DISSERTATION

STATISTICAL MODELING AND INFERENCE FOR COMPLEX-STRUCTURED COUNT DATA WITH APPLICATIONS IN GENOMICS AND SOCIAL SCIENCE

Submitted by

Meng Cao

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2020

Doctoral Committee:

Advisor: Wen Zhou
Co-Advisor: F. Jay Breidt

Don Estep
Mary C. Meyer
Graham Peers

ABSTRACT

STATISTICAL MODELING AND INFERENCE FOR COMPLEX-STRUCTURED COUNT DATA WITH APPLICATIONS IN GENOMICS AND SOCIAL SCIENCE

This dissertation describes models, estimation methods, and testing procedures for count data that build upon classic generalized linear models, including Gaussian, Poisson, and negative binomial regression. The methodological extensions proposed in this dissertation are motivated by complex structures for count data arising in three important classes of scientific problems, from both genomics and sociological contexts. Complexities include large scale, temporal dependence, zero-inflation and other mixture features, and group structure.

The first class of problems involves count data that are collected from longitudinal RNA sequencing (RNA-seq) experiments, where the data consist of tens of thousands of short time series of counts, with replicate time series under treatment and under control. In order to determine if the time course differs between treatment and control, we consider two questions: 1) whether the treatment affects the geometric attributes of the temporal profiles and 2) whether any treatment effect varies over time. To answer the first question, we determine whether there has been a fundamental change in shape by modeling the transformed count data for genes at each time point using a Gaussian distribution, with the mean temporal profile generated by spline models, and introduce a measurement that quantifies the average minimum squared distance between the locations of peaks (or valleys) of each gene's temporal profile across experimental conditions. We then develop a testing framework based on a permutation procedure. Via simulation studies, we show that the proposed test achieves good power while controlling the false discovery rate. We also apply the test to data collected from a light physiology experiment on maize.

To answer the second question, we model the time series of counts for each gene by a Gaussian-Negative Binomial model and introduce a new testing procedure that enjoys the optimality property

of maximum average power. The test allows not only identification of traditional differentially expressed genes but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and expose its good performance via simulation studies. The procedure reveals interesting biological insights when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of a marine diatom.

The second class of problems involves analyzing group-structured sRNA data that consist of independent replicates of counts for each sRNA across experimental conditions. Most existing methods—for both normalization and differential expression—are designed for non-group structured data. These methods may fail to provide correct normalization factors or fail to control FDR. They may lack power and may not be able to make inference on group effects. To address these challenges simultaneously, we introduce an inferential procedure using a group-based negative binomial model and a bootstrap testing method. This procedure not only provides a group-based normalization factor, but also enables group-based differential expression analysis. Our method shows good performance in both simulation studies and analysis of experimental data on roundworm.

The last class of problems is motivated by the study of sensitive behaviors. These problems involve mixture-distributed count data that are collected by a quantitative randomized response technique (QRRT) which guarantees respondent anonymity. We propose a Poisson regression method based on maximum likelihood estimation computed via the EM algorithm. This method allows assessment of the importance of potential drivers of different quantities of non-compliant behavior. The method is illustrated with a case study examining potential drivers of non-compliance with hunting regulations in Sierra Leone.

ACKNOWLEDGEMENTS

Foremost, I would like to sincerely express my gratitude to my dissertation advisers Dr. Wen Zhou and Dr. F. Jay Breidt, for their continuous support, endless discussion and inspirational mentorship throughout my Ph.D. studies. Without their guidance and encouragement, this dissertation would not be possible.

Thank you to my committee members Dr. Don Estep, Dr. Mary C. Meyer, and Dr. Graham Peers for their continuous interest in my research and generously offering their time, support and guidance throughout the preparation and review of this manuscript. A special thank to Dr. Mary C. Meyer for the fruitful discussion and suggestions on the identifiability of the algorithm.

I would also like to thank Dr. Graham Peers, Dr. Tai Montgomery, Dr. Peng Liu, Dr. Jennifer N. Solomon and Dr. Michael C. Gavin for providing experiment data sets, insightful discussions and great collaborations.

Nobody has been more important to me in my Ph.D. studies than my family. I will forever be thankful to my parents for their love, support and encouragement. And my warmest thanks to my husband Xinshuo for his endless support and inspiration. Also, thank you Hannah, my little one. Thank you for using crying and screaming to keep me working.

DEDICATION

I would like to dedicate this dissertation to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

This dissertation describes models, estimation methods, and testing procedures for count data that build upon classic generalized linear models, including Gaussian, Poisson, and negative binomial regression. The methodological extensions proposed in this dissertation are motivated by complex structures for count data arising in three important classes of scientific problems, from both genomics and sociological contexts. Complexities include large scale, temporal dependence, zero-inflation and other mixture features, and group structure.

The first class of problems involves count data arising from RNA sequencing (RNA-seq) experiments [4,5]. Recently, such experiments have been extended to longitudinal studies. Longitudinal RNA-seq experiments can generate large-scale time-course count data, consisting of short time series of counts for each gene, with replicate time series under treatment and under control. The resulting data set consists of tens of thousands of these short time series. It is of scientific interest to study the dynamic patterns of gene expression: to describe the temporal profile for each gene and determine if the time course differs between treatment and control ("differential expression," or DE). Most existing tests are designed to distinguish among conditions based on overall differential patterns across time, though in practice, a variety of complex hypotheses are of more scientific interest. For example, it may be of interest to decide if any treatment effect varies over time ("non-parallel differential expression," or NPDE) or not ("parallel differential expression," or PDE). As another example, it may be of interest to study if the treatment affects the geometric attributes of the temporal profiles, such as the locations of peaks or valleys. Chapter 2 and 3 of this dissertation describe methods for this first class of problems.

In Chapter 2, we consider a novel class of hypotheses that involves geometric attributes of the temporal profiles for each gene. In this setting, we are not interested in overall location shifts or

rescaling of temporal profiles, but in fundamental changes in shape such as moving a peak or valley (for example, $\sin(x)$ and $a\sin(x)+b$ differ only by location shifts and scaling, while $\cos(x)$ has a fundamental change in shape).

To determine whether there has been a fundamental change in shape, we model the transformed count data for genes at each time point using a Gaussian distribution, with the mean temporal profile generated by spline models, and introduce a measurement that quantifies the average minimum squared distance between the locations of peaks (or valleys) of each gene's temporal profile across experimental conditions. We develop a testing framework based on the proposed model via a permutation procedure [6]. This test achieves good power while controlling the false discovery rate (FDR), which is demonstrated by simulation studies and by applying the test to data collected from a light physiology experiment on maize.

Chapter 3 also addresses the first class of problems. We consider hypotheses that identify the NPDE and PDE genes in Chapter 3, as NPDE genes may be of more scientific interest than PDE genes. For example, Sun et al. [7] showed that NPDE genes may provide more information on how the cell responds differently to different treatments. Conditional on a latent Gaussian mixture with evolving means, we model the data (time series of counts for each gene) by negative binomial distributions, introduce a general testing framework based on the proposed model and show that the proposed test enjoys the optimality property of maximum average power. The test allows not only identification of traditional differentially-expressed genes but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and demonstrate its good performance via simulation studies. The procedure reveals interesting biological insights when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of a marine diatom.

The second class of problems arises from the analysis of small RNA (sRNA) data sets. The data consist of independent replicates of counts for each sRNA, with replicates under treatment conditions and under control conditions. In a given experiment, there are tens of thousands of sR-

NAs, and these sRNAs belong to a small number of distinct classes. Similar to RNA-seq analysis, the major goal of sRNA analysis is to detect DE genes across treatments. Because of the multiple classes of sRNAs, both group effect and gene-specific effect may be of scientific interest. For example, Hackenberg et al. [8] found that distinct classes of sRNAs respond differently to drought: the miRNA and rasiRNA groups were down-regulated, while the tsRNAs were up-regulated under drought conditions. Most existing DE analysis methods are designed without considering potential group structure. These methods may lack power, fail to control FDR or not be able to make inference on group effects.

Further complication for the group-structured data is normalization. Existing normalization methods that matches the overall empirical distributions across sample [9] or equates the overall expression levels of genes between samples [10] may fail to provide correct normalization factors for group structured data. In addition, they assume that the majority of genes are not differentially expressed, which is not always true in the sRNA studies due to the group effects.

In Chapter 4, we develop a group-based negative binomial model for sRNA data analysis to address the above challenges. We introduce a testing framework based on the proposed model via bootstrap. We implement the proposed test by developing an efficient algorithm using a weighted generalized linear model. This procedure not only provides a group-based normalization factor, but also conducts group-based differential analysis. To examine the performance of our new method, we perform comprehensive simulation studies. To demonstrate advantages of the new analytical approach for the analysis of sRNA data, we applied it to an experiment to explore the roles of different classes of sRNAs in the roundworm (C. elegans).

Finally, the third class of problems also involves count data and uses inferential procedures developed from generalized linear models, but the scientific motivation comes from the study of sensitive behaviors using randomized response techniques (RRT). These techniques provide anonymity to interviewees who answer sensitive questions. A variation on this approach, the quantitative randomized response technique (QRRT), allows researchers to estimate the frequency or quantity of sensitive behaviors. Researchers are particularly interested in identifying potential

drivers of non-compliant behavior using regression methods. The result data consist of independent, nonnegative counts that are collected using QRRT.

Most existing regression-based methods are developed for multi-category randomized response data, and regression methodology has not been developed for count data from QRRT. To analyze the nonnegative count data produced by QRRT, we develop in Chapter 5 a Poisson regression methodology for QRRT data, based on maximum likelihood estimation computed via the expectation-maximization (EM) algorithm. The method is illustrated with a case study examining potential drivers of non-compliance with hunting regulations in Sierra Leone.

1.2 Outline

The four topics mentioned above have been addressed in four separate chapters of this dissertation, each of which is a co-authored paper. As of this writing, two of these papers have been published in refereed journals.

Each chapter is independent and can be read by itself. This dissertation is organized as follows.

- Chapter 2 is based on the paper: Cao, M., Zhou, W., Liu, P., & Brutnell, T. Differential analysis on dynamical patterns of time course gene expression data.

- Chapter 3: The material in this chapter has appeared as [84]: Cao, M., Zhou, W., Breidt, F. J., & Peers, G. (2019). Large scale maximum average power multiple inference on time-course count data with application to RNA-seq analysis. *Biometrics*. doi.org/10.1111/biom.13144.

- Chapter 4 is based on the paper: Cao, M., Zhou, W., Breidt, F. J., & Montgomery, T. Group structured model with application to small RNA analysis: normalization and differential expression analysis.

- Chapter 5: The material in this chapter has appeared as [142]: Cao, M., Breidt, F. J., Solomon, J. N., Conteh, A., & Gavin, M. C. (2018). Understanding the drivers of sensitive behavior using Poisson regression from quantitative randomized response technique data. *PLOS ONE*, 13(9), e0204433.

The final chapter briefly summarizes the contributions and discusses possible future works.

Chapter 2

Differential Analysis on Dynamical Patterns of Time Course Gene Expression Data

2.1 Introduction

2.1.1 Background and Related Work

Time course RNA-sequencing (RNA-seq) experiments are different from traditional RNA-seq experiments. The gene expression levels are correlated or admit some continuously varying structure over time. This temporal structure provides a more complete picture of general mechanisms yet brings some challenges for modeling and analysis [11]. For instance, the time course sampling is often sparse and irregular due to experimental constraints and true hypotheses of interest are usually composite.

As an example, consider a time course RNA-seq experiment to study the physiology of maize leaves. The objective of this study is to assess the effects of light on different locations on a leaf, particularly, the experiments consider the base, ligule from 3 to 4 cm, ligule from 8 to 9 cm, and the tip position on a leaf (see Figure 2.3). In this experiment, 13 time points at each leaf location are sampled for two different experimental light conditions, constant light and constant dark. At each time point, three biologically independent leaves are randomly picked from genetically identical plants. In total, expression levels of $63,293$ genes were collected from each of the four different leaf sections. The study attempts to identify genes that react to light differently across different leaf sections.

The primary goals of a RNA-seq analysis are 1) identifying differentially expressed (DE) genes; 2) identifying and characterizing changes in the gene expression over time [11, 12]. Efron et al. [13], Eckel et al. [14], and Aryee et al. [11] introduced the univariate empirical Bayes framework for detecting DE genes. As an extension of the univariate model, Tai et al. [15] constructed a

multivariate empirical method, which is applicable to data with both single and multiple conditions. Alternative approaches to account for the temporal structure and to identify DE genes are based on clustering methods [16–18]. However, Xu et al. [19] pointed out that clustering-based methods are easily influenced by the choice of transformations or filtering processes and some existing methods are not able to draw statistical inference.

Lately, researchers are more interested in whether the expression levels of the same gene share similar temporal patterns across different conditions. Genes with changing temporal structure are more interesting than genes with only scaling change. Statistically, it demands more detailed characterization of the changes of expression levels in a gene over time. To address this problem, spline-based methods have been previously proposed [20–22]. Storey et al. [12] extended spline-based method and proposed a method specifically designed for time course experiments. This approach focuses on identifying genes with inconsistent changes on the expression levels over time. A hidden Markov model was developed [23], that focuses on modeling the time dependency. This approach, however, requires the Markov property, which is always hard to be justified in practice. Using a negative binomial mixed-effect model, Sun et al. [7] identified genes with nonparallel and parallel expression profiles which change over time across treatments. They also discussed the importance of identifying the local pattern changes in the mean expression level other than the overall temporal consistency.

The aforementioned methods for analyzing the pattern changes in the temporal expression levels do not address how to capture the differences of the local geometry within the temporal pattern between different conditions, such as shift of peaks or valleys. In fact, they mostly focus on those consistently up-regulated or down-regulated expression profile changes over time across conditions.

2.1.2 Simulated Data

To motivate the new analytical approach, we consider a small simulation to display the comparison of performance of the functional analysis of variance (FANOVA) [24] and our method.

Consider a two-sample local curvature comparison, i.e. comparing the local convexity and concavity of temporal dynamics between two profiles. For each time point t, we generate data $Y_{ijr}(t)$ for replicate r in group j under condition i as following. Given continuous functions $\mu_{ij}(t)$, $Y_{ijr}(t) = \mu_{ij}(t) + \epsilon_{ijr}(t)$, where $i = 1, 2$, $j = 1, 2, 3$, $r = 1, 2, 3$ and $t = t_1, \ldots, t_{T_i}$. Specifically, we set $\mu_{1j}(t) = f_1(t)$ for each j, where $f_1(t) = -3\cos\{(x - \eta_1)/3 \times 2\pi\}$, and let $\mu_{2j}(t) = f_1(t)\mathbb{I}(j = 1) + f_2(t)\mathbb{I}(j = 2) + f_3(t)\mathbb{I}(j = 3)$, where $f_2(t) = -8\cos\{(x - \eta_1)/3 \times 2\pi\}$; and $f_3(t) = -3\cos\{(x - \eta_1)/3 \times 2\pi\}\mathbb{I}(x \leq 9/2 + \eta_1) + (4.47x - 16.72)\mathbb{I}(x > 9/2 + \eta_1)$. Here, $\mathbb{I}(\cdot)$ is the indicator function. Set, $\epsilon_{ijr}(t) \overset{\text{iid}}{\sim} t_{12}$ and $\eta_1 \overset{\text{iid}}{\sim} N(0, 0.5)$. In total, $2,000$ genes are generated, where 60% of genes are from group $j = 1$; 20% of genes are from group $j = 2$. The remaining genes are from group $j = 3$. Therefore, in total, 80% genes share the same locations of peaks and valleys and are considered as the non differentially expressed genes. See Fig 2.1 for an illustration of the mean patterns.

