponential}(\theta_\lambda) \) and \( \mu \sim \text{Exponential}(\theta_\mu) \), where \( \lambda \) and \( \mu \) are independent, \( \theta_\lambda \) and \( \theta_\mu \) are independent, \( \theta_\lambda \in (0, 1) \) and \( \theta_\mu \in (0, \infty) \). Then, changing from stage \( p \) to stage \( r \), we have the probability expression:

\[
P = \Pr\{r|p, \theta_\lambda, \theta_\mu\} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta_\lambda \theta_\mu} e^{-\frac{\lambda}{\lambda} - \frac{\mu}{\mu}} d\lambda d\mu.
\]

(3.2.2)

Correspondingly, the first derivative is:

\[
\frac{\partial P}{\partial \theta_\lambda} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta_\mu} (-\frac{1}{\theta_\lambda^2} + \frac{\lambda}{\theta_\lambda^3}) e^{-\frac{\lambda}{\lambda} - \frac{\mu}{\mu}} d\lambda d\mu.
\]

(3.2.3)
\[
\frac{\partial P}{\partial \theta} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta} \left(-\frac{1}{\theta^2} + \frac{\mu}{\theta^3}\right) e^{-\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.4}
\]

Similarly, the second derivative is:

\[
\frac{\partial^2 P}{\partial \theta^2} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta^2} \left(-\frac{1}{\theta^3} + \frac{2\lambda^2}{\theta^4} + \frac{\mu^2}{\theta^5}\right) e^{-\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.5}
\]

\[
\frac{\partial^2 P}{\partial \theta \mu} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta^2} \left(\frac{1}{\theta^2} + \frac{\lambda}{\theta^3}\right) \left(-\frac{1}{\theta^2} + \frac{\mu}{\theta^3}\right) e^{-\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.6}
\]

\[
\frac{\partial^2 P}{\partial \theta \mu} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \left(-\frac{1}{\theta^2} + \frac{\lambda}{\theta^3}\right) \left(-\frac{1}{\theta^2} + \frac{\mu}{\theta^3}\right) e^{-\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.7}
\]

\[3.2.2 \quad \theta = \theta \lambda / (1 - \theta \lambda)\]

Suppose \(\lambda \sim \text{Exponential}(\theta)\) and \(\mu \sim \text{Exponential}(\theta)\), where \(\theta = \theta \lambda / (1 - \theta \lambda)\), \(\theta \in (0, 1)\), and \(\lambda\) and \(\mu\) are independent. Then, changing from stage \(p\) to stage \(r\), we have the probability expression:

\[
P = \Pr\{r|p, \theta\} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{1}{\theta} \frac{\lambda - \theta}{\theta - \mu} e^{\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.8}
\]

Correspondingly, the first derivative is:

\[
\frac{\partial P}{\partial \theta} = \int_0^\infty \int_0^\infty \Pr\{r|p, \lambda, \mu\} \frac{\lambda^2 - (\lambda + \mu + 2)\theta \lambda + \lambda + \mu}{\theta^4} e^{\frac{\lambda}{\theta} - \frac{\mu}{\theta^2}} d\lambda d\mu. \tag{3.2.9}
\]
Similarly, the second derivative is:

\[
\frac{\partial^2 P}{\partial \theta^2} = \int_0^\infty \int_0^\infty Pr\{r|p, \lambda, \mu\} \left\{ -2\theta^3 + (4\lambda + 4\mu + 6)\theta^2 - (\lambda + \mu + 6)(\lambda + \mu)\theta + (\lambda + \mu)^2 \right\} \theta d\lambda d\mu.
\]

\[e^{\mu - \frac{\lambda + \mu}{\lambda}} d\lambda d\mu. \quad (3.2.10)\]  

3.2.3 Estimation Checking

Similar to the BD model, based on equations 3.2.9 - 3.2.7, we apply a numerical analysis method - Newton-Raphson Method to finding a good approximation of the MLE of the two hyperparameters: \(\theta_\lambda\) and \(\theta_\mu\). Then we take a simulation study to check the performance of our estimation. Table 3.2.1 - 3.2.7 show the average and the corresponding mean square error (MSE) of the estimates over 10 simulations. We can see for most of the values of the true parameters, the estimate mean approaches the true value when increasing sample size, and for all the simulations, the MSE decreases if enlarging sample size. The averaged estimate is close to the real parameter for comparatively large \(N\), like 50, and provides reasonable results for small \(N\), like 10. Based on the results of simulation study in this section, our method provides satisfying estimates of the parameters in the mixed BD model with Exponential prior.

Table 3.2.1: Estimating results across different sample sizes for \(\theta_\lambda = 0.1\) and \(\theta_\mu = \theta_\lambda/(1 - \theta_\lambda)\). Mean(\(\hat{\theta}_\lambda\)) is the average of estimated \(\theta_\lambda\) over 10 simulations, and MSE(\(\hat{\theta}_\lambda\)) is the corresponding mean square error.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean((\hat{\theta}_\lambda))</td>
<td>0.0870</td>
<td>0.1148</td>
<td>0.0931</td>
<td>0.1057</td>
</tr>
<tr>
<td>MSE((\hat{\theta}_\lambda))</td>
<td>0.0094</td>
<td>0.0029</td>
<td>0.0012</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
Table 3.2.2: Estimating results across different sample sizes for $\theta_\lambda = 0.4$ and $\theta_\mu = \theta_\lambda/(1 - \theta_\lambda)$. Mean($\hat{\theta}_\lambda$) is the average of estimated $\theta_\lambda$ over 10 simulations, and MSE($\hat{\theta}_\lambda$) is the corresponding mean square error.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.3339</td>
<td>0.3674</td>
<td>0.3729</td>
<td>0.4025</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.0303</td>
<td>0.0156</td>
<td>0.0056</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

Table 3.2.3: Estimating results across different sample sizes for $\theta_\lambda = 0.7$ and $\theta_\mu = \theta_\lambda/(1 - \theta_\lambda)$. Mean($\hat{\theta}_\lambda$) is the average of estimated $\theta_\lambda$ over 10 simulations, and MSE($\hat{\theta}_\lambda$) is the corresponding mean square error.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.6607</td>
<td>0.6671</td>
<td>0.6727</td>
<td>0.6729</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.0113</td>
<td>0.0063</td>
<td>0.0043</td>
<td>0.0017</td>
</tr>
</tbody>
</table>

Table 3.2.4: Estimating results across different sample sizes for $\theta_\lambda = 0.9$ and $\theta_\mu = \theta_\lambda/(1 - \theta_\lambda)$. Mean($\hat{\theta}_\lambda$) is the average of estimated $\theta_\lambda$ over 10 simulations, and MSE($\hat{\theta}_\lambda$) is the corresponding mean square error.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=5</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.7841</td>
<td>0.8832</td>
<td>0.8826</td>
<td>0.8928</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.0149</td>
<td>0.0038</td>
<td>0.0024</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Table 3.2.5: Estimating results across different sample sizes for $\theta_\lambda = 0.2$ and $\theta_\mu = 1.5$. Mean($\hat{\theta}_\lambda$) and Mean($\hat{\theta}_\mu$) are the averages of estimated $\theta_\lambda$ and $\theta_\mu$ respectively over 10 simulations, and MSE($\hat{\theta}_\lambda$) and MSE($\hat{\theta}_\mu$) are the corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.1573</td>
<td>0.1801</td>
<td>0.1628</td>
</tr>
<tr>
<td>Mean($\hat{\theta}_\mu$)</td>
<td>1.1341</td>
<td>1.5391</td>
<td>1.4864</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.1556</td>
<td>0.0879</td>
<td>0.0557</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\mu$)</td>
<td>0.4240</td>
<td>0.4151</td>
<td>0.2194</td>
</tr>
</tbody>
</table>
Table 3.2.6: Estimating results across different sample sizes for $\theta_\lambda = 0.9$ and $\theta_\mu = 1$. Mean($\hat{\theta}_\lambda$) and Mean($\hat{\theta}_\mu$) are the averages of estimated $\theta_\lambda$ and $\theta_\mu$ respectively over 10 simulations, and MSE($\hat{\theta}_\lambda$) and MSE($\hat{\theta}_\mu$) are the corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.7370</td>
<td>0.8368</td>
<td>0.8829</td>
</tr>
<tr>
<td>Mean($\hat{\theta}_\mu$)</td>
<td>1.0405</td>
<td>1.0174</td>
<td>0.9842</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.2636</td>
<td>0.1805</td>
<td>0.1040</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\mu$)</td>
<td>0.4053</td>
<td>0.2727</td>
<td>0.1773</td>
</tr>
</tbody>
</table>

Table 3.2.7: Estimating results across different sample sizes for $\theta_\lambda = 0.6$ and $\theta_\mu = 0.2$. Mean($\hat{\theta}_\lambda$) and Mean($\hat{\theta}_\mu$) are the averages of estimated $\theta_\lambda$ and $\theta_\mu$ respectively over 10 simulations, and MSE($\hat{\theta}_\lambda$) and MSE($\hat{\theta}_\mu$) are the corresponding mean square errors.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N=10</th>
<th>N=20</th>
<th>N=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean($\hat{\theta}_\lambda$)</td>
<td>0.6258</td>
<td>0.5906</td>
<td>0.5827</td>
</tr>
<tr>
<td>Mean($\hat{\theta}_\mu$)</td>
<td>0.2012</td>
<td>0.1805</td>
<td>0.1950</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\lambda$)</td>
<td>0.2111</td>
<td>0.1360</td>
<td>0.0828</td>
</tr>
<tr>
<td>MSE($\hat{\theta}_\mu$)</td>
<td>0.0934</td>
<td>0.0799</td>
<td>0.0680</td>
</tr>
</tbody>
</table>

3.3 Detecting Differentially Expressed Genes

As in the BD model, suppose $J$ processes follow mixed BD processes with a common prior - Exponential distribution, and have previous states: $p = (p_1, \cdots, p_J)$ and corresponding latter states: $r = (r_1, \cdots, r_J)$. Then, we get the log likelihoods: $g(\theta_\lambda) = \sum_{j=1}^{J} \log(\text{Pr}(r_j|p_j, \theta_\lambda))$ and $g(\theta_\lambda, \theta_\mu) = \sum_{j=1}^{J} \log(\text{Pr}(r_j|p_j, \theta_\lambda, \theta_\mu))$ for the restricted condition ($\theta_\mu = \theta_\lambda/(1-\theta_\lambda)$) and generalized condition ($\theta_\mu \neq \theta_\lambda/(1-\theta_\lambda)$) separately.

Like the BD model, under the mixed BD model (MixBD), differential expression detection for a certain gene between two groups can be achieved from the following null and alternative
hypotheses:

\[ H_0 : \theta_\mu = \frac{\theta_\lambda}{1 - \theta_\lambda} \] vs. \[ H_A : \theta_\mu \neq \frac{\theta_\lambda}{1 - \theta_\lambda} , \]

where \( \theta_\lambda \) and \( \theta_\mu \) are hyperparameters of the Exponential priors for \( \lambda \) and \( \mu \) respectively.

Let \( \hat{\theta} \) be the estimated hyperparameter in the mixed BD model when \( \theta_\mu = \theta_\lambda/(1 - \theta_\lambda) \), while \( \hat{\theta}_\lambda \) and \( \hat{\theta}_\mu \) are estimated for the generalized case. Then \( g(\hat{\theta}) \) and \( g(\hat{\theta}_\lambda, \hat{\theta}_\mu) \) are the log likelihoods at the corresponding MLEs. Comparing the log likelihoods at the MLEs obtained under the two hypotheses, we get the likelihood ratio test (LRT) statistic:

\[ D^2 = -2(g(\hat{\theta}) - g(\hat{\theta}_\lambda, \hat{\theta}_\mu)) , \]

and the LRT \( p \)-value can be acquired through comparing \( D^2 \) to the \( \chi^2_1 \) distribution. Similar to the BD model, when applying this mixed BD model to the real data for DE detection, we need to use the Benjamini-Hochberg procedure [45] to control the false discovery rate.
Simulation study is commonly used to examine the performance of statistical methods, since data is generated with known characteristics and closely mimic the question of interest. In this section, a simulation study was conducted to test the properties of the birth and death model (BDPM) and the mixed birth and death model (MixBD). After that, we want to compare their ability of detecting differential expression (DE) genes to other DE detection methods based on Poisson or Negative Binomial models. To reasonably approach the amount of human genes as well as reduce the computing consumption, 10,000 genes were simulated, from which 2000 genes were simulated as differentially expressed with respect to the simulating distribution, while the remaining 8000 genes were simulated as non-DE genes. Each setting in the following section was independently simulated 10 times, and the comparison results were averaged over the 10 generated data sets. For each simulation setting, the proposed BDPM, the proposed MixBD, the Negative Binomial model from Ander and Huber’s DESeq2 [43] (version 1.6.3), Robinson and Smyth’s edgeR [37] (version 3.8.6), and the two-stage Poisson model (TSPM) from Auer and Doerge’s code (available at http://www.stat.purdue.edu/~doerge/software/TSPM.R) were applied to detecting the differentially expressed genes between two groups. DESeq2 and edgeR are available from
Bioconductor web site (http://www.bioconductor.org/) [44], and all the analyses were employed in the environment R (version 3.1.1).

In order to study the performance of the mentioned five DE testing methods across different numbers of replicates, we simulated different sample sizes for each setting. One of our main interests of using BDPM and MixBD is to detect genes relative to cancer. As we know, the RNA-seq data of cancer usually contains comparatively large sample size, which are paired normal-to-cancer gene expressions of more than ten patients. Also, to mimic the data sets with 7, 21, and 53 patients in chapter 5 we generated sample sizes in both groups as $N = 5$, $N = 10$, $N = 20$, $N = 40$ via simulating 5, 10, 20, and 40 replicates.

To evaluate and compare the detecting results from different methods, we first look at their ability of ranking genes. With approximating exact test or likelihood ratio test, all the five statistical methods studied in this chapter can produce $p$-values for each gene. If we rank the simulated 10,000 genes from the smallest $p$-value to the largest, upper genes are more likely to be differentially expressed between groups than lower genes, according to the testing method. An ideal DE testing method should put all the true DE genes in the top of the list, followed by the non-DE genes. Selecting different thresholds of $p$-values would declare different amounts of DE genes, among which some are truly differentially expressed as true positives, while some are actually non-differentially expressed as false positives. With the known DE information in the simulated data sets, we can calculate the average of true positive rates (TPR) and the average of false positive rates (FPR) across 10 simulations, where TPR and FPR are the proportion of true DE genes (TP/P) and the proportion of non-DE genes (FP/N) respectively that are declared to be differentially expressed. Then a receiver operating characteristic (ROC) curve obtained via plotting the TPR against the FPR at different thresholds of $p$-values, can help us compare the ability of ranking genes in the accurate order of DE from different approaches. DE testing methods with greater area under the ROC curve, rank genes better. Considering the Benjamini-Hochberg (BH) control
of the false discovery rate (FDR), the estimated FDR is calculated by averaging the proportions of declared DE genes that are actually non-DE genes (FP/(TP+FP)) across 10 simulations. Then, at different levels of FDR control using BH method (i.e., different values of nominal FDR), TPR and estimated FDR can be respectively plotted against the nominal FDR to compare the testing power and the type I error rates of the five statistical methods [36].

4.1 Simulation Settings

**Simulation 1** In the first treatment group, gene counts were drawn from a Negative Binomial($\mu_g, \phi_g$) distribution, where the expected mean $\mu_g$ was randomly sampled from an Exponential($\lambda = 50$) distribution and the dispersion parameter $\phi_g$ was fixed as constant 0.5 for all genes. Then for non-DE genes, parameter $\alpha$ was generated from a Uniform(0, 1) distribution, and the gene counts in the second treatment group were sampled from a birth and death process considering equal birth and death rates, given parameter $\alpha$ and the previous gene counts from the first group. For DE genes, parameters $\delta$ and $\beta$ were sampled from Uniform(0, 1) separately, gene expressions in the second group were drawn from a birth and death process considering unequal birth and death rates, given parameters $\delta$, $\beta$ and the previous gene expressions. Under simulation 1, BDPM and MixBD used rounded RPKM as inputs, where to calculate RPKM, gene length was defined as the mean counts across samples for each gene. DESeq2 and edgeR used their built-in method to estimate normalization factors with default parameters. TSPM took the total number of reads per sample as offsets.

**Simulation 2** We adopted simulation 2 in [36], where approximately half of the 10,000 genes were generated from a standard Poisson distribution and the remaining from an overdispersed Poisson distribution with dispersion parameter varying across genes. For each gene, a 0-1
variable \( z \) was sampled from a Bernoulli(0.5) distribution to decide if this gene followed an overdispersed Poisson distribution. \( z = 0 \) meant there was no overdispersion, then counts of non-DE genes were generated from a Poisson(\( \lambda_{\text{overall}} \)) distribution and DE genes were drawn from Poisson(\( \lambda_N \)) and Poisson(\( \lambda_C \)) distributions. In the Poisson distributions, parameters \( \lambda_{\text{overall}} \) (for non-DE genes), \( \lambda_N \) and \( \lambda_C \) (for DE genes) were sampled from a distribution \( \exp(\text{Pareto}(\text{location} = 3, \text{shape} = 7)) \). If \( z = 1 \), variables \( v \), \( v_N \), and \( v_C \) were drawn from distributions \( \text{Gamma}(\frac{\lambda_{\text{overall}}}{\phi - 1}, \frac{\phi - 1}{\lambda_{\text{overall}}}) \), \( \text{Gamma}(\frac{\lambda_N}{\phi - 1}, \frac{\phi - 1}{\lambda_N}) \), and \( \text{Gamma}(\frac{\lambda_C}{\phi - 1}, \frac{\phi - 1}{\lambda_C}) \), where \( \phi \) is drawn from \( \exp(\text{Pareto}(\text{location} = 1, \text{shape} = 3)) - 1 \) on a gene basis. Then non-DE gene counts were generated from Poisson(\( \lambda_{\text{overall}} v \)), and DE gene counts were simulated from Poisson(\( \lambda_N v_N \)) and Poisson(\( \lambda_C v_C \)) separately, where \( \lambda_{\text{overall}}, \lambda_N \) and \( \lambda_C \) were generated as for \( z = 0 \). For the simulated data sets under simulation 2, BDPM and MixBD used rounded RPKM as normalized counts, where gene length was defined as the mean gene expression across samples; DESeq2 and edgeR used their default setting; TSPM used RPKM instead of the total number of reads per sample to normalize the counts in this simulation setting, since the testing performance using RPKM was much better, and TSPM accepts various normalization methods.