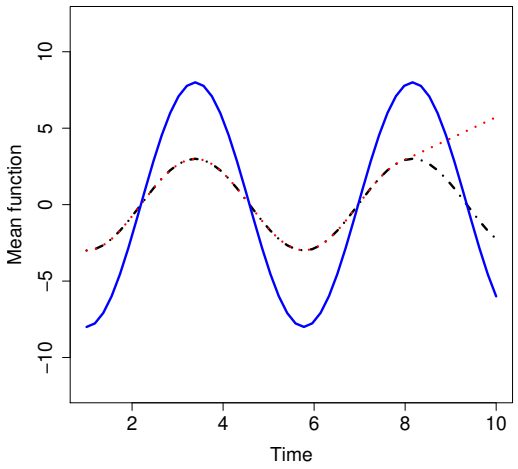
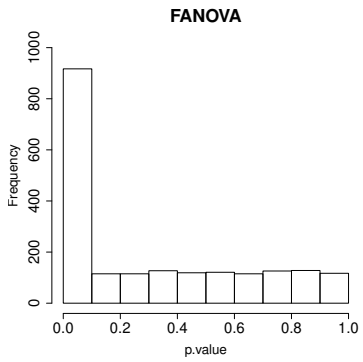
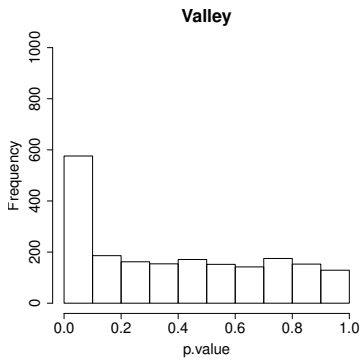


Figure 2.1: Mean Functions: $f_1(t)$: dash-dotted in black, $f_2(t)$: solid line in blue, $f_3(t)$: dot line in red.

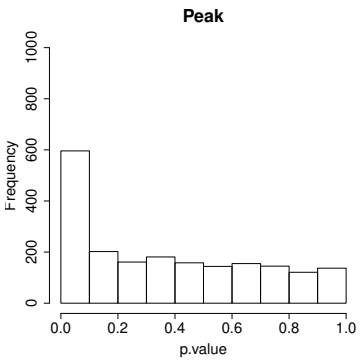
We employed the method to estimate and detect the temporal differences in terms of the locations of peaks and valleys across sections. Fig 2.2 shows the p-value histograms of both FANOVA and our approach. In the convexity and concavity comparisons, we observe that the histogram of p-values of the null genes for FANOVA, shown in Fig 2.2d, does not display the expected uniform



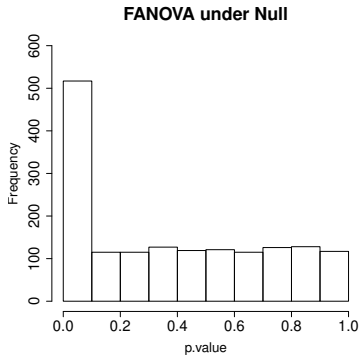
(a) p-value from Functional ANOVA



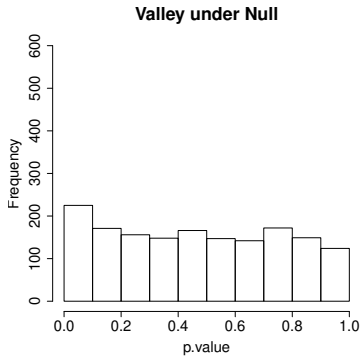
(b) p-value (Valley) from our method



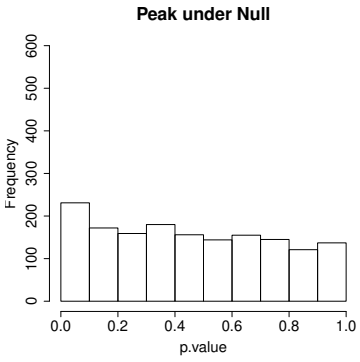
(c) p-value (Peak) from our method



(d) p-value under null from Functional ANOVA



(e) p-value under null (Valley) from our method



(f) p-value under null (Peak) from our method

Figure 2.2: Comparison of p-values. The top row shows the p-value histogram for all data, while the second row shows the p-value histograms for data under the null hypothesis with 1000 permutation.

distribution. This suggests that FANOVA fails to control the false discovery rate (FDR). Compared with FANOVA, the p-values, under null hypothesis of our method for both convexity (valley) and concavity (peak) comparisons are uniformly distributed. It implies that our method controls the FDR (see Fig 2.2e, 2.2f).

The primary goal of this chapter is to define a metric to quantify the local geometry of temporal gene expression profile and propose a reliable test procedure to identify DE genes with differential local geometries between conditions while controlling the FDR. Permutation is employed to facilitate the proposed method while the justification on the proposed metric is detailed in Section 2.2. We model the data using an additive model where limited assumptions are imposed to the mean temporal profile other than regular smoothness conditions. A comprehensive simulation study is

performed in Section 2.3 to examine our method. As an example, we apply the method to analyze time-course diurnal expression data collected from a light physiology experiment on maize. Some biologically interesting genes are discovered and an enrichment analysis is conducted. More technical results and further simulation studies are provided in Appendix A.

2.2 Methodology

Denote $\mathbf{Y}_{gir}(t)$ the reads per kilobase of transcript, per million mapped reads (RPKM) of gene g from the ith condition of the rth replicate at time point t, where $g = 1, \ldots, G$, $i = 1, 2$, $r = 1, \ldots, n_i$ and $t = t_1, \ldots, t_{T_i}$ with integers $T_i > 0$. We assume that the observed data are realizations from the class of twice-differentiable functions $f(t)$. That is, the expected expression level satisfies $\mathrm{E}\{Y_{gir}(t)\} = f_{gi}(t)$ and

$$Y_{gir}(t) = f_{gi}(t) + \epsilon_{gir}, \tag{2.1}$$

where ϵ_{gir} are i.i.d. random errors with zero means. For the mean functions $f_{g1}(t)$ and $f_{g2}(t)$ in the first and second groups, let

$$A_{g1+} = \{t : f'_{g1}(t) = 0 \text{ and } f''_{g1}(t) > 0\},$$

$$A_{g1-} = \{t : f'_{g1}(t) = 0 \text{ and } f''_{g1}(t) < 0\},$$

$$A_{g2+} = \{t : f'_{g2}(t) = 0 \text{ and } f''_{g2}(t) > 0\}$$

and

$$A_{g2-} = \{t : f'_{g2}(t) = 0 \text{ and } f''_{g2}(t) < 0\}$$

be the level-sets of the peak and valley locations of $f_{g1}(t)$ and $f_{g2}(t)$ respectively.

We are interested in testing whether or not $f_{g1}(t)$ and $f_{g2}(t)$ have the same locations of peaks or valleys. The null hypothesis can be written as

$$H_0^g : A_{g1+} = A_{g2+} \text{ and } A_{g1-} = A_{g2-}. \tag{2.2}$$

We propose a metric to measure the deviation from H_0^g in (2.2) by

$$T_{g\cdot}^{(1,2)} = \frac{1}{k_{1g} + k_{2g}} \left[\sum_{i=1}^{k_{1g}} \min_j \left| \theta_{g,i}^{(1)} - \theta_{g,j}^{(2)} \right|^2 + \sum_{j=1}^{k_{2g}} \min_i \left| \theta_{g,j}^{(2)} - \theta_{g,i}^{(1)} \right|^2 \right], \tag{2.3}$$

where $\theta_{g,i}^{(1)} \in A_{g1\cdot}$, $\theta_{g,i}^{(2)} \in A_{g2\cdot}$, $k_{1g} = |A_{g1\cdot}|$ and $k_{2g} = |A_{g2\cdot}|$ are the numbers of changing points for condition 1 and 2 respectively, $\cdot$ is either $+$ or $-$, and then, H_0^g is rejected if and only if $T_{g\cdot}^{(1,2)}$ is large. Metric $T_{g\cdot}^{(1,2)}$ measures the average square distance of either peak locations or valley locations between groups 1 and 2. So, $f_{g1}(t)$ and $f_{g2}(t)$ have the same concavity and convexity if and only if T_{g+} and T_{g-} are both 0. Also, if $k_{1g} \neq k_{2g}$, $f_{g1}(t)$ does not have the same geometry as $f_{g2}(t)$, $T_{g\cdot}^{(1,2)}$ in (2.3) is expected to be large, which will lead to a natural rejection of H_0^g.

2.2.1 A Preliminary Step

To estimate $T_{g\cdot}$ in (2.3), the locations and types of critical points are desired. Many methods of multiple change-point estimation have been developed for that purpose. For example, Fryzlewicz et al. [25] introduced the binary segmentation method and provided an approximate solution to estimate change-points. Although binary segmentation is computationally efficient, it leads to a poor estimation of the number and locations; see [26]. Other likelihood based optimization methods are highly demanding in computations. To overcome these drawbacks, we develop a method to locate the critical points. For our model (2.1), change-points are presented when $f'(t)$ equals to zero; so finding the change-points is equivalent to finding points where the sign of ${f_{gi}}'(t)$ changes. We sample $f'_{gi}(t)$ on equally-spaced grid on $[0, T]$ such that $\theta_\tau = \tau h$, $\tau = 0, 1, \ldots, n$, and $h = T/n$ and define

$$u_\tau = \mathbb{I}\left[\left| \mathbb{I}\left\{ f'_{gi}(\theta_{\tau+1}) > 0 \right\} - \mathbb{I}\left\{ f'_{gi}(\theta_\tau) > 0) \right\} \right| > \eta \right], \tag{2.4}$$

where $\eta > 0$ is small. Here, u_τ indicates where the sign of $f'_{gi}(t)$ changes and any θ_τ that makes u_τ equal to 1 corresponds to a critical point. The total number of change-points is measured by $\sum_{\tau=1}^{n} u_\tau$.

2.2.2 Testing Statistics

To perform the test, we compute the plug-in estimate

$$\widehat{T}_{g\cdot}^{(1,2)} = \frac{1}{\widehat{k}_{g1} + \widehat{k}_{g2}} \left[\sum_{i=1}^{\widehat{k}_{1g}} \min_j \left| \widehat{\theta}_{g,i}^{(1)} - \widehat{\theta}_{g,j}^{(2)} \right|^2 + \sum_{j=1}^{\widehat{k}_{2g}} \min_i \left| \widehat{\theta}_{g,j}^{(2)} - \widehat{\theta}_{g,i}^{(1)} \right|^2 \right], \tag{2.5}$$

for $g = 1, \ldots, G$, $\widehat{A}_{g1\cdot} = \{\widehat{\theta}_{g,i}^{(1)}, i = 1, \ldots, \hat{k}_{1g}\}$ and $\widehat{A}_{g2\cdot} = \{\widehat{\theta}_{g,j}^{(2)}, j = 1, \ldots, \hat{k}_{2g}\}$ where $\widehat{\theta}_{g,i}^{(1)}$ and $\widehat{\theta}_{g,j}^{(2)}$ estimate the critical points of the mean for conditions 1 and 2. Also, $\widehat{k}_{1g} = |\widehat{A}_{g1\cdot}|$ and $\widehat{k}_{2g} = |\widehat{A}_{g2\cdot}|$ are the estimated numbers of critical points for gene g under condition 1 and 2, respectively. Thus, the null hypothesis (2.2) is rejected if and only if $\widehat{T}_{g\cdot}^{(1,2)}$ is large.

2.2.3 Controlling FDR

Benjamini and Hochberg [27] introduced the False Discovery Rate (FDR), which is a popular metric to perform large scale multiple testing such as the traditional DE analysis in genomics. FDR is the expected proportion of false positives among all rejections. A method to control FDR based on comparing p-values with a shrinking threshold is introduced in [27]. Later, among a vast amount of variants to [27], an efficient approach to estimate FDR via adjusted p-values is developed in [28].

For a fixed cut-off value of our test statistics $T_{g\cdot}$, denoted by d, the true FDR and its estimator are $\text{FDR}(d) = \text{FP}(d)/\text{TP}(d)$, and $\widehat{\text{FDR}} = \widehat{\text{FP}}(d)/\text{TP}(d)$, where $\text{FP}(d)$, $\widehat{\text{FP}}(d)$ are the true and estimated number of false positives and $\text{TP}(d)$ is the total number of hypotheses that are rejected at the cut-off value d. In order to estimate the distribution of $T_{g\cdot}$ in (2.5) under the nulls, a permutation procedure is employed. As the gene under alternatives does not share the same distribution with the genes under the null, the empirical distribution of statistics for all genes may fail to estimate

the true null distribution accurately [6]. To overcome these problems a new permutation method is introduced to control FDR [6].

For any $\epsilon > 0$, we define any gene g satisfying $T_{g\cdot} > \epsilon$ to be significant and set TP $= |\{g : T_{g\cdot} > \epsilon\}|$. Let the set of non-significant genes selected by some other statistic F_g (e.g. FANOVA statistics, two-sample t test statistics, and our statistics in (2.5) or the L^2 distance) as $D(\epsilon) = \{g : F_g \leq \epsilon'\}$, and ϵ' is chosen so that $|D(\epsilon)^c|$ = TP. That is ϵ' is determined by the level ϵ and statistics F_g's. Then the observed expression levels are permuted B times. For each permuted dataset b, we compute the null statistic $T_{g\cdot}^{(b)}$ and estimate FP via

$$\widehat{\text{FP}} = \frac{1}{B}\sum_{b=1}^{B} |\{g \in D(\epsilon) : T_{g\cdot}^{(b)} > \epsilon\}|$$

and FDR via

$$\widehat{\text{FDR}}(\epsilon) = \frac{\widehat{\text{FP}}(\epsilon)}{\text{TP}(\epsilon)}. \tag{2.6}$$

A similar procedure is proposed by [6] to improve the accuracy for estimating the null distribution of test statistics. Then, ϵ such that $\widehat{\text{FDR}}(\epsilon) \leq \alpha$ is chosen, where α is a pre-selected nominal FDR level, and genes such that $\widehat{T}_{g\cdot}^{(1,2)} > \epsilon$ are identified as DE. The proposed method is summarized in the following algorithm.

2.3 Monte Carlo Evidence

In this section, we evaluate the proposed method with different FDR estimation methods using two sets of simulation studies.