**Simulation 3** In the first treatment group, gene counts were drawn from a Negative Binomial(\( \mu_g, \phi_g \)) distribution, where the expected mean \( \mu_g \) was randomly sampled from an Exponential(\( \text{mean} = 10 \)) distribution and the dispersion parameter \( \phi_g \) was fixed as constant 1 for all genes. Then for non-DE genes, parameter \( \theta \) was generated from a Uniform(0,1) distribution, and the gene counts in the second treatment group were sampled from a mixed birth and death process considering Exponential priors with \( \exp(\theta) \) and \( \exp(\theta/(1 - \theta)) \) for \( \lambda \) and \( \mu \) respectively, given parameter \( \theta \) and the previous gene counts from the first group. For DE genes, parameters \( \theta_\lambda \) and \( \theta_\mu \) were sampled from Beta(7,7) and Gamma(0.2,0.4) separately, gene expressions in the second group were drawn from a mixed birth and death process considering Exponential priors with \( \exp(\theta_\lambda) \) and \( \exp(\theta_\mu) \) for \( \lambda \) and \( \mu \) respectively,
given parameters $\theta_\lambda$, $\theta_\mu$ and the previous gene expressions. Under simulation 3, BDPM and MixBD used rounded RPKM as inputs, DESeq2 and edgeR stayed with their default setting, and TSPM took the total number of reads per sample as offsets.

### 4.2 Simulation Results

For Simulation 1, where gene counts follow a birth and death process, BDPM and MixBD have significantly higher ability to rank genes in the order of differential expression than the other three methods, while DESeq 2, edgeR, and TSPM have close ranking ability, especially when increasing sample sizes (Figure 4.2.1 and 4.2.2). In other words, BDPM is most likely to rank the DE genes in front of non-DE genes among the five approaches. Similarly, BDPM has the highest power to detect DE and MixBD shows power slightly lower than BDPM, followed by DESeq2, edgeR, and TSPM literally, and the difference in power for the last three methods becomes small when sample sizes are large (Figure 4.2.3). For both ROC curves and power plots, the difference between BDPM, MixBD and the other three methods slightly decreases if we have more replicates. But the degree of decrease is not satisfying, since the difference is still large for $N = 40$. Both BDPM and MixBD have reasonable control of the FDR across different sample sizes, while DESeq2, edgeR, and TSPM have inflated FDRs for small sample sizes, such as $N \leq 10$ (Figure 4.2.4). Applying the Benjamini-Hochberg (BH) procedure to control the FDR at the 5% level and comparing the TPR and the estimated FDR across sample sizes, we can see for all five methods, the TPR is rising with sample sizes, while the true FDR declines for methods DESeq2, edgeR and TSPM, when enlarging sample sizes (Table 4.2.1). Moreover, BDPM and MixBD have significantly higher TPRs than other methods.
Table 4.2.1: Mean true positive rate (TPR) and mean true proportions of false positives among discoveries (V/R) across sample sizes from 10 simulations of setting 1 (under BD model) at controlled 5% FDR via the BH method.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>( N = 5 )</th>
<th>( N = 10 )</th>
<th>( N = 20 )</th>
<th>( N = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDPM TPR (SE)</td>
<td>0.554 (1.27%)</td>
<td>0.714 (1.33%)</td>
<td>0.814 (1.03%)</td>
<td>0.879 (0.78%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.095 (1.02%)</td>
<td>0.101 (0.46%)</td>
<td>0.116 (0.75%)</td>
<td>0.147 (0.95%)</td>
</tr>
<tr>
<td>DESeq2 TPR (SE)</td>
<td>0.081 (0.66%)</td>
<td>0.18 (1.41%)</td>
<td>0.302 (1.17%)</td>
<td>0.451 (1.75%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.323 (3.59%)</td>
<td>0.237 (2.66%)</td>
<td>0.131 (1.48%)</td>
<td>0.077 (1.13%)</td>
</tr>
<tr>
<td>edgeR TPR (SE)</td>
<td>0.028 (0.44%)</td>
<td>0.143 (1.31%)</td>
<td>0.286 (1.18%)</td>
<td>0.446 (1.75%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.428 (4.89%)</td>
<td>0.174 (2.99%)</td>
<td>0.07 (1.27%)</td>
<td>0.049 (0.97%)</td>
</tr>
<tr>
<td>TSPM TPR (SE)</td>
<td>0.014 (0.17%)</td>
<td>0.078 (1.31%)</td>
<td>0.242 (1.32%)</td>
<td>0.416 (1.89%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.714 (4.08%)</td>
<td>0.27 (3.73%)</td>
<td>0.074 (1.66%)</td>
<td>0.045 (1.00%)</td>
</tr>
<tr>
<td>MixBD TPR (SE)</td>
<td>0.391 (2.04%)</td>
<td>0.602 (1.27%)</td>
<td>0.717 (1.03%)</td>
<td>0.783 (1.15%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.161 (1.21%)</td>
<td>0.152 (0.55%)</td>
<td>0.135 (0.59%)</td>
<td>0.165 (0.94%)</td>
</tr>
</tbody>
</table>

For Simulation 2, where genes are drawn from a mixture of standard and overdispersed Poisson distributions, the DE ranking ability of DESeq2, edgeR, and TSPM are slightly higher than BDPM’s and MixBD’s, and the dominance becomes faint for \( N = 40 \) (Figure 4.2.5). The ranking difference between BDPM and those three methods is scarce for genes from a Poisson distribution, but comparatively larger for genes from an overdispersed Poisson distribution, while MixBD shows a competitive ranking ability as those three methods, under a comparatively higher FPR, even for the overdispersed Poisson distribution (Figure 4.2.8). The power curves are very similar to the ROC curves, and similar comparison conclusions can be drawn (Figure 4.2.6 and 4.2.9). It worth mentioning that MixBD performs as good as the three methods for Poisson simulation, according to figures of ROC and power curves.

All five methods demonstrate control of FDRs for small sample size (e.g. \( N \leq 10 \)), But MixBD tends to underestimate the FDR when \( N \geq 20 \). This may be caused by the complexity and uncertainty of MixBD. On the other hand, BDPM shows good control across replicates, but not as conservative as those three methods except for \( N = 40 \) (Figure 4.2.7). BDPM has slightly higher true FDRs than expected for small sample sizes (\( N \leq 10 \)) and
Figure 4.2.1: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 1 where genes follow a birth and death process, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR smaller than 0.05. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.2: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 1 where genes follow a birth and death process, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR smaller than 0.015. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.3: Power plots comparing the DE detection power of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 1 where genes follow a birth and death process, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the nominal FDR. Higher curves indicate higher power of detecting DE genes.
Figure 4.2.4: Estimated FDRs comparing the true FDR control of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 1 where genes follow a birth and death process, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated FDR and the horizontal axis is for the nominal FDR. The yellow dash represents the estimated FDR exactly meeting the nominal FDR. Curves above the yellow dash indicate more false discoveries than stated, while curves below the yellow dash indicate less false discoveries than declared.
lower true FDRs for $N \geq 20$, while DESeq2 and edgeR become less conservative with more replicates, and TSPM has comparatively stable control of the FDR across different sample sizes. Except MixBD, the other four methods have higher true FDRs for standard Poisson simulations than for overdispersed Poisson, i.e., those four methods are more conservative for overdispersed genes counts (Figure 4.2.10). Controlling the FDR at the 5% level, we find that BDPM and MixBD have just faintly lower TPRs than the other three methods, especially for genes from a standard Poisson distribution, and the difference is indistinguishable when replicates number is as large as 40 (Table 4.2.2 - 4.2.4).

Table 4.2.2: Mean true positive rate (TPR) and mean true proportions of false positives among discoveries (V/R) across sample sizes from 10 simulations, including both overdispersed and standard Poisson simulations, of setting 2 at controlled 5% FDR via the BH method.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$N = 5$</th>
<th>$N = 10$</th>
<th>$N = 20$</th>
<th>$N = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BDPM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPR (SE)</td>
<td>0.567 (1.25%)</td>
<td>0.6520 (0.93%)</td>
<td>0.703 (1.13%)</td>
<td>0.833 (0.82%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.083 (1.02%)</td>
<td>0.0650 (0.84%)</td>
<td>0.049 (0.51%)</td>
<td>0.027 (0.56%)</td>
</tr>
<tr>
<td><strong>DESeq2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPR (SE)</td>
<td>0.730 (2.50%)</td>
<td>0.7900 (0.80%)</td>
<td>0.841 (0.80%)</td>
<td>0.883 (0.64%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.008 (0.21%)</td>
<td>0.0170 (0.40%)</td>
<td>0.022 (0.18%)</td>
<td>0.031 (0.74%)</td>
</tr>
<tr>
<td><strong>edgeR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPR (SE)</td>
<td>0.741 (1.17%)</td>
<td>0.8010 (0.71%)</td>
<td>0.844 (0.82%)</td>
<td>0.885 (0.60%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.029 (0.47%)</td>
<td>0.0300 (0.53%)</td>
<td>0.030 (0.42%)</td>
<td>0.032 (0.38%)</td>
</tr>
<tr>
<td><strong>TSPM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPR (SE)</td>
<td>0.715 (1.68%)</td>
<td>0.7930 (0.74%)</td>
<td>0.845 (0.81%)</td>
<td>0.887 (0.53%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.070 (1.02%)</td>
<td>0.0470 (0.80%)</td>
<td>0.045 (0.63%)</td>
<td>0.050 (1.03%)</td>
</tr>
<tr>
<td><strong>MixBD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPR (SE)</td>
<td>0.650 (2.20%)</td>
<td>0.7470 (0.67%)</td>
<td>0.811 (0.81%)</td>
<td>0.878 (0.57%)</td>
</tr>
<tr>
<td>V/R (SE)</td>
<td>0.131 (1.74%)</td>
<td>0.0960 (1.32%)</td>
<td>0.130 (1.17%)</td>
<td>0.422 (3.45%)</td>
</tr>
</tbody>
</table>
Figure 4.2.5: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 2 where data is simulated from a mixture of overdispersed and standard Poisson distributions, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.6: Power plots comparing the DE detection power of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 2 where data is simulated from a mixture of overdispersed and standard Poisson distributions, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the nominal FDR. Higher curves indicate higher power of detecting DE genes.
Figure 4.2.7: Estimated FDRs comparing the true FDR control of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 2 where data is simulated from a mixture of overdispersed and standard Poisson distributions, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated FDR and the horizontal axis is for the nominal FDR. The yellow dash represents the estimated FDR exactly meeting the nominal FDR. Curves above the yellow dash indicate more false discoveries than stated, while curves below the yellow dash indicate less false discoveries than declared.
Table 4.2.3: Mean true positive rate (TPR) and mean true proportions of false positives among discoveries (V/R) across sample sizes from 10 overdispersed Poisson simulations of setting 2 at controlled 5% FDR via the BH method.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$N = 5$</th>
<th>$N = 10$</th>
<th>$N = 20$</th>
<th>$N = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDPM</td>
<td>TPR (SE)</td>
<td>0.700 (2.00%)</td>
<td>0.732 (1.65%)</td>
<td>0.732 (1.90%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.032 (0.55%)</td>
<td>0.054 (1.35%)</td>
<td>0.046 (0.78%)</td>
</tr>
<tr>
<td>DESeq2</td>
<td>TPR (SE)</td>
<td>0.971 (0.49%)</td>
<td>0.982 (0.45%)</td>
<td>0.987 (0.60%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.006 (0.22%)</td>
<td>0.014 (0.51%)</td>
<td>0.021 (0.52%)</td>
</tr>
<tr>
<td>edgeR</td>
<td>TPR (SE)</td>
<td>0.968 (0.39%)</td>
<td>0.979 (0.44%)</td>
<td>0.982 (0.65%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.024 (0.59%)</td>
<td>0.024 (0.62%)</td>
<td>0.027 (0.57%)</td>
</tr>
<tr>
<td>TSPM</td>
<td>TPR (SE)</td>
<td>0.970 (0.45%)</td>
<td>0.986 (0.48%)</td>
<td>0.990 (0.54%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.030 (0.68%)</td>
<td>0.036 (0.92%)</td>
<td>0.042 (1.06%)</td>
</tr>
<tr>
<td>MixBD</td>
<td>TPR (SE)</td>
<td>0.844 (2.07%)</td>
<td>0.099 (1.78%)</td>
<td>0.147 (1.92%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.095 (2.75%)</td>
<td>0.099 (1.78%)</td>
<td>0.147 (1.92%)</td>
</tr>
</tbody>
</table>