2.3.1 Simulation Settings

We generate time-course data from (2.1). Specifically, the datasets contain $G = 2,500$ genes from 2 treatments, with $n = 6, 9$ or 15 replicates and $T = 10$ or 15 time points. For each setting, 80% of the genes are randomly chosen with common true mean functions $f_{g1}(t) = f_{g2}(t) = f_1(t)$. The other 20% of the genes are generated with different true mean functions in each condition,

Data: Gene expression data $\{\mathbf{Y}_{gi}(t)\}_{g=1,i=1}^{G,I}$ and Nominal FDR level α
Result: DE gene list.
for $g = 1, \ldots, G$ **do**
    Fit smooth spline model and record the estimated first and second derivative.
    Find the number and location of critical points using (2.4).
    Define the critical point type and calculate (2.5).
end
Find null set genes $D(\epsilon)$ who are detected by test statistics F_g.
for $b = 1, \ldots, B$ **do**
    Permute treatment condition label on null set genes $D(\epsilon)$ and store the permuted test statistics $\{T_{i\cdot}^{(b)}\}_{i \in D(\epsilon)}$.
    Compute $\widehat{\text{FDR}}(\epsilon)$ in (2.6).
end
Choose ϵ to control $\widehat{\text{FDR}}(\epsilon) \leq \alpha$ and identify DE genes wherever $T_{g\cdot} > \epsilon$.

i.e. $f_{g1}(t) = f_1(t)$ and $f_{g2}(t) = f_2(t)$. We apply different procedures to estimate FDR, and for comparison, the empirical FDRs and powers (averaged over 100 replicates) are reported for each procedure at nominal levels $0.05, 0.075, 0.1$ and 0.15.

Simulation A: Valley only

In (2.1), $f_1(t) = (t - 2 - \eta_1)^2$ and $f_2(t) = (t - 3.5 - \eta_1)^2$, where $\eta_1 \sim N(0, \sigma_1^2)$, and $\sigma_1 = 0.5$. We consider two different measurement error settings: (1) $\epsilon_{gir}(t) = 1/\sigma \times \tau_{gir}(t)$ where $\tau_{gir}(t) \sim t_{12}$, $\sigma = 0.5$ and (2) $\epsilon_{gir}(t) \sim N(0, 0.7)$.

Simulation B: Both peaks and valleys with trigonometric functions

In (2.1), $f_1(t) = -3\cos\{(x - \eta_1)/3 \times 2\pi\}$, and $f_2(t) = -3\cos\{(x - \eta_1)/3 \times 2\pi\}\,\mathbb{I}(x \leq 9/2 + \eta_1) + (4.47x - 16.72)\mathbb{I}(x > 9/2 + \eta_1)$, where $\eta_1 \sim N(0, 0.16^2)$ and $\epsilon_{gir}(t) \sim N(0, \sigma^2)$ with $\sigma^2 = 0.5$ or 1.

2.3.2 Simulation Results

For comparison, these procedures to estimate $\widehat{\text{FP}}$ in section 2.2.3 are considered: 1) null genes are not pre-selected and the standard permutation procedure are applied 2) null genes are selected

by F_g, where F_g is the FANOVA statistic, and 3) null genes are selected by T_g, where T_g is our metric in (2.5). Simulation results are shown in Tables 2.1–2.3 and A.1–A.3 as Standard, FANOVA based, and T_g based, respectively.

For each gene, we estimate the mean function, $f_{gi}(t)$ and calculate the test statistics by the method in Section 2.2.2. The empirical FDRs and the empirical power are shown, at nominal FDR level, in Tables 2.1 and A.1 for simulation A. The simulation results for setting B are shown in Tables 2.2, 2.3, A.2 and A.3.

Since the mean function under setting A only contains a valley, we report the simulation results for valley location only. When T increases from 10 to 15, the average empirical powers for all three methods increase. As the number of replicates n increases from 6 to 15, the average empirical powers increase as well. Applying our T_g to select null genes tends to have more false positives compared with the FANOVA-based method and standard method. The standard method is too conservative and not obtain enough power. Among these three methods, the estimated FDR using FANOVA-based method is very close to the nominal values and achieves the optimal power. We report similar tables for setting B, with both peak and valley results. With $\sigma = 0.5$, from Tables 2.2 and 2.3, we observe the empirical powers of testing valley locations are much lower than empirical powers for testing differences in peak locations. This is because the difference between $f_1(t)$ and $f_2(t)$ in setting B is the peak location. Similar results are shown in Table A.2 and A.3. With the increasing error σ, the estimated FDRs do not change, while the average empirical powers decrease because of the larger variation. Simulation results for setting A and B indicate that the best permutation approach to estimate FDR is using F_g as the FANOVA statistics. So, we use the FANOVA based approach for further application to diurnal differential expression analysis.

2.4 An Application to Diurnal Differential Expression Analysis on Maize Study

Many tissues have circadian clocks that generate transcriptional rhythms. These are important for the daily timing and physiological processes [29]. Furbank and Taylor [30] show that most

Table 2.1: Simulation A: Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (A) with $\sigma = 0.5$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.049	0.071	0.094	0.148	0.050	0.073	0.099	0.137
		Power	0.964	0.968	0.970	0.973	0.972	0.975	0.976	0.978
	T_g based	FDR	0.055	0.081	0.111	0.175	0.064	0.083	0.117	0.162
		Power	0.966	0.969	0.971	0.975	0.974	0.975	0.977	0.980
	Standard	FDR	0.014	0.011	0.010	0.010	0.064	0.083	0.117	0.162
		Power	0.561	0.715	0.771	0.826	0.596	0.739	0.787	0.835
9	FANOVA based	FDR	0.045	0.070	0.097	0.141	0.046	0.068	0.084	0.144
		Power	0.971	0.974	0.976	0.978	0.979	0.980	0.981	0.983
	T_g based	FDR	0.052	0.084	0.097	0.166	0.055	0.084	0.107	0.144
		Power	0.972	0.975	0.976	0.979	0.980	0.981	0.982	0.983
	Standard	FDR	0.007	0.007	0.006	0.006	0.008	0.008	0.007	0.007
		Power	0.758	0.807	0.855	0.879	0.763	0.811	0.859	0.883
15	FANOVA based	FDR	0.045	0.067	0.088	0.133	0.045	0.064	0.094	0.150
		Power	0.979	0.981	0.982	0.984	0.982	0.984	0.985	0.987
	T_g based	FDR	0.051	0.070	0.095	0.156	0.049	0.065	0.095	0.164
		Power	0.980	0.981	0.982	0.984	0.983	0.984	0.985	0.987
	Standard	FDR	0.005	0.005	0.005	0.005	0.006	0.006	0.006	0.006
		Power	0.862	0.886	0.906	0.929	0.860	0.886	0.905	0.929

plants use the C_3 pathway of photosynthesis, also known as the photosynthetic carbon reduction cycle, to convert light energy into chemical energy. The C_4 pathway is a complex adaptation of the C_3 pathway that overcomes the limitation of photorespiration that is found in a diverse collection of species. In 2005, Storey et. al. [12] address the temporal approach to study DE genes. In this section, in order to study the transcriptional rhythms generated by circadian clocks, we analyze the local geometric changes of the mean expression level change using the method in Section 2.2. This is equivalent to identifying the DE genes with different temporal patterns across conditions. We perform analysis on the general genes and C_4 genes separately. We further determine whether there

Table 2.2: Simulation B: (Peak) Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (B) with $\sigma = 0.5$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.050	0.069	0.083	0.136	0.043	0.078	0.078	0.145
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	T_g based	FDR	0.058	0.083	0.105	0.180	0.043	0.078	0.078	0.145
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Standard	FDR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Power	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000
9	FANOVA based	FDR	0.048	0.068	0.084	0.154	0.047	0.047	0.047	0.120
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	T_g based	FDR	0.057	0.084	0.120	0.176	0.047	0.047	0.047	0.120
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Standard	FDR	0.000	0.000	0.014	0.028	0.000	0.000	0.000	0.001
		Power	0.000	0.000	0.999	1.000	0.000	0.000	0.000	1.000
15	FANOVA based	FDR	0.046	0.046	0.074	0.139	0.026	0.026	0.026	0.120
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	T_g based	FDR	0.046	0.074	0.074	0.139	0.026	0.026	0.026	0.120
		Power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Standard	FDR	0.011	0.017	0.021	0.046	0.000	0.000	0.001	0.004
		Power	1.000	1.000	1.000	1.000	0.000	0.000	1.000	1.000

Table 2.3: Simulation B: (Valley) Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (B) with $\sigma = 0.5$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.020	0.041	0.057	0.109	0.000	0.001	0.010	0.090
		Power	0.495	0.532	0.554	0.598	0.404	0.491	0.566	0.666
	T_g based	FDR	0.041	0.057	0.080	0.155	0.032	0.050	0.090	0.147
		Power	0.532	0.554	0.576	0.626	0.614	0.637	0.666	0.696
	Standard	FDR	0.000	0.002	0.006	0.020	0.000	0.000	0.001	0.010
		Power	0.305	0.386	0.436	0.495	0.348	0.432	0.491	0.566
9	FANOVA based	FDR	0.027	0.048	0.048	0.078	0.000	0.021	0.060	0.141
		Power	0.598	0.626	0.626	0.653	0.546	0.666	0.709	0.750
	T_g based	FDR	0.048	0.078	0.078	0.155	0.044	0.077	0.095	0.141
		Power	0.626	0.653	0.653	0.695	0.696	0.719	0.730	0.750
	Standard	FDR	0.000	0.002	0.010	0.027	0.000	0.000	0.002	0.021
		Power	0.434	0.496	0.555	0.598	0.474	0.546	0.587	0.666
15	FANOVA based	FDR	0.034	0.053	0.074	0.122	0.025	0.059	0.085	0.123
		Power	0.691	0.711	0.725	0.750	0.745	0.771	0.785	0.800
	T_g based	FDR	0.046	0.074	0.099	0.137	0.048	0.071	0.085	0.145
		Power	0.705	0.725	0.740	0.758	0.765	0.778	0.785	0.808
	Standard	FDR	0.001	0.004	0.010	0.043	0.000	0.001	0.006	0.025
		Power	0.570	0.620	0.646	0.702	0.612	0.664	0.707	0.745

are any significant differences in number of DE genes between the general genes and C_4-specific genes via an enrichment analysis.

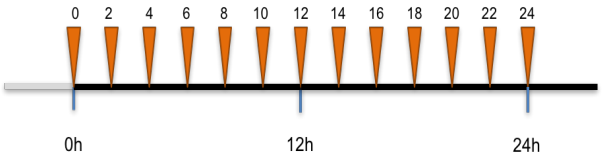
2.4.1 Real Data Description

As discussed in Section 2.1, the objective of the maize study is to assess effects of light on four different leaf sections: the base, ligule from 3 to 4 cm, ligule from 8 to 9 cm, and the tip position on a leaf. In this experiment, measurements at 13 points are obtained under two different light conditions, constant light and constant dark. At each time point, three biological replicate leaves are randomly picked from genetically identical plants under each light condition. In total, $63,293$ genes' expression levels are collected from each of the four different leaf sections (see Fig 2.3). We focus on the data collected from the constant dark and constant light conditions, the protocol of which can be generalized to data collected from the general diurnal experiment. We identify DE genes by comparing the differences of convexity and concavity of the mean functions between different sections. The genes in the base section are very noisy and not informative for local pattern analysis, so we focus our analysis on data from sections $2, 3$ and 4.

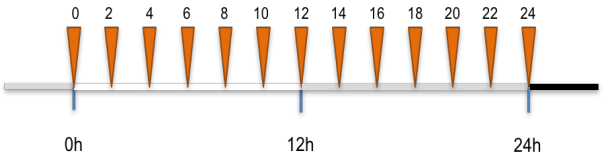
Filtering

In general, there are three classifications for genes: 1) the temporal profiles of the gene expression in all sections have constant or linear patterns, 2) one or two sections have nonlinear temporal profiles of the gene expression and 3) the temporal profiles of the gene expression in all sections have nonlinear patterns. Much work has been devoted to test differences for the first type of data, such as comparing the estimated slope directly. Methods for comparing a non-linear pattern to a linear pattern have also been developed in [29]. The third type of genes is our primary interest.

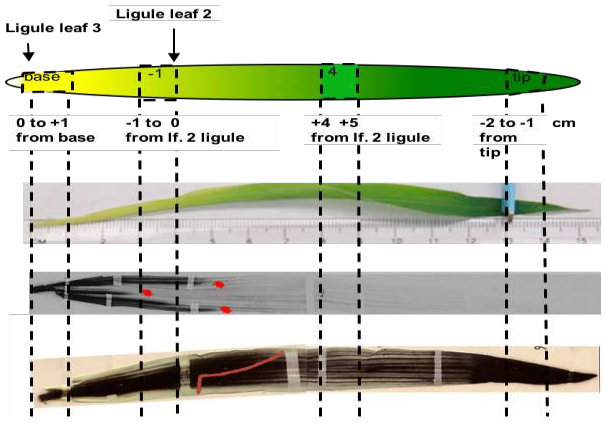
That is, genes with nonlinear patterns for all sections are selected for analysis. We first apply the goodness of fit method in [29] to filter genes with constant or linear only patterns by testing the null hypothesis, $H_0 : \mu(t) = \beta_0 + \beta_1 t$, versus the alternative that the mean function has a nonlinear pattern. Upon filtering, $10,330$ genes remain for further analysis under the constant dark condition and $9,063$ genes under the constant light condition.



(a) Experiment time points under constant dark



(b) Experiment time points under constant light



(c) Treatment section, namely the base, 1cm, 4cm, and tip position on the leaf

Figure 2.3: Experiment Description: (a) and (b) within 24 hours experiment time, the number of genes will be measured every 2 hours, so we have 13 time points in total. (c) Under each experimental condition (light or dark), four sections in one leaf have been measured.

2.4.2 Summary Analysis

To identify DE genes with different local geometry across sections we consider the null hypothesis for gene g as

$$H_0^g : A_{gl+} = A_{gm+} \text{ and } A_{gl-} = A_{gm-} \ \ \forall\, l \neq m.$$

By extending the two-sample test statistics $T_{g\cdot}^{(1,2)}$ in (2.5) to more than two samples, we consider the average paired "distance" $\overline{T}_{g\cdot} = \sum_l \sum_{m \neq l} T_{g\cdot}^{(l,m)}/C$, where C is the number of distinct combinations of sections. The null hypothesis $H_0^{g,\epsilon}$ is rejected whenever $\overline{T}_{g\cdot}$ is large. Applying the proposed method, we identify $1,658$ DE genes out of $10,330$ genes under constant dark experiment and $1,020$ genes out of $9,063$ genes under constant light experiment at an 0.05 nominal FDR level. During the 24 hours experiment time, the local geometries of temporal profiles of these DE

genes change across sections. Also, under light condition, out of 112 C_4 genes (55 without isoform), we find 15 (7 without isoform) of them to be differentially expressed. On the other hand, for dark data, within these 73 genes (39 without isoform) C_4 genes, we find 22 genes (13 without isoform) to be DE.

Table 2.4: C_4/PS genes under Light: significant DE C_4/PS genes under light. * means DE genes in both light and dark condition

gene ID	Method	both DE	Annotation
GRMZM2G359038	V		uncharacterized LOC100272602
GRMZM2G040933	P	*	plastidic general dicarboxylate transporter
GRMZM2G086258	P	*	plastidic general dicarboxylate transporter
GRMZM2G153920	P		sorbitol transporter
GRMZM2G374812	P		hexose carrier protein HEX6
GRMZM2G113033	P		ribulose bisphosphate carboxylase small subunit 2
GRMZM2G463280	V		Phosphoribulokinase

Tables A.5 and A.6 display the top 10 significant DE genes, ordered by the the average "distance" $\overline{T}_{g\cdot}$, under both the constant light and the constant dark conditions. As our method compares either valley or peak locations, 20 genes are listed (10 for each). The method column includes the type of critical point that is used, indicated as V for valley and P for peak. Fig A.2 displays the heatmap of the top 10 significant DE genes under both the constant light and the constant dark conditions. For example, from the heatmap of Gene GRMZM2G074672_T01 in Fig A.2a, we observe that not only the scale but the shape of the temporal profile are different between sections.

Tables 2.4 and A.7 present the results for DE C_4 genes under constant light and constant dark conditions. We find that genes GRMZM2G086258 and GRMZM2G040933 are DE under both conditions. Fig A.1 includes the log scale RPKM data plot with the smoothed mean curve of gene GRMZM2G086258. From this figure, we see that the mean in section 2 has a different pattern compared with section 3 and 4 under both the constant light and the constant dark conditions. As the time increases under the constant light condition, the log scale gene expression level decreases in both sections 3 and 4. At the opposite extreme, the gene expression level increases in section 2

(see Fig A.1a). Similarly, the gene expression level decreases in section 2 as shown in the red curve in Fig A.1b, the exact opposite of both the blue and the green curves. However, the local geometry in section 3 is similar to that in section 4 for this gene under both light conditions. Heatmaps of gene GRMZM2G040933 in Fig A.2a and A.2b indicate the mean patterns of sections 2, 3 and 4 are different for both the constant light and the constant dark conditions.

2.4.3 Enrichment Analysis

Gene set enrichment analysis is a powerful and revealing follow-up step for RNA-seq analysis. By carefully inspecting predefined gene sets, we can verify statistical discoveries and more importantly, identify critical pathways that are responsive to treatment variations. For this analysis, the goal is to determine if the C_4 gene is more likely to have different local patterns in different sections than non C_4 genes. We conduct the Fisher's exact test to perform the enrichment analysis. The null hypothesis is that the proportion of the C_4 gene to be DE is less than or equal to the proportion of other genes

$$H_0 : P_{C_4} \leq P_{all} \text{ vs } H_a : P_{C_4} > P_{all}. \tag{2.7}$$

We found that the proportion of being DE is highly enriched for the C_4 gene under constant dark condition (p-value $= 0.01$), while there is no significant enrichment under the constant light condition (p-value $= 0.508$).

2.5 Discussion

Given the small sample size and large number of potentially composite hypotheses depending on the biological questions of interest, analysis on the differential geometric pattern of the time-course RNA-seq data is challenging. In this chapter, we focus on testing whether or not the treatment affects the geometric attributes of the temporal profile of the gene expressions. This approach helps us identify interesting genes and pathways. Further biology study needs to be conducted for verification of the identified DE genes.

In order to assess detailed differences of the local curvature over time between different conditions, we propose a metric to quantify the changes of the geometric attributes of the temporal profile and a multiple inference procedure to identify DE genes. In particular, we model the transformed count data for genes at each time point using a spline model and develop a testing framework based on a permutation procedure. We demonstrate that the proposed test achieves satisfactory power and has its FDR controlled via simulation studies. We apply the proposed method to the data collected from a light physiology experiment on maize. Though the proposed method focuses on the context of comparing the RPKM data, our framework is also applicable to other types of gene expression data. Also, the framework shown in this chapter gives a general approach to build local geometric tests in multiple hypothesis testing problems.

Chapter 3

Large Scale Maximum Average Power Multiple Inference on Time-Course Count Data with Application to RNA-Seq Analysis

3.1 Introduction

Studying transcriptomes through the sequencing of RNA (RNA-seq) has revolutionized biology and medical science. RNA-seq experiments have been employed to detect allele-specific expression and novel biomarkers, to assist medical prognosis, and to explore how global expression profiles alter in different biological environments. RNA-seq is affordable and allows for more samples; in particular, longitudinal studies are feasible, and can reveal dynamical biological patterns for thousands of genes simultaneously.

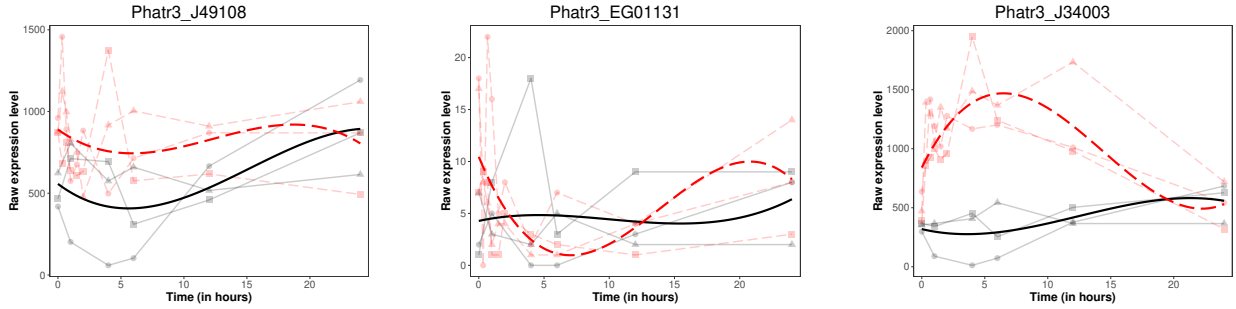
As with traditional RNA-seq analysis [3, 31, 32], the major goal of time-course RNA-seq is to detect differentially expressed (DE) genes across treatments. Genes differentially expressed over time are of particular interest. An early and popular analysis for time-course RNA-seq data is to model logarithmic fold change (LFC) as a function of time points as categorical factors [3,32]. This approach does not account for possible smoothness in the mean dynamics and may fail to identify DE genes with complex temporal profiles [7]. An extension of the original model for time-course data in microarray experiments is the negative binomial employed by `maSigPro-GLM`, which accounts for temporal structure by a polynomial in time [33]. Similarly, `splineTC` uses cubic splines to model time-course transcriptome data and applies empirical Bayes moderated F-statistics for DE analysis [34]. [23] account for temporal structure with a hidden Markov model. [35] introduces an MCMC sampling procedure to identify temporally DE genes. The one-sample problem for time-course RNA-seq data, which focuses on identifying genes or gene sets with significant temporal dynamics, has also been studied [36, 37]. Based on a family of flexible parametric func-

tions, ImpulseDE2 performs both one-sample and two-sample analyses on time-course RNA-seq data [38]. Using a negative binomial mixed-effect model (NBMM), [7] analyze time-course read counts of genes at the exon level and identify DE genes using the Kullback-Leibler distance ratio. They also discuss the importance of testing a variety of composite hypotheses other than the overall temporal pattern; namely, the nonparallel differentially expressed (NPDE) and parallel differentially expressed (PDE) genes, which are modeled through time by treatment interactions. Other methods for analyzing time-course RNA-seq data include [39–41]; see [42] for a review.

The aforementioned methods for analyzing time-course RNA-seq data may suffer from low power in practice. Though some approaches—such as the local regression model employed by [31] to estimate the dispersion parameters—adopt the idea of borrowing information across genes from the traditional RNA-seq analysis, the improvement on power is not completely theoretically justified and is unsatisfactory in practice. To illustrate, we considered the fission yeast data set used by [3] to show the performance of DESeq2 in analyzing time-course RNA-seq data. Besides DESeq2, we also fitted the expression of each gene to a negative binomial model and identified DE genes using the resulting likelihood ratio statistics and the standard [27] procedure. We refer to this procedure as LRT. With the nominal false discovery rate (FDR) controlled at 0.05, our proposed method identified 128 temporally DE genes out of $7,039$ genes while DESeq2 identified 85 DE genes and the LRT only identified 27 DE genes. See Appendix B for further discussion of this example. In addition to the lack of theoretical guarantees on the power to detect DE genes using time-course RNA-seq data, the FDR control of existing methods is not well studied. Simulation and empirical studies like [43], [44], [45], and [46] show that both DESeq and edgeR are conservative in some cases while liberal in others for traditional RNA-seq analysis, and similar observations are reported for analyzing time-course RNA-seq data [7, 47]. Our comprehensive simulations support the same conclusions.

Furthermore, most existing methods, such as maSigPro-GLM, splineTC, DEseq2, and edgeR, are mainly designed for overall differential patterns across time among conditions. But other differential expression profiles, such as PDE and NPDE genes [7], can provide a more subtle

and accurate characterization of underlying biological processes. Section 4.4 describes analysis of time-course RNA-seq data collected from a novel light physiology experiment [2] on the marine diatom *Phaeodactylum tricornutum* (*P.t.*). The experiment investigates the gene transcriptional response of *P.t.* during the process of photoacclimation in order to better understand how photosynthetics are regulated during the shifts of light conditions from low to high levels. Using our proposal, we detected more DE genes than existing methods and identified different types of DE profiles, as displayed in Figure 3.1. Specifically, gene Phatr3_J49108 primarily exhibits a relative overall mean shift on the expression between the high and low light levels, gene Phatr3_EG01131 displays altered temporal patterns between these two light levels while the overall mean levels remain unchanged, and gene Phatr3_J34003 possesses a more complex DE profile.



(a) PDE with no time-by-treatment

(b) NPDE with only time-by-treatment

(c) NPDE with both treatment and time-by-treatment

Figure 3.1: Three types of DE profiles (labeled in the caption of each plot) across time identified by the proposed method in Section 3.3 using the RNA-seq data for *P.t.* from a time-course experiment on the effect of light intensity on the algae physiology and molecular mechanism. Two light conditions are examined in this experiment, low and high light levels. Red and black curves represent the group with low and high light levels, respectively. The y-axis are raw expression levels.

To address the above challenges systematically, motivated by [48–50], we develop an optimal test in Section 3.2 for time-course RNA-seq data analysis: the test achieves maximum power averaged across all genes for which null hypotheses are false, while controlling the FDR. Use of the test relies on a model, which we develop in Section 3.3. Read counts for a gene at each time point have negative binomial distributions, with the mean temporal profile generated by a latent mixture Gaussian process. By modeling different types of DE profiles with mixture components, we are able to

draw inference on a variety of composite hypotheses other than the simple overall DE hypothesis. In addition, the model naturally adapts to non-equally spaced time points across conditions. We carefully establish the identifiability of the proposed mixture model Appendix B and implement the proposed test by developing an efficient algorithm to estimate the model parameters, using the gradient expectation-maximization (EM) algorithm and quasi-Monte Carlo integration Appendix B. We perform comprehensive simulation studies in Section 3.5 and Appendix B to demonstrate the advantages of our method in comparison to existing methods. In Section 4.4, we analyze the *Phaeodactylum* light data using our method. Some biologically interesting genes and critical pathways, which are potentially related to photosynthesis regulation, are discovered. Discussion follows in Section 6.

3.2 A Maximum Average Power Test

Let $\mathbf{Y}_g$ denote the vector (combined across all experimental conditions, time points, and replicates) of read counts for gene g ($g = 1, \ldots, G$) from an RNA-seq experiment. Assume that $\mathbf{Y}_g$ follows a distribution parameterized by generic $\boldsymbol{\eta}_g$ and $\boldsymbol{\tau}_g$, where $\boldsymbol{\eta}_g$ models the potential differential expression across conditions. Specifically, considering Δ_0 and Δ_1 as the null and alternative sets for $\boldsymbol{\eta}_g$, we identify DE genes by testing, for each g,

$$H_0^g : \boldsymbol{\eta}_g \in \Delta_0 \text{ vs. } H_1^g : \boldsymbol{\eta}_g \in \Delta_1. \tag{3.1}$$

Sets Δ_0, Δ_1 can be defined flexibly to reflect a variety of biological questions of interest. Without loss of generality, suppose that the first G_0 null hypotheses are true while others are false: with respect to a common dominating measure $\nu(\cdot)$, the densities $f(\cdot|\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)$ are null for $g = 1, 2, \ldots, G_0$ and alternative for $g = G_0 + 1, G_0 + 2, \ldots, G$. Employing these densities and Neyman-Pearson arguments, [49] developed an optimal discovery procedure (ODP) to identify DE genes, in which null hypothesis H_0^g is rejected if and only if $S(\mathbf{Y}_g) \leq \lambda$ where $S(\mathbf{Y}) = \{\sum_{g=1}^{G_0} f(\mathbf{Y}|\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)\}\{\sum_{g=G_0+1}^{G} f(\mathbf{Y}|\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)\}^{-1}$. Let the expected false and true positives be

$\text{EFP}(\Gamma) = \sum_{g=1}^{G_0} \int_\Gamma f(\mathbf{Y}|\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)\, d\nu(\mathbf{Y})$ and $\text{ETP}(\Gamma) = \sum_{g=G_0+1}^{G} \int_\Gamma f(\mathbf{Y}|\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)\, d\nu(\mathbf{Y})$, where $\Gamma = \{\mathbf{Y}_g : S(\mathbf{Y}_g) \leq \lambda\}$ is the significance region with respect to λ. [49] shows that the ODP achieves maximum ETP among all procedures controlling the EFP. However, ODP implementation relies on estimates of G_0 and identification of the nulls in advance, which may be challenging for time-course RNA-seq data. Further, ODP needs G^2 likelihood evaluations, which is computationally demanding.

To circumvent these challenges with ODP, we adopt the idea of borrowing information across genes, which is commonly used in traditional DE analysis to improve power [48, 51]. Specifically, consider $(\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)$'s to be independent random vectors with finite second moments from a common probability distribution μ_0 if H_0^g is true and μ_1 otherwise. Assume that $G_0 \to \infty$, $G - G_0 \to \infty$ with $G_0/G \to p_0 < 1$. Then

$$\text{EFP}(\Gamma)/G_0 \xrightarrow{\text{p}} \int\int_\Gamma f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau})\, d\nu(\mathbf{Y})\, d\mu_0(\boldsymbol{\eta}, \boldsymbol{\tau}) := \mathbb{P}\left(\mathbf{Y} \in \Gamma | \Delta_0\right), \tag{3.2}$$

$$\text{ETP}(\Gamma)/(G - G_0) \xrightarrow{\text{p}} \int\int_\Gamma f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau})\, d\nu(\mathbf{Y})\, d\mu_1(\boldsymbol{\eta}, \boldsymbol{\tau}) := \mathbb{P}\left(\mathbf{Y} \in \Gamma | \Delta_1\right). \tag{3.3}$$

[48] refer to (3.2) as the average type I error and (3.3) as the average power.