Table 4.2.4: Mean true positive rate (TPR) and mean true proportions of false positives among discoveries (V/R) across sample sizes from 10 standard Poisson simulations of setting 2 at controlled 5% FDR via the BH method.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$N = 5$</th>
<th>$N = 10$</th>
<th>$N = 20$</th>
<th>$N = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDPM</td>
<td>TPR (SE)</td>
<td>0.442 (1.37%)</td>
<td>0.576 (1.18%)</td>
<td>0.676 (0.99%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.149 (2.21%)</td>
<td>0.078 (0.89%)</td>
<td>0.051 (0.73%)</td>
</tr>
<tr>
<td>DESeq2</td>
<td>TPR (SE)</td>
<td>0.489 (4.01%)</td>
<td>0.603 (1.04%)</td>
<td>0.700 (1.39%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.011 (0.61%)</td>
<td>0.020 (0.55%)</td>
<td>0.023 (0.69%)</td>
</tr>
<tr>
<td>edgeR</td>
<td>TPR (SE)</td>
<td>0.521 (1.66%)</td>
<td>0.627 (1.24%)</td>
<td>0.710 (1.56%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.039 (0.81%)</td>
<td>0.039 (0.77%)</td>
<td>0.035 (0.64%)</td>
</tr>
<tr>
<td>TSPM</td>
<td>TPR (SE)</td>
<td>0.475 (2.29%)</td>
<td>0.609 (1.04%)</td>
<td>0.707 (1.29%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.140 (1.72%)</td>
<td>0.062 (1.24%)</td>
<td>0.048 (0.72%)</td>
</tr>
<tr>
<td>MixBD</td>
<td>TPR (SE)</td>
<td>0.466 (2.39%)</td>
<td>0.587 (1.30%)</td>
<td>0.685 (1.12%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE)</td>
<td>0.187 (1.91%)</td>
<td>0.091 (1.22%)</td>
<td>0.102 (1.80%)</td>
</tr>
</tbody>
</table>
Figure 4.2.8: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), and TSPM (red) and MixBD (orange) based on 10 simulations from setting 2, considering genes from overdispersed (solid lines) and standard (dashed lines) Poisson distributions separately, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.9: Power plots comparing the DE detection power of BDPM (black), DESeq2 (green), edgeR (blue), and TSPM (red) and MixBD (orange) based on 10 simulations from setting 2, considering genes from overdispersed (solid lines) and standard (dashed lines) Poisson distributions separately, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the nominal FDR. Higher curves indicate higher power of detecting DE genes.
Figure 4.2.10: Estimated FDRs comparing the true FDR control of BDPM (black), DESeq2 (green), edgeR (blue), and TSPM (red) and MixBD (orange) based on 10 simulations from setting 2, considering genes from overdispersed (solid lines) and standard (dashed lines) Poisson distributions separately, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated FDR and the horizontal axis is for the nominal FDR. The yellow dash represents the estimated FDR exactly meeting the nominal FDR. Curves above the yellow dash indicate more false discoveries than stated, while curves below the yellow dash indicate less false discoveries than declared.
For Simulation 3, where gene counts follow a mixed birth and death process with Exponential
priors, when sample size is large enough (e.g. $N \geq 10$), MixBD have significantly higher
ability to rank genes in the order of differential expression than the other four methods, while
BDPM, DESeq2, edgeR, and TSPM have close ranking ability (Figure 4.2.11 and 4.2.12). In
other words, MixBD is most likely to rank the DE genes in front of non-DE genes among
the five approaches, with a comparatively high FDR (such as FDR > 0.005) and enough
replicates. Similarly, when $N \geq 10$, MixBD has the highest power to detect DE and BDPM
shows power lower than MixBD, but much higher than the three remaining methods, and the
difference in power increases with sample size, since DESeq2, edgeR and TSPM have stably
low power across different sample sizes (Figure 4.2.13). Based on ROC curves and power
plots, neither BDPM nor MixBD shows better performance than the other three methods,
if sample size is small (e.g. $N \leq 5$). From Figure 4.2.14 we can see MixBD have acceptable
control of the FDR across different sample sizes, while BDPM tends to underestimate FDR,
and the other three methods always overestimate it. Applying the Benjamini-Hochberg (BH)
procedure to control the FDR at the 5% level and comparing the TPR and the estimated
FDR across sample sizes, we can see for all five methods, the TPR is rising with sample
sizes, although the growth is not significant for methods: DESeq2, edgeR and TSPM, as
showed in Table 4.2.5. The true FDR tends to be greater than 5% for BDPM and MixBD,
while the true FDR stays stably low across different replicates for the other three methods.
Table 4.2.5: Mean true positive rate (TPR) and mean true proportions of false positives among discoveries (V/R) across sample sizes from 10 simulations of setting 3 (under Mixed BD model) at controlled 5% FDR via the BH method.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>N = 5</th>
<th>N = 10</th>
<th>N = 20</th>
<th>N = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDPM</td>
<td>TPR (SE) 0.004 (0.45%)</td>
<td>0.257 (0.80%)</td>
<td>0.497 (1.55%)</td>
<td>0.600 (0.79%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE) 0.092 (11.29%)</td>
<td>0.220 (1.51%)</td>
<td>0.214 (1.25%)</td>
<td>0.217 (0.69%)</td>
</tr>
<tr>
<td>DESeq2</td>
<td>TPR (SE) 0.222 (0.74%)</td>
<td>0.259 (0.99%)</td>
<td>0.261 (1.26%)</td>
<td>0.261 (0.85%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE) 0.014 (0.61%)</td>
<td>0.015 (0.68%)</td>
<td>0.021 (0.65%)</td>
<td>0.026 (0.50%)</td>
</tr>
<tr>
<td>edgeR</td>
<td>TPR (SE) 0.224 (0.63%)</td>
<td>0.259 (0.99%)</td>
<td>0.261 (1.26%)</td>
<td>0.261 (0.86%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE) 0.015 (0.59%)</td>
<td>0.015 (0.68%)</td>
<td>0.021 (0.65%)</td>
<td>0.026 (0.51%)</td>
</tr>
<tr>
<td>TSPM</td>
<td>TPR (SE) 0.098 (0.9%)</td>
<td>0.257 (1.05%)</td>
<td>0.264 (1.33%)</td>
<td>0.272 (1.01%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE) 0.019 (0.83%)</td>
<td>0.017 (0.85%)</td>
<td>0.023 (0.83%)</td>
<td>0.026 (0.57%)</td>
</tr>
<tr>
<td>MixBD</td>
<td>TPR (SE) 0.059 (1.61%)</td>
<td>0.572 (1.23%)</td>
<td>0.795 (0.77%)</td>
<td>0.864 (0.76%)</td>
</tr>
<tr>
<td></td>
<td>V/R (SE) 0.031 (2.27%)</td>
<td>0.130 (0.84%)</td>
<td>0.102 (0.60%)</td>
<td>0.118 (0.50%)</td>
</tr>
</tbody>
</table>
Figure 4.2.11: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 3 where genes follow a Mixed birth and death process with Exponential prior, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR smaller than 0.05. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.12: ROC curves comparing the DE ranking ability of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 3 where genes follow a Mixed birth and death process with Exponential prior, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the estimated FPR smaller than 0.015. Higher curves indicate higher ability to rank DE genes correctly.
Figure 4.2.13: Power plots comparing the DE detection power of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 3 where genes follow a Mixed birth and death process with Exponential prior, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated TPR and the horizontal axis is for the nominal FDR. Higher curves indicate higher power of detecting DE genes.
Figure 4.2.14: Estimated FDRs comparing the true FDR control of BDPM (black), DESeq2 (green), edgeR (blue), TSPM (red) and MixBD (orange) based on 10 simulations from setting 3 where genes follow a Mixed birth and death process with Exponential prior, with 5, 10, 20, and 40 replicates. The vertical axis is for the estimated FDR and the horizontal axis is for the nominal FDR. The yellow dash represents the estimated FDR exactly meeting the nominal FDR. Curves above the yellow dash indicate more false discoveries than stated, while curves below the yellow dash indicate less false discoveries than declared.
With $N = 5$, plotting histograms of difference and ratio between simulation parameters from one of ten simulations for setting 1, 2 and 3 separately, we can evaluate how big the expected counts difference is between two treatment groups. Figure 4.2.15 indicates that the simulated DE genes may not have adequately large difference in counts between two groups for Simulation 1, since the difference between simulation parameters is centered around 0, and the ratio has a high density around 1, which may lead to unsatisfying performance of DESeq2, edgeR, and TSPM in simulations based on the BD process. Histograms of Simulation 2 tends to provide similar indication, but the difference between two parameters is bigger, which may make it easier for finding DE genes simulated from a Poisson distribution for all methods.