Analogous to the ODP, we propose a test statistic $\delta^*_{\text{MAP}}(\mathbf{Y}) = p_0 \int f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau})\, d\mu_0(\boldsymbol{\eta}, \boldsymbol{\tau}) \{(1 - p_0) \int f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau})\, d\mu_1(\boldsymbol{\eta}, \boldsymbol{\tau})\}^{-1}$ for (3.1) and H_0^g is rejected if and only if $\delta^*_{\text{MAP}}(\mathbf{Y}_g) \leq \lambda$ for some $0 \leq \lambda < \infty$. Statistic $\delta^*_{\text{MAP}}(\mathbf{Y})$ mimics $S(\mathbf{Y})$ in the ODP and maximizes the average power while controlling the average type I error, and is also equivalent to the maximum average power testing statistic introduced by [48] and [50].

Furthermore, assume that $(\boldsymbol{\eta}_g, \boldsymbol{\tau}_g)$ are independent and identically distributed from a common mixture $\pi(\boldsymbol{\eta}, \boldsymbol{\tau}) = p_0 \pi_0(\boldsymbol{\eta}, \boldsymbol{\tau}) + (1 - p_0)\pi_1(\boldsymbol{\eta}, \boldsymbol{\tau})$. Let $\Theta = \Delta_0 \cup \Delta_1$ and Ω denote the parameter spaces for $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$, respectively. Assume that $\pi_0(\boldsymbol{\eta}, \boldsymbol{\tau})$ has support $\Delta_0 \times \Omega$ and $\pi_1(\boldsymbol{\eta}, \boldsymbol{\tau})$ has support $\Delta_1 \times \Omega$, and Δ_0 is zero-measure under $\pi_1(\boldsymbol{\eta}, \boldsymbol{\tau})$ and Δ_1 is zero-measure under $\pi_0(\boldsymbol{\eta}, \boldsymbol{\tau})$. Then, a statistic equivalent to $\delta^*_{\text{MAP}}(\mathbf{Y})$ yet providing a natural way to estimate FDR is

$$\delta_{\text{MAP}}(\mathbf{Y}) = \left\{ \int_{\Delta_0 \times \Omega} f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau}) d\mu(\boldsymbol{\eta}, \boldsymbol{\tau}) \right\} \left\{ \int_{\Theta \times \Omega} f(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\tau}) d\mu(\boldsymbol{\eta}, \boldsymbol{\tau}) \right\}^{-1}. \tag{3.4}$$

Theorem 3 in [48] states that among all tests controlling the FDR, a Neyman-Pearson test maximizes the average power. Hence, the tests based on $\delta^*_{\text{MAP}}(\mathbf{Y})$ and (3.4) also maximize the average power while controlling the FDR.

Given $\pi(\boldsymbol{\eta}, \boldsymbol{\tau})$ above, $\delta_{\text{MAP}}(\mathbf{Y}) = \mathbb{P}(\mathbf{Y}, \Delta_0)\,\mathbb{P}^{-1}(\mathbf{Y}) = \mathbb{P}(\Delta_0|\mathbf{Y})$ is the posterior probability of Δ_0 given $\mathbf{Y}$. The (posterior) expected number of false positives for a given rejection region Γ is $\sum_g \delta_{\text{MAP}}(\mathbf{Y}_g)\mathbb{I}(\mathbf{Y}_g \in \Gamma) = \sum_g \mathbb{P}(\Delta_0|\mathbf{Y}_g)\mathbb{I}(\mathbf{Y}_g \in \Gamma)$. As is standard [48, 49], we estimate the FDR by the ratio of expected false positives and expected positives, which is $\widehat{\text{FDR}}_\Gamma = \sum_g \delta_{\text{MAP}}(\mathbf{Y}_g)\mathbb{I}(\mathbf{Y}_g \in \Gamma)/\{\sum_g \mathbb{I}(\mathbf{Y}_g \in \Gamma)\}$. Therefore, rejection region $\Gamma(\alpha) = \{\mathbf{Y}_g : \delta_{\text{MAP}}(\mathbf{Y}_g) \leq \lambda_\alpha\}$ is chosen with $\widehat{\text{FDR}}_\Gamma \leq \alpha$ for a nominal level α and it defines the maximum average power test for (3.1).

3.3 Methodology

3.3.1 Data Model

Let $Y_{gij}(t)$ denote the number of reads mapped to gene g from the jth replicate in treatment group i at time point t, where $g = 1, \ldots, G$, $i = 1, 2$, $j = 1, \ldots, n_i$, and $t = t_1, \ldots, t_{T_i}$ with integers $T_i > 0$. A widely applicable model for traditional RNA-seq analysis is the negative binomial (NB), which provides extra flexibility to model count data with large variations yet includes the popular Poisson model as a special case [50, 52]. We therefore model $Y_{gij}(t)|\lambda_{gij}(t)$ as independent NB $(\lambda_{gij}(t), \phi_g)$ across genes g, treatments i, replicates j and time points t, where ϕ_g is the dispersion parameter and

$$\lambda_{gij}(t) = \mathbb{E}\{Y_{gij}(t)|\lambda_{gij}(t)\} = S_{ij} \exp\left\{\eta_{g1i} + \mathbf{B}'(t)\boldsymbol{\eta}_{g2i} + \mathbf{B}'(t)\boldsymbol{\tau}_g\right\} \tag{3.5}$$

with some q-dimensional orthogonal basis functions $\mathbf{B}'(t) = (b_1(t), \ldots, b_q(t))$, $\eta_{g1i} = \eta_{g1}\mathbb{I}_{\{i=2\}}$, $\boldsymbol{\eta}_{g2i} = \boldsymbol{\eta}_{g2}\mathbb{I}_{\{i=2\}}$ with $\boldsymbol{\eta}_{g2}, \boldsymbol{\tau}_g \in \mathbb{R}^q$. In (3.5), $\mathbf{B}'(t)\boldsymbol{\tau}_g$ models the mean temporal pattern for time-

course expressions and S_{ij} are normalization factors (treated as known constants in practice) that adjust for varying sequencing depths and other technical effects across replicates.

Here, η_{g1} is the relative mean shift of gene expression between treatments and $\boldsymbol{\eta}_{g2}$ characterizes potential interactions between the temporal patterns and treatments. If $\boldsymbol{\eta}_{g2}$ is not zero, the relative differences between $\lambda_{g1j}(t)$ and $\lambda_{g2j}(t)$ display NPDE patterns as introduced in [7]. PDE may hold when $\boldsymbol{\eta}_{g2} = \mathbf{0}$, for which the relative differences between $\lambda_{g1j}(t)$ and $\lambda_{g2j}(t)$ do not vary over time [7].

Like the negative binomial model for time series of counts in [53], model (3.5) has a smooth mean temporal pattern $\lambda_{gij}(t)$, conditional on a latent, zero-mean Gaussian process $\mathcal{G}(t)$. By Mercer's Theorem [54], $\mathcal{G}(t)$ admits a series representation that converges almost surely. In practice, a finite expansion $\sum_{k=1}^{K} Z_k b_k(t)$ approximates $\mathcal{G}(t)$ well for even relatively small K, where $b_k(t)$'s are the eigenfunctions corresponding to the covariance structure of $\mathcal{G}(t)$ and eigenvalues Z_k's are independent normal random variables, which motivates model (3.5). This link to general Gaussian processes reflects the flexibility of the proposed model. Since $\eta_{g1}, \boldsymbol{\eta}_{g2}$ and $\boldsymbol{\tau}_g$ mimic the eigenvalues Z_k, they are similarly modeled using normal distributions in Section 3.3.2 below.

3.3.2 Latent Model and Hypotheses

Under model (3.5), testing (3.1) reduces to testing $H_0^g : \eta_{g1}, \boldsymbol{\eta}_{g2} \in \Delta_0$ versus $H_1^g : \eta_{g1}, \boldsymbol{\eta}_{g2} \in \Delta_1$. To employ the proposed maximum average power test defined by $\Gamma(\alpha)$ in Section 3.2, the derivation of $\delta_{\text{MAP}}(\mathbf{Y})$ in (3.4) assumes information on $\pi(\boldsymbol{\eta}, \boldsymbol{\tau})$, where $\boldsymbol{\eta} = (\eta_1, \boldsymbol{\eta}_2)$ for our model. Motivated from the discussion in Section 3.3, we consider $(\eta_{g1}, \boldsymbol{\eta}_{g2}, \boldsymbol{\tau}_g)$ as independent normal random vectors. Given the large number of tests to be performed, we balance the computational burden and model flexibility by specifying

$$(\eta_{g1}, \boldsymbol{\eta}_{g2}, \boldsymbol{\tau}_g) \sim \pi(\boldsymbol{\eta}, \boldsymbol{\tau}) = \sum_{k=1}^{K} p_k \mathcal{N}\left(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Lambda}^{(k)}\right) \tag{3.6}$$

where $\boldsymbol{\mu}^{(k)} = (\mu_1^{(k)}, 0, \ldots, 0)'$ and $\boldsymbol{\Lambda}^{(k)} = \text{diag}(\sigma_1^{2,(k)}, \mathbf{M}^{(k)}, \boldsymbol{\Psi})$ with diagonal matrices $\mathbf{M}^{(k)}$ and $\boldsymbol{\Psi}$. Together, (3.5) and (3.6) are the proposed *K-component latent Gaussian-Negative Binomial*

model. By varying the number of components K, proportions p_k, and parameters of the components, the proposed model possesses ample flexibility.

The number of components K needs to be specified for estimation and inference with (3.5) and (3.6). While some practical guidance is discussed in literature [50, 55, 56], we motivate the choice of K from the important classification of DE genes into PDE and NPDE, as in [7]. PDE and NPDE genes have been observed in real time-course RNA-seq studies, such as the light physiology experiment discussed in Sections 3.1 and 4.4. To model these two types of differential expression, we consider $K = 4$ components: (i) genes without DE, $\eta_{g1} = 0$ and $\boldsymbol{\eta}_{g2} = \mathbf{0}$, that is $\sigma_1^{2,(1)} = 0$ and the diagonals of $\mathbf{M}^{(1)}$ are zeros so that $(\eta_{g1}, \boldsymbol{\eta}_{g2})$ have degenerate marginal distributions with point mass at zeros; similarly, (ii) NPDE genes with only time-by-treatment interaction, $\eta_{g1} = 0$ and $\boldsymbol{\eta}_{g2} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$; (iii) PDE genes, $\eta_{g1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\boldsymbol{\eta}_{g2} = \mathbf{0}$; and (iv) NPDE genes with both treatment and time-by-treatment effects, $\eta_{g1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\boldsymbol{\eta}_{g2} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$. In view of $\delta_{\text{MAP}}(\mathbf{Y})$ in (3.4), the null set Δ_0 in (3.1) can then be specified to correspond to components of the mixture. For example,

$$\Delta_0^{\text{Mean}} = \{\eta_{g1} = 0\},\ \Delta_0^{\text{NPDE}} = \{\boldsymbol{\eta}_{g2} = \mathbf{0}\},\ \text{and}\ \Delta_0^{\text{DE}} = \Delta_0^{\text{Mean}} \cap \Delta_0^{\text{NPDE}} \tag{3.7}$$

correspond to the first and second components, the first and third components, and only the first component in above model, respectively. The alternatives to these nulls correspond to biologically interesting hypotheses: any mean shift (with or without NPDE) is the alternative to Δ_0^{Mean}, any NPDE is the alternative to Δ_0^{NPDE}, and any DE is the alternative to Δ_0^{DE}.

Estimates of the unknown parameters μ_1, σ_1^2, the diagonal entries of $\boldsymbol{\Psi}$ and $\mathbf{M}$, and the proportions p_k are needed to conduct the test. In the Appendix B, we detail a quasi-Monte Carlo integration-assisted gradient EM algorithm for estimation. Specify Δ_0 as in (3.7) and let $\boldsymbol{\beta} = (\eta_1, \boldsymbol{\eta}_2, \boldsymbol{\tau})$. To perform the test, we compute the plug-in estimate

$$\widehat{\delta}_{\text{MAP}}(\mathbf{Y}) = \int_{\mathbb{R}^q} \int_{\Delta_0} f(\mathbf{Y}|\mathbf{x}'\boldsymbol{\beta}, \widehat{\phi}) d\widehat{\pi}(\boldsymbol{\beta}) \{ \int_{\mathbb{R}^q} \int_{\boldsymbol{\Omega}} f(\mathbf{Y}|\mathbf{x}'\boldsymbol{\beta}, \widehat{\phi}) d\widehat{\pi}(\boldsymbol{\beta}) \}^{-1},$$

where $f(\cdot|\mathbf{x}'\boldsymbol{\beta}, \widehat{\phi})$ is the negative binomial density with estimated dispersion $\widehat{\phi}$ and matrix $\mathbf{x}_{(2q+1)\times(T_1+T_2)} = [\{\mathbf{0}_{T_1\times 1}\ \ \mathbf{B}_1'(t)\ \ \mathbf{0}_{T_1\times q}\}'\ \ \{\mathbf{1}_{T_2\times 1}\ \ \mathbf{B}_2(t)\ \ \mathbf{B}_2(t)\}']$ for which

$$\mathbf{B}_i(t) = (B_1(t), \ldots, B_q(t))_{t=t_1}^{t_{T_i}},$$

and

$$d\widehat{\pi}(\boldsymbol{\beta}) = \sum_{k=1}^{4} \widehat{p}_k \varphi(\boldsymbol{\beta}|\widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\boldsymbol{\Lambda}}^{(k)}) d\boldsymbol{\beta}$$

with multivariate normal density φ. Then, as in Section 3.2, $\widehat{\lambda}_\alpha$ is chosen so that $\widehat{\text{FDR}}_\Gamma \leq \alpha$ with $\delta_{\text{MAP}}(\mathbf{Y}_g)$ replaced by $\widehat{\delta}_{\text{MAP}}(\mathbf{Y}_g)$ for each g, and using $\widehat{\Gamma}(\alpha) = \{\mathbf{Y}_g : \widehat{\delta}_{\text{MAP}}(\mathbf{Y}_g) \leq \widehat{\lambda}_\alpha\}$, we can identify DE genes for the hypothesis defined by Δ_0 and related biological questions.

The proposed K-component latent Gaussian-Negative Binomial model in (3.5) and (3.6) is a finite mixture model for which identifiability must be established in order to draw meaningful statistical conclusions [57–59]. In Appendix B, we show that our proposed model admits finite identifiability [58] provided the number of basis functions q satisfies $2q + 1 \leq T_1 + T_2$.

3.4 Proof of Identifiablility

The proposed K-component latent Gaussian-Negative Binomial model is a finite mixture model, for which identifiability must be established in order to draw meaningful statistical conclusions. In his pioneering work, [57] formally defined identifiability for finite mixtures. Focusing on some specific families of distributions, [58] provided a sufficient and necessary condition on the identifiability of generated finite mixtures. [60] showed the identifiability of finite mixtures from binomial, Poisson, and Gaussian distributions under some conditions. Identifiability of finite mixtures of regression models has been explored in [61] and [55]. Modifying the results from [57], [62] showed the identifiability of finite mixtures of Log-gamma and reverse Log-gamma distributions. [59] extended the finite identifiability of [57] to strong identifiability, which involves higher-order derivatives of the densities. A more general type of k-strong identifiability was proposed by [63], which unifies other types of identifiability. Strong identifiability [59, 63] is only applicable to classes

of mixtures parameterized by parameters of single type, such as location or scale. Therefore, to explore the identifiability of our proposed model, we focus on finite identifiability.