In Simulation 3, the ratio does not concentrate on 1, but the difference between $\theta_\mu$ and $\theta_\lambda$ is still small, which may bring some difficulty in DE detection.

Since BDPM overperforms other DE detection methods to a large extent when gene counts follow a birth and death process, but not significantly inferior when genes are drawn from the mixture of standard and overdispersed Poisson distributions, especially for standard Poisson, it demonstrates satisfying ability of detecting DE genes when the distribution of gene counts is uncertain and may not be overdispersed Poisson or Negative Binomial distributions. This may also be a desirable property for analyzing RNA-seq data with little overdispersion, such as data with only technical replicates. Similarly, MixBD performs best for mixed-BD-simulated data, and keeps fair performance for BD or Poisson simulations. Thus, MixBD is also a good choice for DE detection, since there always exists uncertainty of the distribution of RNA-seq data.
Figure 4.2.15: Histograms displaying the distributions of difference and ratio between the simulation parameters for DE genes, where $\lambda t$ and $\mu t$ are birth and death parameters from the birth and death process in simulation 1, $\lambda_N$ and $\lambda_C$ are expected gene counts of the first and the second group respectively from the Poisson distribution in Simulation 2, and $\theta_\lambda$ and $\theta_\mu$ are the hyperparameters in the Exponential prior for birth and death rates separately from the mixed birth and death processes in Simulation 3.
Chapter 5

Application

We need to apply the proposed models: birth and death model (BDPM) and mixed birth and death model (MixBD) to the real data set, and compare their abilities of fitting and differential expression detection to these of models based on Negative Binomial (NB) or overdispersed Poisson distributions. Three sets of paired normal-to-tumor transcriptomic data, studying three types of cancer and covering three levels of sample size, are retrieved from the GEO database [40] with ID: TCGA data archive (https://tcga-data.nci.nih.gov/tcga/), and then used for the application study in this article.

The LUAD data contains 7 lung adenocarcinoma patients’ gene expression in both normal and tumor issue, covering 20532 genes with tissue source site 55 (International Genomics Consortium). The HNSC data includes the paired normal to tumor gene expression from 21 patients with head and neck squamous cell carcinoma from tissue source site CV (MD Anderson Cancer Center), containing 20532 genes in total. The BRCA data has 53 breast invasive carcinoma patients’ one-to-one normal and tumor tissues gene expression, including 20532 genes with tissue source site BH (University of Pittsburgh).
All the three data sets were produced from the paired-end sequencing on Illumina HiSeq 2000 sequencers, where reads were aligned to the reference transcriptome released on the University of California Santa Cruz (UCSC) Genome Browser website, with subsequent mapping to the whole genome UCSC hg19 based on GRCh37. Read counts and RPKM per composite gene was calculated using the SeqWare framework via the RNASeqAlignmentBWA workflow by researchers Dr. Neil D. Hayes, Dr. Charles M. Perou, Dr. Derek D. Chiang, Dr. George Wu, and Dr. Alan P. Hoyle from the University of North Carolina at Chapel Hill.

First, to get the data meaningful for applying the differential expression test, we need to remove genes non-expressed, i.e., genes have RPKM 0 across all patients. Second, to make the computing of a BD process efficient, we think it is appropriate to focus on genes with average RPKM less than 1000. Actually, the last step does not remove many genes. Only about 50 genes need to be removed, but the computing process is largely quickened. Third, since the BD process requires both previous (normal) and current (tumor) states to be integer, we have to round the RPKM values in each data set before analyzing. last, we decide to remove the extremely low-expressed genes, especially those having at least one patient with RPKM in normal tissue less than 0.5, i.e., rounded RPKM is 0. Because one of the requirements for a BD process is that the start state should not be 0, since it is impossible to have any births or deaths from 0.

After the data manipulation, we keep 13710 genes in the LUAD data with 7 patients, 10537 genes in the HNSC data with 21 patients, and 11833 genes in the BRCA data with 53 patients.
5.1 Goodness-of-fit Test

Before applying the proposed models and the traditional models to detecting differentially expressed genes, we want to see how well the assumed distribution fits the data and hence compare the fitting ability among those models. Thus, we decide to do the goodness-of-fit test to each data set first.

Considering the goodness-of-fit test, the first idea coming into mind is usually Pearson $\chi^2$ test. For Pearson $\chi^2$ test, we have to divide the ordered observations into $k$ mutually exclusive classes, and compare the frequency of the observed outcomes in each class to the frequency under the hypothesised distribution. The test statistic has an approximate $\chi^2$ distribution with $k - p - 1$ degrees of freedom under the null hypothesis, where $p$ is the number of estimated parameters. For NB, BD and mixed BD models, we need to estimate two parameters ($p = 2$). Thus, in order to get at least 1 degree of freedom, it requires $k \geq 4$, i.e., we have to divide the observations into at least 4 classes. On the other side, it requires the estimated frequency not below 5 to make the approximate $\chi^2$ distribution hold. Hence, for the LUAD and HNSC data with only 7 and 21 patients correspondingly, Pearson $\chi^2$ goodness-of-fit is not appropriate.

Then we consider the Kolmogorov-Smirnov test, which is one of the most popular nonparametric goodness-of-fit tests and can handle data with small sample size. The test statistics is:

$$D = \sup_x |F_0(x) - F_{\text{data}}(x)|,$$  \hspace{1cm} (5.1.1)

where $x$ is different data points, $F_0(x)$ is the cumulative distribution function under null hypothesis and $F_{\text{data}}(x)$ is the empirical distribution function. For both continuous and discontinuous distributions in a null hypothesis of the Kolmogorov-Smirnov (K-S) test, the formula of the test statistic $D$ stays the same. But the general Kolmogorov-Smirnov table (which is used for finding a $p$-value) is designed for a continuous distribution, and applying the
table to a discontinuous distribution may make the \( p \)-value conservative. Since our purpose is to compare the fitting of the proposed models with the traditional models, employing conservative \( p \)-values for all the five models may be acceptable and may not affect the comparison much.\cite{41}

To solve the problem of conservative \( p \)-values, we can use bootstrap method to decide the critical value for the corresponding K-S statistic. Suppose we want to do the K-S test for each gene at \( \alpha \) significance level. After getting the estimated parameters and K-S statistic, we can generate \( N \) bootstrap samples from the estimated parameters and the corresponding model, and hence calculate the K-S statistic for each of the \( N \) samples. With \( N \) bootstrap K-S statistics, we actually have an empirical distribution of the K-S statistic for the individual gene, and we can use the \((1 - \alpha)\) quantile value of the \( N \) K-S statistics as the critical value for the K-S test, which we can call a bootstrap critical value.

Thus, we decide to apply the Kolmogorov-Smirnov test with corresponding bootstrap critical values to Poisson, NB, BD and mixed BD models, to test how well they fit the three data sets. We tried \( N = 100 \) and \( N = 1000 \) for both Poisson and NB models, and we find there is not a big difference in the bootstrap critical value between the two bootstrap sizes, neither does the proportion of genes lack of fit. A larger bootstrap sample size, e.g. \( N = 1000 \), makes the percentage of lack-of-fit genes slightly lower, usually difference no more than 1%, and makes the fitting seem slightly better. We think \( N = 100 \) provides a bootstrap critical value correct enough, and to save computing time, we decide to use 100 bootstrap samples in the K-S test for the four types of distributions.

Else, we must notice that we can only get the conditional distribution for BD and mixed BD models, since the distribution of the gene expression in the tumor issue depends on the expression in the normal issue. We understand it is an advantage of our BD models, since they consider the change from normal to tumor and take the individual difference into account. In order to get an uniform cumulative distribution \( F_0(x) \) appropriate for calculating
the test statistic $D$, we decide to assign a common normal expression to all the patients, which is the mean of the normal RPKM across patients. Then, each patient’s gene expression in the tumor is modeled as changing from a same expression value in the normal, for which we are able to get $F_0(x)$ from the uniform conditional distribution. It is worth mentioning that in this way, the fitting ability of the BD model and the mixed BD model are reduced, since the individual difference cannot be measured. Considering this, we think the fitting of the real BD and mixed BD models should be better than what we get in the proposed test, since we replace the individual normal expression with the average expression in the proposed test.

For the LUAD data with 7 patients, after removing those non-expressed or low-expressed genes, we fit Poisson, Negative Binomial, BD model and mixed BD model separately to the 13710 genes, and get the below goodness-of-fit results based on the Kolmogorove-Smirnov test at a 0.05 significance level, where we use the bootstrap critical value from the 100 bootstrap samples, that is, we consider genes with the test statistic larger than the corresponding bootstrap critical value as lack of fit.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>3441</td>
<td>1747</td>
<td>1118</td>
<td>1443</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>25.10%</td>
<td>12.74%</td>
<td>8.15%</td>
<td>10.53%</td>
</tr>
</tbody>
</table>

Table 5.1.1: LUAD data with 7 patients: Kolmogorov-Smirnov goodness-of-fit test results.

After that, We want to go further to compare the fitting of a birth and death (BD) process, a mixed BD process, and a Negative Binomial (NB) distribution at different expression levels to see if there is any difference in the objective preference. We calculate the average gene expression from the 7 patients in the tumor group, and divide the 13710 genes into three categories with the average tumor expression less than 50, between 50 and 100, and larger than 100. Then we compare the number and corresponding proportion of genes not satisfying the fitted model within each category. In the following tables, NB stands for
Table 5.1.2: LUAD data with 7 patients: Kolmogorov-Smirnov test results within different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0, 50)</td>
<td>12522</td>
<td>1659</td>
<td>1026</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>688</td>
<td>55</td>
<td>50</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>500</td>
<td>33</td>
<td>43</td>
</tr>
</tbody>
</table>

Negative Binomial distribution, BD stands for birth and death process and MixBD stands for mixed birth and death process.

We also want to study among those lack-of-fit genes under one model, how many of them can be fitted by the other model. That is, the number of corresponding of genes which cannot be well fitted by the NB model, but can be measured using the BD model and the mixed BD model.

Table 5.1.3: LUAD data with 7 patients (NB v.s. BD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>1747</td>
<td>1591</td>
<td>91.07%</td>
</tr>
<tr>
<td>BD</td>
<td>1117</td>
<td>961</td>
<td>86.03%</td>
</tr>
</tbody>
</table>

Table 5.1.4: LUAD data with 7 patients (NB v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>1747</td>
<td>1579</td>
<td>90.38%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1443</td>
<td>1275</td>
<td>88.36%</td>
</tr>
</tbody>
</table>

Similarly, we test the fitting of models of interest on the 10537 analysis-appropriate genes in the HNSC data with 21 patients and the 11833 filtered genes in the BRCA data with 53 patients, using the corresponding bootstrap critical values from 100 bootstrap samples.
Table 5.1.5: LUAD data with 7 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>1117</td>
<td>985</td>
<td>88.18%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1443</td>
<td>1311</td>
<td>90.85%</td>
</tr>
</tbody>
</table>

Table 5.1.6: HNSC data with 21 patients: Kolmogorov-Smirnov goodness-of-fit test results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>4221</td>
<td>1472</td>
<td>829</td>
<td>1637</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>40.06%</td>
<td>13.97%</td>
<td>7.87%</td>
<td>15.54%</td>
</tr>
</tbody>
</table>

Table 5.1.7: HNSC data with 21 patients: Kolmogorov-Smirnov test results within different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0, 50)</td>
<td>9322</td>
<td>1363</td>
<td>730</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>682</td>
<td>53</td>
<td>58</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>533</td>
<td>56</td>
<td>41</td>
</tr>
</tbody>
</table>

Table 5.1.8: HNSC data with 21 patients (NB v.s. BD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>1472</td>
<td>1352</td>
<td>91.85%</td>
</tr>
<tr>
<td>BD</td>
<td>829</td>
<td>709</td>
<td>85.52%</td>
</tr>
</tbody>
</table>

From Kolmogorov-Smirnov test results, we find: 1) as expected, generally speaking, the fitting ability decreases when the sample size increases for all the four models, i.e., the proportion of genes lack of fit rises with the number of patients in the study (Table 5.1.1, 5.1.6 and 5.1.11). 2) Poisson model has a poor fit to the RNASeq data with biological replications, while NB model, BD model and mixed BD model fit the RNASeq data much
Table 5.1.9: HNSC data with 21 patients (NB v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>1472</td>
<td>1260</td>
<td>85.60%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1637</td>
<td>1425</td>
<td>87.05%</td>
</tr>
</tbody>
</table>

Table 5.1.10: HNSC data with 21 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>829</td>
<td>699</td>
<td>84.32%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1637</td>
<td>1507</td>
<td>92.06%</td>
</tr>
</tbody>
</table>

Table 5.1.11: BRCA data with 53 patients: Kolmogorov-Smirnov goodness-of-fit test results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>6008</td>
<td>2707</td>
<td>1435</td>
<td>3572</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>50.77%</td>
<td>22.88%</td>
<td>12.13%</td>
<td>30.19%</td>
</tr>
</tbody>
</table>

Table 5.1.12: BRCA data with 53 patients: Kolmogorov-Smirnov test results within different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0, 50)</td>
<td>10572</td>
<td>2420</td>
<td>1272</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>752</td>
<td>172</td>
<td>99</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>509</td>
<td>115</td>
<td>64</td>
</tr>
</tbody>
</table>

better than Poisson model. 3) The last three models (NB, BD and mixed BD) have reliable fittings to the studied data, but BD model explains the data best, especially when sample size is increasing, and NB fits large data slightly better than mixed BD. In detail, the difference between the number of BD’s lack-of-fit genes and NB’s increases with the sample size, and
Table 5.1.13: BRCA data with 53 patients (NB v.s. BD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2707</td>
<td>2367</td>
<td>87.44%</td>
</tr>
<tr>
<td>BD</td>
<td>1435</td>
<td>1095</td>
<td>76.31%</td>
</tr>
</tbody>
</table>

Table 5.1.14: BRCA data with 53 patients (NB v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2707</td>
<td>1910</td>
<td>70.56%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>3572</td>
<td>2775</td>
<td>77.69%</td>
</tr>
</tbody>
</table>

Table 5.1.15: BRCA data with 53 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from Kolmogorov-Smirnov test.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>1435</td>
<td>993</td>
<td>69.20%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>3572</td>
<td>3130</td>
<td>87.63%</td>
</tr>
</tbody>
</table>

NB almost doubles the number of lack-of-fit genes in BD for the 53-patient data (Table 5.1.1, 5.1.6 and 5.1.11). 4) Generally speaking, BD model remains the best fitting across the three RPKM ranges, however, this superiority becomes less obvious for higher gene expressions (RPKM > 50) under comparatively smaller sample size (7-patient and 21-patient data). In other words, BD and mixed BD models have stable fitting ability across different gene expression levels, while NB model performs better for comparatively highly expressed genes in the studied data (Table 5.1.2, 5.1.7 and 5.1.12). 5) For the three models: NB, BD and mixed BD, among genes that do not satisfy one model, there is a large amount, about 75% – 90%, of those genes fitting the other two models, and the proportion tends to decline as the sample size increases. Moreover, more genes unsatisfying NB model (Table 5.1.3).
or mixed BD model (Table 5.1.5, 5.1.10 and 5.1.15) can be explained by BD model, i.e., for each data set, the proportion of BD-fit genes among the NB-non-fit genes or the MixBD-non-fit genes is greater than that of the BD-non-fit but NB-fit or MixBD-fit genes.

It is sometimes controversial of using the Kolmogorov-Smirnov method for the goodness of fit test, since it is too conservative with measuring only the biggest difference between the hypothesized cumulative distribution and empirical cumulative distribution. Then we consider adjusting the Kolmogorov-Smirnov test by taking all the data points into account instead of only the point with the biggest difference in cumulative distribution. In that way, we can gain more power to detect genes lack of fit. A simple way is to add up the absolute differences across all the data points. Thus, using the sum, we get an adjusted Kolmogorov-Smirnov test with the following test statistic:

$$D_{\text{adjust}} = \sum_x |F_0(x) - F_{\text{data}}(x)|,$$

(5.1.2)

where $x$, $F_0(x)$ and $F_{\text{data}}(x)$ have the same interpretations as in Equation 5.1.1.

Calculating the adjusted K-S statistic from Equation 5.1.2 and following the previous process of K-S tests for genes in the LUAD data, the HNSC data and the BRCA data, we can get the results of adjusted K-S tests from 100 bootstrap samples. From the adjusted K-S tests results in Table 5.1.16 - 5.1.30 we can get similar conclusions as from the K-S tests, except that 1) the general fitting of mixed BD model is better than NB model’s (Table 5.1.16, 5.1.21 and 5.1.26); 2) the proportion of non-NB-fit genes which can be fitted by mixed BD model is higher than the proportion of non-MixBD-fit but NB-fit genes (Table 5.1.19, 5.1.24 and 5.1.29).

The possible reasons that the mixed BD model does not have fitting results good enough, like the BD model, as we expected, may be 1) the prior - Exponential distribution is
not appropriate; 2) the use of a common starting point (the average of normal gene counts
across patients) make the mixed BD model lose its ability to catch the individual information
(normal-cancer pairs), and for a model as complex as MixBD, it may affect the function
seriously.