3.4.1 Main Results

Let $\mathbf{L}$ denote a $m \times m$ lower triangular matrix and consider a family of distributions

$$\mathcal{F} = \{F(\mathbf{Y};\boldsymbol{\theta}) : \mathbf{Y} \in \mathbb{R}^T, \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Lambda}), \boldsymbol{\mu} \in \Theta \subseteq \mathbb{R}^m, \boldsymbol{\Lambda} = \mathbf{L}\mathbf{L}' \in \Omega \subset \mathcal{S}_m^{++}\}, \tag{3.8}$$

$$F(\mathbf{Y};\boldsymbol{\theta}) = \int_{\mathbb{R}^m} f(\mathbf{Y}|\mathbf{x}'\boldsymbol{\beta},\phi)\varphi(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Lambda})d\boldsymbol{\beta} = \prod_{t=1}^{T}\int_{\mathbb{R}^m} f(y_t|\mathbf{x}_t'\boldsymbol{\beta},\phi)\varphi(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\Lambda})d\boldsymbol{\beta},$$

where $f(y_t|\mathbf{x}_t'\boldsymbol{\beta},\phi)$ is the negative binomial density with $\mathbf{x}_t'$ the tth row of $T \times m$ design matrix $\mathbf{x}'$ and dispersion parameter ϕ, $\varphi(\cdot|\boldsymbol{\mu},\boldsymbol{\Lambda})$ is the multivariate normal density, and $\mathcal{S}_m^{++}$ is the family of $m \times m$ symmetric positive definite matrices. At most $m(m+1)/2$ entries of $\mathbf{L}$ are nonzero. Hence, $\mathcal{F}$ is indexed by points in a Borel subset of $\mathbb{R}^m \times \mathbb{R}^{m(m+1)/2}$.

Definition 3.4.1. The set of all finite mixtures generated from $\mathcal{F}$ is $\mathcal{H} = \{H(\mathbf{Y},\boldsymbol{\Theta}) : H(\mathbf{Y},\boldsymbol{\Theta}) = \sum_{k=1}^{K} p_k F(\mathbf{Y};\boldsymbol{\theta}_k), \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_\ell \text{ if } k \neq \ell, \text{ where } k,\ell = 1,2,\ldots,K, p_k > 0, \sum_{k=1}^{K} p_k = 1\}$, where $K = 1,2,\ldots$.

With $m = 2q+1$, $T = T_1 + T_2$, and $\mathbf{x}'$ and $(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Lambda}^{(k)})$'s specified in Section 3.2 in the main chapter, the proposed model is indeed a finite mixture generated by $\mathcal{F}$ with $K = 4$.

Definition 3.4.2. [58]. Consider

$$H(\mathbf{Y},\{\boldsymbol{\theta}_k\}_k) = \sum_{k=1}^{K} p_k F(\mathbf{Y};\boldsymbol{\theta}_k)$$

and

$$H'(\mathbf{Y},\{\boldsymbol{\theta}'_\ell\}_\ell) = \sum_{\ell=1}^{L} p'_\ell F(\mathbf{Y};\boldsymbol{\theta}'_\ell).$$

If $H(\mathbf{Y},\{\boldsymbol{\theta}_k\}_k) = H'(\mathbf{Y},\{\boldsymbol{\theta}'_\ell\}_\ell)$ implies $K = L$, and for each k there is an ℓ such that $p_k = p'_\ell$ and $\boldsymbol{\theta}_k = \boldsymbol{\theta}'_\ell$, then $\mathcal{F}$ generates identifiable finite mixtures or admits finite identifiability.

Theorem 3.4.3. *Consider a full column rank* $\mathbf{x}'$*. All finite mixtures generated by the family* $\mathcal{F}$ *in* (3.8) *are identifiable in the sense of Definition 3.4.2 provided the existence of an index set* $\mathbb{S} \subseteq \{1, \ldots, T\}$*, such that for all* $k = 1, \ldots, K$*, pairs*

$$\left(\left\{ \sum_{s \in \mathbb{S}} \mathbf{x}_s' \right\} \mathbf{\Lambda}_k \left\{ \sum_{s \in \mathbb{S}} \mathbf{x}_s' \right\}', \left\{ \sum_{s \in \mathbb{S}} \mathbf{x}_s' \right\} \boldsymbol{\mu}_k \right)$$

are distinct, where $\mathbf{x}_s'$ *is the rows of* $\mathbf{x}'$ *for* $s = 1, \ldots, T$*.*

The full column rank condition on $\mathbf{x}'$ is commonly used for exploring identifiability of mixture models, for example; it is sufficient for identifiability of a class of Poisson regression mixtures [64]. For our model, this condition guides us to choose the number of basis functions q so that $2q + 1 \leq T_1 + T_2$. For the *distinct pairs condition*, consider the "location mixtures" scenario in which $\mathbf{\Lambda}_k$'s are the same and $\boldsymbol{\mu}_k$'s are different.The finite mixtures generated by $\mathcal{F}$ are then identifiable if for all $k = 1, \ldots, K$, there exists an index set $\mathbb{S}$ such that $\sum_{s \in \mathbb{S}} \mathbf{x}_s' \boldsymbol{\mu}_k$ are different, which is equivalent to $\sum_{s \in \mathbb{S}} \mathbf{x}_s$ lying outside a finite number of hyperplanes for such an $\mathbb{S}$.

To complete the identifiability of our proposed model, we define mapping $\mathbf{g} : \Theta^* \times \Omega^* \to \Theta \times \Omega$ such that $\mathbf{g}(\boldsymbol{\nu}, \mathbf{\Sigma}) = (g_1(\boldsymbol{\nu}, \mathbf{\Sigma}), g_2(\boldsymbol{\nu}, \mathbf{\Sigma})) = (\boldsymbol{\mu}, \mathbf{\Lambda})$, where $\Theta^* \subset \mathbb{R}^{m'}$ with $m' \leq m$ and $\Omega^* \subset \mathcal{S}_m^{++}$. Then, we make the following assumption.

Condition 3.4.4. It holds $g(\boldsymbol{\nu}_1, \mathbf{\Sigma}_1) \neq g(\boldsymbol{\nu}_2, \mathbf{\Sigma}_2)$ whenever $(\boldsymbol{\nu}_1, \mathbf{\Sigma}_1) \neq (\boldsymbol{\nu}_2, \mathbf{\Sigma}_2)$ in $\Theta^* \times \Omega^*$.

Condition 3.4.4 and its variants have been used to study identifiability in transformed parameter space and widely discussed in the literature [65].

Theorem 3.4.5. *Consider a family of density functions* $\{G(\mathbf{Y}; \boldsymbol{\nu}, \mathbf{\Sigma}) : \boldsymbol{\nu} \in \Theta^*, \mathbf{\Sigma} \in \Omega^*\}$ *where* $G(\mathbf{Y}; \boldsymbol{\nu}, \mathbf{\Sigma}) = F(\mathbf{Y}; \mathbf{g}(\boldsymbol{\nu}, \mathbf{\Sigma}))$*. Assume that all finite mixtures generated by the family* $\mathcal{F} = \{F(\mathbf{Y}; \boldsymbol{\theta}) : \boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{\Lambda}), \boldsymbol{\mu} \in \Theta, \mathbf{\Lambda} \in \Omega\}$ *are identifiable. Then, all finite mixture generated by the family* $\{G(\mathbf{Y}; \boldsymbol{\nu}, \mathbf{\Sigma}) : \boldsymbol{\nu} \in \Theta^*, \mathbf{\Sigma} \in \Omega^*\}$ *are identifiable under Condition 3.4.4.*

For our proposed model with parameters $\mu_1, \sigma_1^2, \mathbf{\Psi}, \mathbf{M}$, $(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k)$ in Theorem 3.4.3 coincides with $(\boldsymbol{\mu}^{(k)}, \mathbf{\Lambda}^{(k)})$ in Section 3.3 in the main chapter for $k = 1, 2, 3, 4$. That is, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 =$

$\mathbf{0}_{(2q+1)\times 1}$, $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = \left[\mu_1, \mathbf{0}_{1\times(2q)}\right]'$, where $\mu_1 \neq 0$;

$$\boldsymbol{\Lambda}_1 = \begin{pmatrix} 0 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \boldsymbol{\Psi} \end{pmatrix}, \quad \boldsymbol{\Lambda}_2 = \begin{pmatrix} 0 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{M} & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \boldsymbol{\Psi} \end{pmatrix}$$

and

$$\boldsymbol{\Lambda}_3 = \begin{pmatrix} \sigma_1^2 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \boldsymbol{\Psi} \end{pmatrix}, \quad \boldsymbol{\Lambda}_4 = \begin{pmatrix} \sigma_1^2 & \mathbf{0}_{1\times q} & \mathbf{0}_{1\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{M} & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times 1} & \mathbf{0}_{q\times q} & \boldsymbol{\Psi} \end{pmatrix},$$

where $\sigma_1^2 > 0$, $\boldsymbol{\Psi}$ and $\mathbf{M}$ are diagonal matrices with positive diagonals. It is easy to see that the transformation between $(\mu_1, \sigma_1^2, \boldsymbol{\Psi}, \mathbf{M})$ and $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$'s satisfies Condition 3.4.4. In addition, for $\mathbf{x}'$ given in Section 3.2 in the main chapter, the full column rank condition is satisfied and it is easy to see that there exists $\mathbb{S}$ such that all entries in $\sum_{s\in\mathbb{S}} \mathbf{x}'_s$ are not zero, which guarantees distinct $\left\{\sum_{s\in\mathbb{S}} \mathbf{x}'_s\right\} \boldsymbol{\Lambda}_k \left\{\sum_{s\in\mathbb{S}} \mathbf{x}'_s\right\}'$ for all $k = 1, \ldots, K = 4$. Therefore, Theorems 3.4.3 and 3.4.5 establish the identifiability of our proposed model.

3.4.2 Proof of Theorem 3.4.3

Assume that $\mathcal{F}$ is not identifiable. Then, there exists $M \geq 1$ and nonzero $\alpha_j \in \mathbb{R}$ for $j = 1, \ldots, M$ such that

$$\sum_{j=1}^{M} \alpha_j F(\mathbf{Y}; \boldsymbol{\theta}_j) = \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^m} \prod_{t=1}^{T} f(y_t | \mathbf{x}'_t \boldsymbol{\beta}, \phi) \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta} = 0, \tag{3.9}$$

where $F(\mathbf{Y}, \boldsymbol{\theta}_j) \in \mathcal{F}$ for each j. For $\mathbb{S} \subseteq \{1, \ldots, T\}$ and $u \in \mathbb{Z}^+ \cup \{0\}$, integrating $\prod_{s\in\mathbb{S}} y_s^u$ on both sides of (3.9) and using Fubini's Theorem leads to

$$
\begin{aligned}
0 &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(\prod_{s \in \mathbb{S}} y_s^u \right) \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^m} \prod_{t=1}^{T} f(y_t | \mathbf{x}_t' \boldsymbol{\beta}, \phi) \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta} dy_1 \cdots dy_{s \in \mathbb{S}} \cdots dy_T \\
&= \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^p} \prod_{s \in \mathbb{S}} \left\{ \int_{\mathbb{R}} y_s^u f(y_s | \mathbf{x}_s' \boldsymbol{\beta}, \phi) dy_s \right\} \left\{ \prod_{t \notin \mathbb{S}} \int_{\mathbb{R}} f(y_t | \mathbf{x}_t' \boldsymbol{\beta}, \phi) dy_i \right\} \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta} \\
&= \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^m} \prod_{s \in \mathbb{S}} \mathbb{E} \left(y_s^u | \mathbf{x}_s' \boldsymbol{\beta}, \phi \right) \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta}.
\end{aligned}
$$

By Lemma 3.4.6 and induction, for each $u \in \mathbb{Z}^+ \cup \{0\}$,

$$
\begin{aligned}
0 &= \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^m} \prod_{s \in \mathbb{S}} \exp\left(u \mathbf{x}_s' \boldsymbol{\beta} \right) \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta} \\
&= \sum_{j=1}^{M} \alpha_j \int_{\mathbb{R}^m} \exp\left(u \sum_{s \in \mathbb{S}} \mathbf{x}_s' \boldsymbol{\beta} \right) \varphi(\boldsymbol{\beta} | \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) d\boldsymbol{\beta} \\
&= \sum_{j=1}^{M} \alpha_j \exp\left\{ \frac{u^2}{2} \left(\sum_{s \in \mathbb{S}} \mathbf{x}_s' \right) \boldsymbol{\Lambda}_j \left(\sum_{s \in \mathbb{S}} \mathbf{x}_s' \right)' + u \left(\sum_{s \in \mathbb{S}} \mathbf{x}_s' \right) \boldsymbol{\mu}_j \right\}
\end{aligned}
\tag{3.10}
$$

with $\alpha_j \neq 0$ for some j. Given distinct pairs $(\{\sum_{s \in \mathbb{S}} \mathbf{x}_s'\} \Lambda_j \{\sum_{s \in \mathbb{S}} \mathbf{x}_s'\}', \{\sum_{s \in \mathbb{S}} \mathbf{x}_s'\} \boldsymbol{\mu}_j) := (s_j^2, m_j)$ for $j = 1, \ldots, M$, (3.10) shows that the family of normal distributions $\{\mathcal{N}(m_j, s_j^2)\}_{j=1}^M$ is linearly dependent and therefore not identifiable by the main theorem of [58]. However, as shown by [57], the class of all finite mixtures of univariate normal distributions is identifiable. Therefore, we reach a contradiction. Hence, $\mathcal{F}$ must be linearly independent and generates identifiable mixtures.

3.4.3 Proof of Theorem 3.4.5

Proof. Consider different pairs $(\boldsymbol{\nu}_1, \boldsymbol{\Sigma}_1), \ldots, (\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k) \in \Theta^* \times \Omega^*$ for $k \geq 2$. Assume that there exists $\alpha_i \in \mathbb{R}$ such that

$$
\sum_{i=1}^{k} \alpha_i G(\mathbf{Y}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i) = 0. \tag{3.11}
$$

Let $(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) := \mathrm{g}(\boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i)$ for each $i \leq k$. By Condition 4.3, $(\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_1), \ldots, (\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ are all distinct. Equation (3.11) therefore yields

$$\sum_{i=1}^{k} \alpha_i F(\mathbf{Y}; \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = 0. \tag{3.12}$$

It is known that all finite mixtures generated by the family $\{F(\mathbf{Y}; \boldsymbol{\theta}) : \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Lambda})\}$ are identifiable if and only if the elements in this family are linearly independent [58]. Equation (3.12) implies that $\alpha_i = 0$ for each i. Therefore, (3.11) implies that all finite mixtures generated from $\{G(\mathbf{Y}; \boldsymbol{\nu}, \boldsymbol{\Sigma}) : \boldsymbol{\nu} \in \Theta^*, \boldsymbol{\Sigma} \in \Omega^*\}$ are identifiable. □

3.4.4 A Technical Lemma

Lemma 3.4.6. The nth-order moment of the negative binomial distribution is an n^{th}-degree polynomial in its first-order moment.