Table 5.1.16: LUAD data with 7 patients: adjusted Kolmogorov-Smirnov goodness-of-fit
tests results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>3590</td>
<td>2898</td>
<td>1242</td>
<td>1546</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>26.19%</td>
<td>21.14%</td>
<td>9.06%</td>
<td>11.28%</td>
</tr>
</tbody>
</table>

Table 5.1.17: LUAD data with 7 patients: adjusted Kolmogorov-Smirnov tests results within
different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0, 50)</td>
<td>12522</td>
<td>2778</td>
<td>1141</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>688</td>
<td>66</td>
<td>54</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>500</td>
<td>54</td>
<td>47</td>
</tr>
</tbody>
</table>

Table 5.1.18: LUAD data with 7 patients (NB v.s. BD): genes not fitting one model but
fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2898</td>
<td>2640</td>
<td>91.10%</td>
</tr>
<tr>
<td>BD</td>
<td>1241</td>
<td>983</td>
<td>79.21%</td>
</tr>
</tbody>
</table>

Table 5.1.19: LUAD data with 7 patients (NB v.s. MixBD): genes not fitting one model but
fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2898</td>
<td>2593</td>
<td>89.48%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1546</td>
<td>1241</td>
<td>80.27%</td>
</tr>
</tbody>
</table>
Table 5.1.20: LUAD data with 7 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>1241</td>
<td>1101</td>
<td>88.72%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1546</td>
<td>1406</td>
<td>90.94%</td>
</tr>
</tbody>
</table>

Table 5.1.21: HNSC data with 21 patients: adjusted Kolmogorov-Smirnov goodness-of-fit tests results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>4675</td>
<td>2450</td>
<td>994</td>
<td>1629</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>44.37%</td>
<td>23.25%</td>
<td>9.43%</td>
<td>15.46%</td>
</tr>
</tbody>
</table>

Table 5.1.22: HNSC data with 21 patients: adjusted Kolmogorov-Smirnov tests results within different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0, 50)</td>
<td>9322</td>
<td>2276</td>
<td>879</td>
</tr>
<tr>
<td>(50, 100)</td>
<td>682</td>
<td>73</td>
<td>63</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>533</td>
<td>101</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 5.1.23: HNSC data with 21 patients (NB v.s. BD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2450</td>
<td>2213</td>
<td>90.33%</td>
</tr>
<tr>
<td>BD</td>
<td>994</td>
<td>757</td>
<td>76.16%</td>
</tr>
</tbody>
</table>

To visually see the fitting of NB, BD and mixed BD models, we can draw plots of cumulative distribution under the three models. For each data set, we randomly choose a gene satisfying all the three models and a gene satisfying none of the models, and fit NB, BD and mixed.
Table 5.1.24: HNSC data with 21 patients (NB v.s. MixBD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>2450</td>
<td>2064</td>
<td>84.24%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1629</td>
<td>1243</td>
<td>76.30%</td>
</tr>
</tbody>
</table>

Table 5.1.25: HNSC data with 21 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>994</td>
<td>833</td>
<td>83.80%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>1629</td>
<td>1468</td>
<td>90.12%</td>
</tr>
</tbody>
</table>

Table 5.1.26: BRCA data with 53 patients: adjusted Kolmogorov-Smirnov goodness-of-fit tests results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Poisson</th>
<th>NB</th>
<th>BD</th>
<th>Mixed BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes lack of fit</td>
<td>6721</td>
<td>4635</td>
<td>1993</td>
<td>3755</td>
</tr>
<tr>
<td>Proportion lack of fit</td>
<td>56.80%</td>
<td>39.17%</td>
<td>16.84%</td>
<td>31.73%</td>
</tr>
</tbody>
</table>

Table 5.1.27: BRCA data with 53 patients: adjusted Kolmogorov-Smirnov tests results within different RPKM ranges in terms of the average tumor expression.

<table>
<thead>
<tr>
<th>RPKM range</th>
<th>Total genes</th>
<th>No. of genes lack of fit</th>
<th>Proportion lack of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NB</td>
<td>BD</td>
</tr>
<tr>
<td>(0,50)</td>
<td>10572</td>
<td>4218</td>
<td>1781</td>
</tr>
<tr>
<td>(50,100)</td>
<td>752</td>
<td>250</td>
<td>132</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>509</td>
<td>167</td>
<td>80</td>
</tr>
</tbody>
</table>

BD models to each gene and get the cumulative distributions. Then we draw the empirical cumulative distribution curve and cumulative distribution curves from NB, BD and mixed BD models separately for each gene. From Figure 5.1.1 we can see that all the three models
Table 5.1.28: BRCA data with 53 patients (NB v.s. BD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>4635</td>
<td>3798</td>
<td>81.94%</td>
</tr>
<tr>
<td>BD</td>
<td>1993</td>
<td>1156</td>
<td>58.00%</td>
</tr>
</tbody>
</table>

Table 5.1.29: BRCA data with 53 patients (NB v.s. MixBD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>4635</td>
<td>3209</td>
<td>69.23%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>3728</td>
<td>2302</td>
<td>61.75%</td>
</tr>
</tbody>
</table>

Table 5.1.30: BRCA data with 53 patients (BD v.s. MixBD): genes not fitting one model but fitting the other from adjusted Kolmogorov-Smirnov tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of genes lack of fit</th>
<th>No. of genes fitting the other</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD</td>
<td>1993</td>
<td>1349</td>
<td>67.69%</td>
</tr>
<tr>
<td>Mixed BD</td>
<td>3728</td>
<td>3084</td>
<td>82.73%</td>
</tr>
</tbody>
</table>

have curves close to the empirical curve, especially when enlarging the sample size, although the fitting of mixed BD model for small-size data (e.g. LUAD data) is not as good as other models. Furthermore, the curves from NB and BD models are very close to each other in most cases, although not exactly coinciding, and we know that the cumulative probabilities from NB and BD models at each data point are extremely close. It worth noticing that, even though considering the non-fit genes, the cumulative distributions from the three models are not terribly deviated from the empirical distribution, and both K-S and adjusted K-S tests results indicate that the three models can still explain the data to a considerable extent under the non-fit situation, since the test statistic in the rejected situation is larger than but not much larger than the critical value.
Figure 5.1.1: Cumulative distribution curves for fit and non-fit genes in LUAD, HNSC and BRCA data.
5.2 Data Analyses

To test the performance of BDPM and MixBD, we considered the same data: LUAD, HNSC and BRCA. We performed the differential expression (DE) tests to these data sets in the same way as described in the section of Simulation Study: using RPKM for BDPM and MixBD, and raw counts for DESeq2, edgeR with default settings, and raw counts normalized with total number of reads per sample for TSPM.

The 7-patient BRCA data kept 13710 genes for analyses, from which 5779, 4792, 4735, 4391 and 2506 genes were detected as differentially expressed separately by DESeq2, BDPM, TSPM, edgeR and MixBD descendingly, where 1483 genes were detected by all the methods, 3105 shared by BDPM and the three BD-unrelated methods, and 1530 shared by MixBD and the three methods. Among the detected genes in DESeq2, TSPM and edgeR, 4332 (90%), 3713 (78%), and 3760 (86%) genes were also found by BDPM respectively (Figure 5.2.1), while 2357 (49%), 1738 (38%), and 2055 (47%) were shared with MixBD correspondingly (Figure 5.2.2). Comparing the two proposed methods, there are 2280 genes detected as DE by both BDPM and MixBD, which is 91% of the DE genes detected by MixBD, and 48% of BDPM’s (Figure 5.2.3).

In the HNSC data including 21 patients and 10537 genes, BDPM found the most DE genes (5622), followed by MixBD (5489), DESeq2 (5002), edgeR (4577), and TSPM (4537). Among those detected genes, DESeq2 had 4084 (82%) genes also detected by BDPM, and edgeR and TSPM had 3875 (85%) and 3647 (80%) DE genes shared with BDPM (Figure 5.2.1). Correspondingly, DESeq2, edgeR and TSPM shared 2940 (59%), 2721 (59%) and 2560 (56%) DE genes with MixBD (Figure 5.2.2). In general, all the five approaches found the same 2419 DE genes, and MixBD detected the most DE genes (1059) that were undetected for other approaches (Figure 5.2.3). Moreover, for the 4389 DE genes shared by the two BD-related methods, they are 78% and 80% of BDPM’s and MixBD’s detected DE genes respectively.
For the BRCA data of 53 patients, after filtering and manipulating, 11833 were remained for testing. Among the five methods, BDPM detected the most DE genes (8423), followed by DESeq2 (8360), edgeR (8194), TSPM (7840) and MixBD (7176) successively, where 5887 DE genes were shared by the five methods (Figure 5.2.3). Moreover, BDPM successfully found 7821 (94%), 7882 (96%), and 7562 (96%) DE genes from genes that were detected by DESeq2, edgeR, and TSPM respectively (Figure 5.2.1), while MixBD found 6880 (82%), 6801 (83%) and 5917 (75%) correspondingly (Figure 5.2.2). In addition, there are 6632 DE genes shared by both BDPM and MixBD, which occupies 79% and 92% of the DE genes detected from BDPM and MixBD separately.

Considering the analyses of the three data sets together, all the five methods detected more DE genes when sample sizes (the number of patients in this cancer study) increased. On one hand, both BDPM and MixBD shared most of DE genes with the three traditional methods, which supports the credibility and the stability of the two BD-related methods. On the other hand, both proposed methods discovered a considerable amount of DE genes that could not be detected by the three remaining methods, which can be explained by the completely different distribution included in BDPM and MixBD. That means, with the proposed methods - BDPM and MixBD, we are able to find some new information from genes and discover some new genes with importance we never realized before. In addition, there are some difference between the results from BDPM and MixBD, which makes MixBD seem not to be as stable as BDPM. A possible answer is the mixed model may highly depend on the prior distribution, and different data may need different priors.
Figure 5.2.1: Venn diagrams of the number of differentially expressed genes detected by BDPM, DESeq2, edgeR and TSPM, controlling false discovery rate at 5% level with Benjamin-Hochberg method. The four DE approaches were employed in LUAD (7 patients), HNSC (21 patients) and BRCA (53 patients) data, with 13710, 10537 and 11833 genes separately.