Proof. Consider parametrization of the negative binomial distribution by probability of success p and predefined number of failures r. The corresponding moment generating function is $M(t) = p^r\{u(t)\}^{-r}$, where $u(t) = 1 - (1-p)e^t$. Then,

$$M^{(1)}(t) := \frac{\partial M(t)}{\partial t} = p^r(-r)u^{-r-1}\frac{\partial u}{\partial t},$$

and

$$M^{(2)}(t) := \frac{\partial M^{(1)}(t)}{\partial t} = M^{(1)}(t) + p^r(-r)(-r-1)u^{-r-2}\left(\frac{\partial u}{\partial t}\right)^2.$$

By induction, assume that for all $i = 2, \ldots, n$,

$$M^{(i)}(t) = P_i\left(M^{(1)}(t), \ldots, M^{(i-1)}(t)\right) + p^r(-1)^i\frac{\Gamma(r+i)}{\Gamma(r)}u^{-r-i}\left(\frac{\partial u}{\partial t}\right)^i,$$

where $P_i\left(M^{(1)}(t), \ldots, M^{(i-1)}(t)\right)$ is a linear combination of $M^{(1)}(t), \ldots, M^{(i-1)}(t)$. It yields

$$\begin{aligned} M^{(n+1)}(t) = {} & P_n\left(M^{(2)}(t), \ldots, M^{(n)}(t)\right) + n\left\{M^{(n)}(t) - P_n\left(M^{(1)}(t), \ldots, M^{(n-1)}(t)\right)\right\} \\ & + p^r(-1)^{n+1}\frac{\Gamma(r+n+1)}{\Gamma(r)}u^{-r-n-1}\left(\frac{\partial u}{\partial t}\right)^{n+1}. \end{aligned}$$

Therefore, for any $n > 1$, we have

$$M^{(n)}(t) := P_n\left(M^{(1)}(t), \ldots, M^{(n-1)}(t)\right) + p^r(-1)^n \frac{\Gamma(r+n)}{\Gamma(r)} u^{-r-n} \left(\frac{\partial u}{\partial t}\right)^n.$$

On the other hand, the first-order moment of negative binomial random variable Y is $\mathbb{E}(Y) = M^{(1)}(0) = r(1-p)p^{-1}$. Since $u(0) = p$, $\partial u/\partial t|_{t=0} = -(1-p)$, we have

$$\mathbb{E}(Y^n) = M^{(n)}(0) = P_n\left(M^{(1)}(0), \ldots, M^{(n-1)}(0)\right) + \frac{\Gamma(r+n)}{\Gamma(r)}(-1)^n \left\{\frac{\mathbb{E}(Y)}{r}\right\}^n,$$

where $P_n\left(M^{(1)}(0), \ldots, M^{(n-1)}(0)\right)$ is a polynomial in $\mathbb{E}(Y)$ of degree $n-1$. □

3.5 Monte Carlo Evidence

In this section, we examine the performance of the proposed method and other existing approaches, including `DESeq2`, `edgeR` [66], `ImpulseDE2`, `maSigPro-GLM`, and `splineTC`, using two extensive sets of simulation studies. We generate time-course count data according to (3.5), (3.6) and a variety of Δ_0's.

3.5.1 Simulation Setting

For Settings A and B, detailed below, we simulate 100 independent datasets of time-course count data, with each dataset containing $G = 1,000$ genes, two treatment groups, $T = 6$ or $T = 10$ time points, and $r = 3$ or $r = 6$ replicates. Three types of basis functions with $q = 2$ or $q = 3$ are considered for estimating (3.5): the basis functions for the traditional Gaussian kernel (see [54] and the Appendix) denoted GA_2 and GA_3; the orthogonal Fourier basis functions, denoted FO_2 and FO_3; and the orthogonal polynomial bases, denoted PL_2 and PL_3. Also, we set $S_{ij} \equiv 1$.

Setting A: In (3.5), $q = 2$ and the true basis functions are PL_2. Parameters $\eta_{g1}, \boldsymbol{\eta}_{g2}$ and $\boldsymbol{\tau}_g$ are drawn from (3.6), where $\mu_1 = 2$ or 4, $\sigma_1^2 = 1$ or 2, $\boldsymbol{\Psi} = \text{diag}\{1, 1\}$ and $\mathbf{M} = \text{diag}\{3, 2\}$.

Setting B: In (3.5), $q = 3$ and the true basis functions are PL_3. Parameters $\eta_{g1}, \boldsymbol{\eta}_{g2}$ and $\boldsymbol{\tau}_g$ are drawn from (3.6), where $\mu_1 = 2$ or 4, $\sigma_1^2 = 1$ or 2, $\boldsymbol{\Psi} = \text{diag}\{1, 1, 1\}$ and $\mathbf{M} = \text{diag}\{3, 2, 1\}$.

For each simulated dataset, we fit (3.5) with each of the six basis types and conduct the test accordingly. We also conduct the *true test*, using the true basis functions (e.g., PL_2 in setting A) and known parameters, and the *oracle test*, using the true basis functions but unknown parameters. `ImpulseDE2` and `splineTC` use their own mean models; we use the true basis functions for all other competing methods. Dispersion parameters are estimated, by borrowing information across genes if applicable, for competing methods except `maSigPro-GLM`, for which the true dispersion parameters are used, and `splineTC`, for which no dispersion is assumed. For our proposed method, the moment estimator for dispersion is developed and detailed in the Appendix. Extra simulation results for our method with different dispersion estimators yield similar results; see the Appendix B (Figure S.7 in Section C.3 and related discussion). Empirical FDRs and powers averaged over 100 replications at nominal levels $0.01, 0.025, 0.05$ 0.075 and 0.1 are reported for each method.

For each setting, 50% of genes are drawn from the null component of (3.6), 20% from NPDE with only time-by-treatment, 20% from PDE with no time-by-treatment; and 10% from NPDE with both treatment and time-by-treatment. In Appendix B, additional results for different proportions of DE genes are reported and show similar patterns (Figure S.7).

3.5.2 Results

We consider the three null hypotheses in (3.7): any temporal DE is the alternative to Δ_0^{DE}, relative mean shift is the alternative to Δ_0^{Mean}, and NPDE is the alternative to Δ_0^{NPDE}. For testing Δ_0^{DE} we compare the performance of our proposal to all five competitors. By specifying corresponding contrasts or likelihood ratio statistics, `edgeR` and `DESeq2` can test the other two composite hypotheses and are compared with our method as well.

Figures 3.2–3.3 display results for the proposed method using the GA_3 basis, *true test*, *oracle test*, and the five competing methods. Each point on the figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. The closer the point is to the corresponding vertical dashed line, the more the

empirical and nominal FDR levels coincide. Results for the proposed method with other basis functions are qualitatively similar; see Figures S.1–S.6 in the Appendix.

Most points on these plots are to the left of corresponding vertical dashed lines, which suggests that most methods have their empirical FDRs controlled. Compared to our proposed method, all other competitors are less powerful, particularly for small number of replicates r and lower μ. The *true test* with known parameters and the *oracle test* without information on parameters perform the best overall, and are almost indistinguishable for most cases, reflecting the reliability of the estimation procedure described in the Appendix. Though number and type of the basis functions are all misspecified, our proposal still provides satisfactory results, reflecting the expected flexibility and robustness of the method (Figures S.1-S.6 in the Appendix). In addition, as the number of replicates r increases, the deviations between the proposed method with misspecified bases and the *true test* quickly diminish. For Setting A with PL_2 as the bases, all methods have the nominal FDRs controlled, as shown in Figure 3.2. For Setting B with PL_3, Figure 3.3 shows that `maSigPro-GLM` may not provide a satisfactory control on the FDR for all levels as it essentially models the mean dynamics by a less flexible quadratic regression. It is interesting to observe that in both figures, compared to `edgeR` and `ImpulseDE2`, `DESeq2` and `splineTC` have smaller empirical FDRs but comparable powers. Similar patterns also display in Tables 3.1 and 3.2.

Results for testing against Δ_0^{Mean} and Δ_0^{NPDE} under Setting A are reported in Tables 3.1 and 3.2, respectively. Those for Setting B are deferred to the Tables S.1–S.4 in the Appendix. Empirical FDRs are close to the nominal level for most tests for both hypotheses. The empirical powers for testing against Δ_0^{Mean} increase as μ_1 increases from 2 to 4 for all methods, while they are not substantially improved by increasing T and r. On the other hand, similar to testing the interactions in functional ANOVA [7, 67], testing for NPDE is more challenging and the power is expected to be lower. However, as displayed in Table 3.2, as well as in Tables S.3–S.4 in the Appendix, the powers for testing NPDE genes of the proposed method increase in T and r for all bases and are unsurprisingly not influenced by mean parameter μ_1. Furthermore, the performance of the proposal is not affected much by the choice of the bases. In addition, our proposed method

outperforms `edgeR` and `DESeq2` in power, particularly for testing NPDE genes. These numerical results confirm the theoretical guidance in Section 3.2 on the optimality of the proposed method, as well as its flexibility on testing a variety of composite and biologically interesting hypotheses.

3.6 Analysis of the *Phaeodactylum* Light Experiment

We apply our method to *P.t.* transcriptomics data from the light experiment introduced in Section 3.1. Most algae and cyanobacteria, including *P.t.*, undergo some major changes in cell biochemistry via a process termed photoacclimation. Growth in high light environments leads to low photosynthetic pigment per cell, accumulation of fats and carbohydrates, and an upregulated oxidative stress response. Growth in low light environments leads to high pigmentation of cells in order to capture more light and major increase in structural lipids associated with the chloroplast [68].

Peers et al. [2] investigated the gene transcriptional response during photoacclimation, to better understand regulation of photosynthetic and catabolic metabolisms during the shift of a day-night cycle. Cultures of the diatom *P.t.* were submitted to a single step change in light from excess light fluxes to low light fluxes that nearly limit the growth rate. Samples for transcriptomics were taken over a 24-hour period. After mapping to the genome, $12,319$ candidate genes were to be analyzed from the two groups with three replicates within each. Data for both groups were collected at 0, 60, 240, and 360 minutes as well as 12 and 24 hours, while data for the low light group were also collected at 20, 40, 90, and 120 minutes.

We first filtered and normalized the data following standard procedures, described in detail in Appendix B. After filtering, $10,597$ genes remain for further analysis.

For DE analysis, the nominal FDR level is set at 0.05 and we use polynomial basis functions with $q = 2$ for $\mathbf{B}(t)$. Results for all basis functions discussed in Section 3.5 yield similar results; for example, the majority of DE genes ($\sim 91\%$) for the hypotheses of interest are identified by all six basis functions. We identify $3,869$ overall temporally differential expressed genes, within which $3,771$ genes have relative mean shift between two groups (reject Δ_0^{Mean}) and 98 are NPDE (reject

Δ_0^{NPDE}). Figure S.8 displays the top 10 selected genes identified with both relative mean shift and NPDE. The top 10 genes selected with only relative mean shifts are displayed in Figure S.8 in Appendix B .

As a large proportion of predicted gene models for *P.t.* still encode genes of unknown function [69], functions of the top 10 differentially regulated genes in Figure S.8 are not completely documented. However, gene Phatr3_J47271 is an experimentally-verified critical component of the photosynthetic pigment biosynthesis pathway and our finding on its differential regulation supports the large dynamical change in pigmentation observed during photoacclimation [68]. In addition, genes Phatr3_J50183, Phatr3_J49693, and Phatr3_EG01882 are all highly related to photoacclimation. From Figure S.8, their temporal differential expression patterns are much more sophisticated than simple mean shift targeted by the traditional RNA-seq analysis. The new tests have revealed potentially critical pathways for better understanding photoacclimation in general. Follow-up analysis via gene set enrichment for three critical photoacclimation-related pathways is discussed in Appendix B.

3.7 Discussion

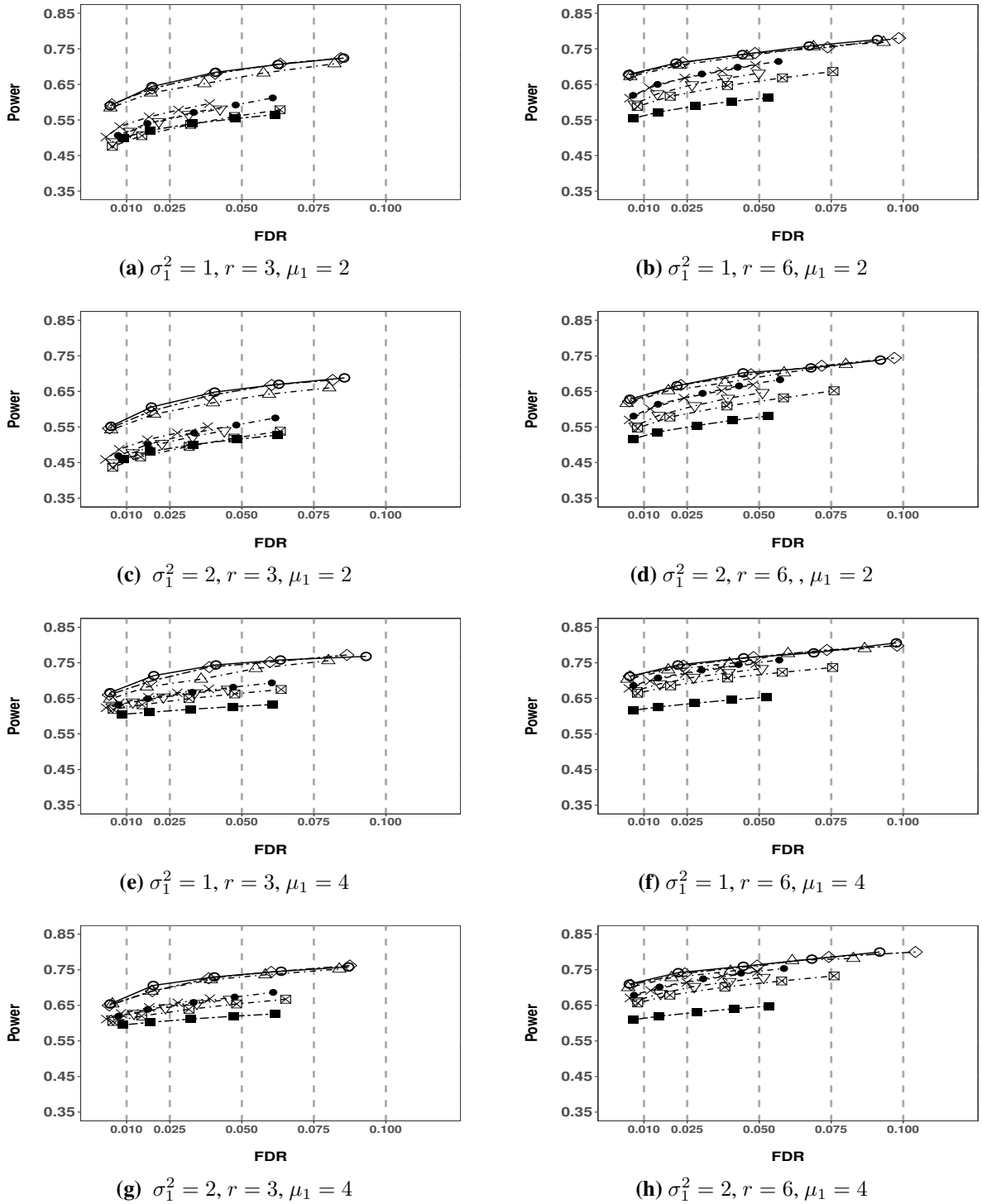
We have proposed an inferential procedure for time-course RNA-seq data that maximizes average power and controls FDR. Novel features of our approach include smooth and flexible modeling of the mean dynamic (3.5) within genes, conditional on a latent zero-mean Gaussian process; and flexible mixture modeling (3.6) of coefficients across genes. Taken together, the within-gene/across-gene model (our K-component latent Gaussian-Negative Binomial model) allows feasible estimation of unknown model parameters, natural borrowing of information across genes for both the mean and variance instead of the dispersion only, and straightforward, one-step testing of general composite null hypotheses of great biological interest. By contrast, existing pipelines such as `edgeR` and `DESeq2` have to rely on a two-step procedure for testing some composite hypotheses. Additional simulations in the Appendix B (see Table S.5) show the superior performance of our proposal over existing methods for testing these more general composite null hypotheses.