Figure 5.2.2: Venn diagrams of the number of differentially expressed genes detected by MixBD, DESeq2, edgeR and TSPM, controlling false discovery rate at 5% level with Benjamin-Hochberg method. The four DE approaches were employed in LUAD (7 patients), HNSC (21 patients) and BRCA (53 patients) data, with 13710, 10537 and 11833 genes separately.
Figure 5.2.3: Venn diagrams of the number of differentially expressed genes detected by BDPM, MixBD, DESeq2, edgeR and TSPM, controlling false discovery rate at 5% level with Benjamin-Hochberg method. The four DE approaches were employed in LUAD (7 patients), HNSC (21 patients) and BRCA (53 patients) data, with 13710, 10537 and 11833 genes separately.
Chapter 6

Discussion

The birth and death model (BDPM) presented in this dissertation comes with some assumptions to simplify the modeling and computing process. First, the change between two treatment groups is considered as a birth and death process. This assumption requires the data to be paired, i.e., observations of an individual should be available in both groups. It is sensible in many cases, such as cancer study, comparison between stem and root of the same plant. Take cancer study for example, it is reasonable to consider the change of DNA copy numbers or gene expressions from the normal tissue to the tumor tissue following a birth and death process. Second, for each biological entity (e.g., a gene) in the data, all the individuals are assumed to follow the same BD process, i.e., they share the same birth and death rates. This is similar to the assumption behind other DE testing models, where gene read counts are considered coming from the common Negative Binomial or overdispersed Poisson distributions. Third, individuals are assumed independent. With this assumption, we can get the joint likelihood through multiplying the individual BD process probabilities of individual BD process, and the parameter estimating becomes possible. Back to the cancer example, it is acceptable to make an assumption of independent patients to simplify the problem.
As a generalization of BDPM, the mixed birth and death processes model with Exponential prior (MixBD) has similar assumptions. First, changing from the first group to the second group is assumed as a mixture of several birth and death processes with a common prior. Second, under each biological entity, this model assume all the individuals follow a common mixed BD process. Third, individuals are considered independent. Fourth, the prior is selected as Exponential distribution, where parameters birth rate and death rate come from Exponential distributions.

A distinct advantage as well as a main characteristic of both BDPM and MixBD is that it is based on the birth and death process, instead of some frequently used distributions like Poisson and Negative Binomial. In the homogeneous BD process, we assume the birth and death rates do not change for a given gene. Because we are unable to observe the expression of the given gene at different time points. That is, we cannot model birth and death rates at different time. Thus, we use the overall birth and death rates in BDPM. According to the simulation results, BDPM and MixBD show good performance in data generated from all the three simulation settings: a BD process, a mixed BD process and a mixture of standard and overdispersed Poisson distribution, while other methods examined (DESeq2, edgeR, TSPM) perform slightly better for overdispersed-Poisson-simulated gene counts, but much worse than BDPM in detecting differentially expressed genes if data is obtained from a BD or a mixed BD process. Most of the parametric DE testing methods is developed from a Poisson distribution, with Negative Binomial as an extension to account the overdispersion. Because the mapping process in the sequencing technology can be theoretically consider a Poisson distribution. However, no work has been done to prove that gene counts with biological replicates follow a Negative binomial or overdispersed Poisson distribution. From the goodness-of-fit tests on the RNA-seq data of lung adenocarcinoma, head and neck squamous cell carcinoma, and breast invasive carcinoma, we find that the BD process fit the data better than the NB distribution, which at least indicates that there are some real data modeled better with a BD process, although other data may still fit better under Poisson or NB distributions.
(The mixed BD process with an Exponential prior does not perform a fitting as good as the BD process, although not worse than the NB distribution. This is possibly because of the assumption of Exponential prior not appropriate for the selected data.) Thus, we think BDPM and MixBD have a bright prospect of application in RNA-seq data or other related technologies, considering its reliable DE detecting behavior under different simulated distributions.

Another important advantage of BDPM and MixBD is their ability to model the individual change. None of the current DE testing methods, except the proposed one in this dissertation, takes the individual information into consideration, even given the paired data. Most methods treat the replicates within a group as random samples from a distribution (some nonparametric approaches analyze those replicates as a group directly), and then compare the expected means between different groups. There is a waste of information if we directly ignore the individual relationship between two groups. It is a great strength of BDPM and MixBD able to take advantage of the individual message in the paired data. Take cancer study for example, it is very possible that the gene expressions vary across different patients (that is, some individuals may have higher/lower gene expressions in both normal and tumor groups compared to other individuals), while the pattern of change from normal to tumor remains consistent. Thus, it is more reasonable to do a two-class comparison by modeling the individual change, like what BDPM and MixBD do.

Since BDPM models the group-to-group move individually, unlike those BD-unrelated methods, it does not have to deal with the problem of overdispersion within a group. In the simulation study, BDPM performs reasonably on simulations from both standard and overdispersed Poisson distributions, and other methods examined cannot exceed BDPM in DE detection if the actual distribution is a standard Poisson. BDPM’s stable performance with or without the existence of overdispersion, may make it more appropriate than some statistical methods designed for overdispersed distributions (e.g., DESeq2 and edgeR). Because for some data,
there is a considerable amount of real counts lack of overdispersion. Slightly different from BDPM, MixBD considers a mixture of several BD processes with the birth and death parameters following a prior - Exponential distribution, which may account for the overdispersion in the data to some extent. Because of this, from the simulation results of setting 2, we can see MixBD is more like DESeq2, edgeR and TSPM, performing better for overdispersed data.

BDPM and MixBD themselves do not include any normalization procedure. So they can accommodate different normalizing methods. For the real data analyses, RPKM is used because it can be directly retrieved from the GEO database and there are not many extremely large values after this normalization. BDPM and MixBD can work with other normalization techniques besides RPKM, if most values are within a 1 to 1000 range after normalization and rounding. This will not be a problem of concern, if some more appropriate data are used, such as DNA copy numbers, instead of gene expressions. Other types of data suitable for BDPM and MixBD are expected to come in the near future, with the constantly updated technology.

As the proposed models considering a totally different distribution from traditional models’, BDPM and MixBD also have some limitations in application. First, the BD-related methods only work for two-class comparison. In other words, BDPM and MixBD can only be used to test the difference between two groups, since the used stochastic process measures the change between two stages. However, this will not make BDPM or MixBD less competitive, considering most of the biological studies focus on two groups, such as the cancer study which we are most interested in applying our methods to. Second, the input data should be integers and not too large (e.g., > 10,000), due to the properties of the BD process. Extremely large values will make the proposed methods, especially MixBD, work slowly. Thus, for the RNA-seq data we use, we have to do the rounding first, and then remove genes with extremely high expressions. But there are not many genes like that after normalization, and using the
DNA copy number data can avoid the inconvenience of rounding and removing. With the development of new technologies, we believe that more accurate and smaller-valued data will be accessible, and there will be no difficulty of applying BDPM and MixBD. Third, as with many DE testing methods, in order to apply BDPM and MixBD, the sample size should be large enough to provide reasonable estimates of the parameters. In order to estimate the birth and death rates of the BD process, or the hyperparameters in the mixed BD processes, there should be at least three replicates in each group to get one degree of freedom. As the cost of sequencing gets lower, obtaining large sample sizes should not be a problem. Moreover, for many biological studies (e.g., cancer study), enough replicates are very necessary for drawing conclusions.

The simulation study implies that when there are no more than five replicates, like TSPM, BDPM may overestimate the significance for data actually coming from a standard Poisson distribution, compared to DESeq2 and edgeR. The performance of BDPM is more stable than other methods examined for BD-process-simulated data across different sample sizes, and the problem of overestimated significance disappears when the underlying distribution is overdispersed Poisson. We need pay attention to that BDPM tends to provide slight over-estimates of significance even though the true change satisfies a BD process, and the over-estimation of significance becomes serious for a simulated mixed BD process. Thus, we think BDPM can perform well if the actual distribution of data is a BD process or Poisson, show acceptable performance for a mixed BD process. MixBD works as good as BDPM for the simulation of a BD process, and performs best among all methods for the mixed-BD-simulated data, but have serious over-estimating of significance for Poisson-simulated data with sample size greater than 20. Else, similar to BDPM, MixBD has a problem of over-estimation under its own simulation. Considering these, we think MixBD can work fair if data actually follows a BD or mixed BD process, and handle medium-size data from Poisson distribution.
Chapter 7

Future work

1. The BDPM considers a simplified case by directly applying a simple birth and death process between two groups. Because the probability expression derived in section 2.1, which includes a Poisson distribution to measure the difference between the true and observed gene expressions, is too complicate and inappropriate for computation. Thus, we want to improve our model from the biological aspect and add other biology information into our model, to make BDPM fit the actual situation better.

2. We are interested in applying the birth and death process to the study of clustering. Take the cancer data for example, if we have the knowledge of cancer stages, we can model a BD process to study the pattern within each stage. Then, with a new patient coming, given his or her information, we can decide which stage that patient belongs to. If we have no information of the stages, we can employ a BD process to clustering the patients into different cancer stages.

3. We hope to try other prior distributions for the mixed BD model, and find a prior better than Exponential distribution to fit the RNAseq data.
4. We plan to optimize the computing algorithm used in our programs, to make BDPM and MixBD work faster and more efficient. We also want to write packages for BDPM and MixBD, to make them user-friendly. If the planned studies of generalization and clustering are successful, we need to provide a package including all the optional methods, and offer the users more choices to fit a variety of data.

5. We want to employ BDPM, MixBD and other DE detecting methods to other data, to study their performance in application. Different methods may be appropriate for different types of data. Thus, we may want to try various forms of data and find one that BDPM works on best and one MixBD works on best. This exploration may also help predicting the performance of BDPM and MixBD on future data from new technology.
Bibliography