Our proposed methodology is easy to implement. However, as suggested by a numerical study reported in the Appendix B, it incurs some computational cost on parameter estimation via gradient-EM and quasi-Monte Carlo integration, which can be partly offset by parallel computations in practice. In addition, when the true alternatives are not at all smooth, such as an abrupt change in mean for a single time point, our method may have compromising powers, which can also be partly resolved by using different basis functions as mentioned in Section D.4 in the Appendix B.

Our test is numerically indistinguishable from the *oracle* and *true* tests when the number of replicates r is relatively large, while the difference is more obvious for small r. One reason for this is the estimation of gene-wise dispersion ϕ. We employ a moment estimate for ϕ inducing extra variations for our test. It has been long recognized that the unstable estimation of the dispersion in the small RNA-seq experiments may result in poor control of FDR while a good estimate of this parameter is challenging in practice [31, 46, 52]. Existing pipelines usually address this challenge by borrowing information across genes via an empirical Bayes approach. In the Appendix B, additional simulations (see Figure S.7) are reported for our test combined with different estimates on the dispersion, including the empirical dispersion estimator in Section B.1.2 (ED), the common dispersion estimator (CD), and the local regression approach used in `DESeq2`. While the empirical powers of our test with the three estimates of dispersion are similar with minor differences, the FDR control of ED is slightly inflated when $r = 3$ and the DE proportion is small. When r increases, the performances among different approaches are comparable. This relatively robust performance with respect to different approaches can be possibly explained as follows. Compared to the traditional RNA-seq trials with only r replicates, dispersion estimation for time-course experiments may leverage more from rT samples collected at T time points. Also, `edgeR` is less conservative than `DESeq2` in general, which is due to the robust empirical Bayes method employed in `edgeR` to estimate the prior degree of freedom for the weighted likelihood as it reduces the informativeness of prior distribution for outlier genes [70,71]. In contrast, using a data-adaptive control on the shrinkage, `DESeq2` controls the FDR better than `edgeR` for small r. Employing the

`edgeR` approach to estimate gene-wise dispersion may improve our method for large r in practice and is left to future study.

Though the proposed method focuses on two-sample problems, it is readily generalized to $M \geq 3$ treatments by extending the specification of $\lambda_{gij}(t)$ in (3.5) and the mixture distribution on its coefficients in (3.6). A further extension is to allow the temporal dynamics to vary continuously with respect to some explanatory variable $\mathcal{Z}$, such as age or blood pressure. Extending the mean specification (3.5) to this continuous case is straightforward, but the resulting continuous mixture distributions on model coefficients lead to more involved questions of identifiability and estimation. We leave these as topics for future investigation.



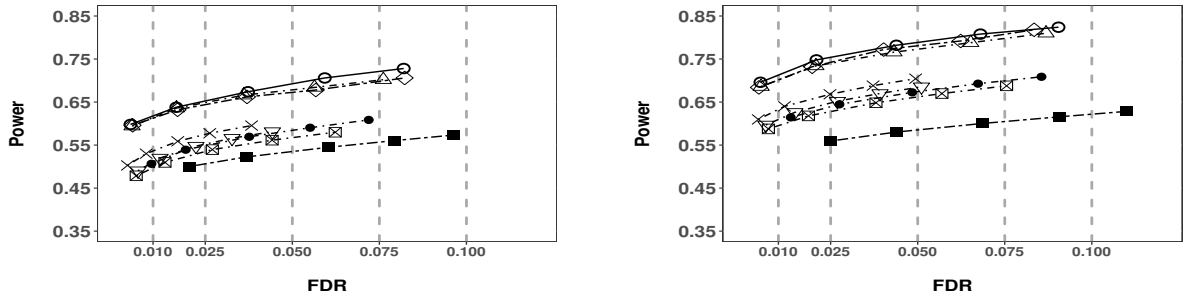
(a) $\sigma_1^2 = 1, r = 3, \mu_1 = 2$ **(b)** $\sigma_1^2 = 1, r = 6, \mu_1 = 2$

(c) $\sigma_1^2 = 2, r = 3, \mu_1 = 2$ **(d)** $\sigma_1^2 = 2, r = 6, , \mu_1 = 2$

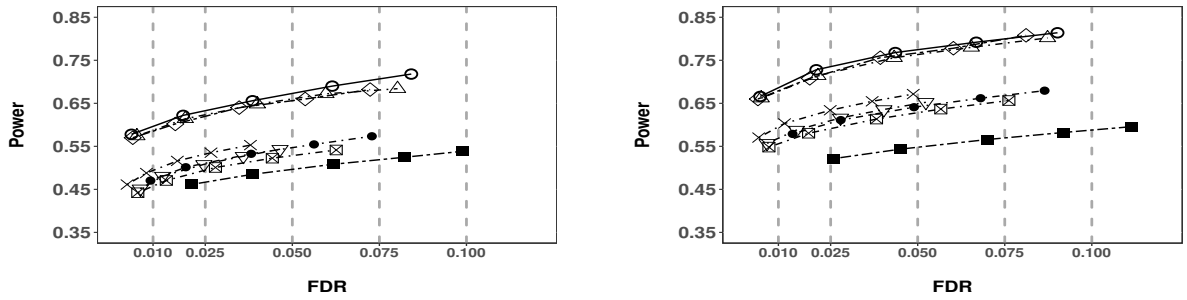
(e) $\sigma_1^2 = 1, r = 3, \mu_1 = 4$ **(f)** $\sigma_1^2 = 1, r = 6, \mu_1 = 4$

(g) $\sigma_1^2 = 2, r = 3, \mu_1 = 4$ **(h)** $\sigma_1^2 = 2, r = 6, \mu_1 = 4$

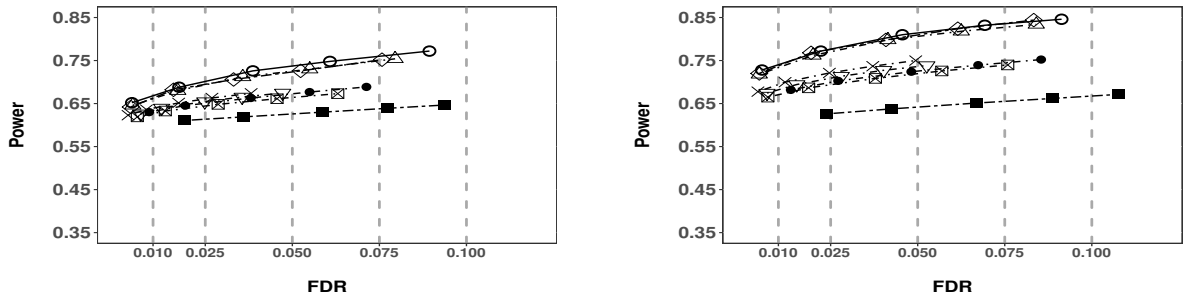
Figure 3.2: Empirical FDRs and powers for testing the overall temporal DE genes by the proposed method using GA_3 basis (Δ), compared to those of the oracle test ($\circ$), the true test ($\diamond$), `maSigPro-GLM`($\blacksquare$), `edgeR` ($\bullet$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$) and `DESeq2` ($\times$) for Setting A. Each point on figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. All plots are for $T = 10$. Results are based on 100 replications.



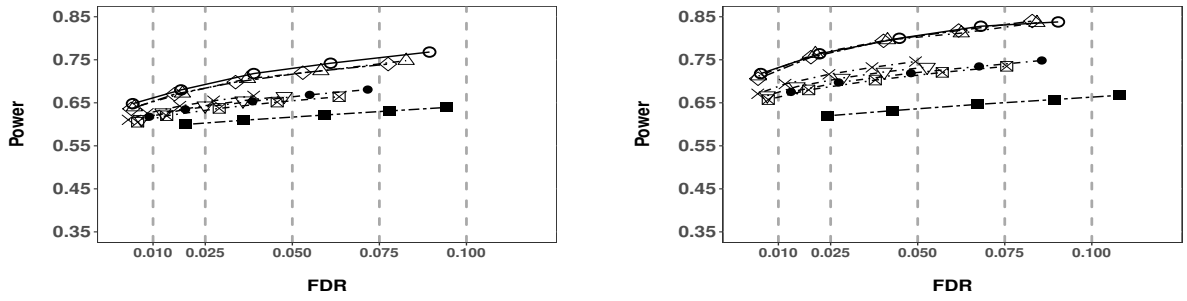
(a) $\sigma_1^2 = 1, r = 3, \mu_1 = 2$ **(b)** $\sigma_1^2 = 1, r = 6, \mu_1 = 2$



(c) $\sigma_1^2 = 2, r = 3, \mu_1 = 2$ **(d)** $\sigma_1^2 = 2, r = 6, \mu_1 = 2$



(e) $\sigma_1^2 = 1, r = 3, \mu_1 = 4$ **(f)** $\sigma_1^2 = 1, r = 6, \mu_1 = 4$



(g) $\sigma_1^2 = 2, r = 3, \mu_1 = 4$ **(h)** $\sigma_1^2 = 2, r = 6, \mu_1 = 4$

Figure 3.3: Empirical FDRs and powers for testing the overall temporal DE genes by the proposed method using GA_3 basis (Δ), compared to those of the oracle test ($\circ$), the true test ($\diamond$), `maSigPro-GLM`($\blacksquare$), `edgeR` ($\bullet$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$) and `DESeq2` ($\times$) for Setting B. Each point on figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. All plots are for $T = 10$. Results are based on 100 replications.

Table 3.1: Comparison of empirical FDRs and powers for testing DE genes with relative mean shift by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting A. In simulations, μ_1, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

	(T, r, σ_1^2)	$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
					$\mu_1 = 2$				
GA_2	FDR	0.053	0.061	0.055	0.052	0.052	0.055	0.036	0.044
	Power	0.917	0.800	0.937	0.893	0.933	0.853	0.967	0.890
GA_3	FDR	0.057	0.066	0.061	0.056	0.056	0.059	0.035	0.045
	Power	0.917	0.790	0.933	0.893	0.930	0.853	0.967	0.880
FO_2	FDR	0.053	0.075	0.072	0.073	0.053	0.078	0.047	0.045
	Power	0.920	0.800	0.933	0.890	0.933	0.860	0.963	0.887
FO_3	FDR	0.058	0.072	0.068	0.060	0.057	0.070	0.044	0.045
	Power	0.920	0.793	0.937	0.893	0.930	0.853	0.960	0.880
PL_3	FDR	0.057	0.051	0.051	0.046	0.057	0.050	0.045	0.045
	Power	0.920	0.790	0.930	0.893	0.930	0.857	0.967	0.880
Oracle	FDR	0.052	0.049	0.048	0.045	0.052	0.047	0.056	0.045
	Power	0.920	0.797	0.933	0.890	0.933	0.853	0.967	0.887
True	FDR	0.043	0.044	0.044	0.045	0.045	0.045	0.046	0.047
	Power	0.913	0.803	0.933	0.890	0.930	0.853	0.967	0.887
`edgeR`	FDR	0.072	0.072	0.049	0.049	0.051	0.052	0.045	0.044
	Power	0.796	0.741	0.882	0.820	0.862	0.801	0.919	0.864
`DESeq2`	FDR	0.021	0.022	0.032	0.031	0.031	0.030	0.036	0.036
	Power	0.799	0.730	0.881	0.814	0.861	0.792	0.919	0.861
					$\mu_1 = 4$				
GA_2	FDR	0.067	0.069	0.017	0.051	0.038	0.052	0.006	0.050
	Power	1.000	0.997	1.000	1.000	1.000	0.990	1.000	0.997
GA_3	FDR	0.084	0.081	0.054	0.058	0.058	0.062	0.012	0.047
	Power	1.000	0.997	1.000	1.000	1.000	0.990	1.000	0.997
FO_2	FDR	0.060	0.067	0.027	0.055	0.041	0.061	0.043	0.060
	Power	1.000	0.997	1.000	1.000	1.000	0.987	1.000	0.993
FO_3	FDR	0.083	0.079	0.046	0.055	0.053	0.060	0.012	0.057
	Power	1.000	0.997	1.000	1.000	1.000	0.987	1.000	0.997
PL_3	FDR	0.072	0.068	0.051	0.053	0.058	0.060	0.012	0.048
	Power	1.000	0.997	1.000	1.000	1.000	0.990	1.000	0.993
Oracle	FDR	0.064	0.061	0.019	0.051	0.038	0.048	0.012	0.048
	Power	1.000	0.997	1.000	1.000	1.000	0.990	1.000	0.997
True	FDR	0.021	0.045	0.008	0.049	0.011	0.047	0.005	0.044
	Power	1.000	0.997	1.000	1.000	1.000	0.987	1.000	0.997
`edgeR`	FDR	0.068	0.068	0.049	0.049	0.051	0.051	0.045	0.045
	Power	0.997	0.977	0.999	0.989	0.999	0.985	1.000	0.992
`DESeq2`	FDR	0.022	0.022	0.032	0.032	0.031	0.031	0.036	0.036
	Power	0.996	0.975	0.999	0.988	0.999	0.985	1.000	0.992

Table 3.2: Comparison of empirical FDRs and powers for testing NPDE genes by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting A. In simulations, μ_1, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

	(T, r, σ_1^2)	$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
					$\mu_1 = 2$				
GA_2	FDR	0.025	0.027	0.040	0.038	0.031	0.036	0.036	0.043
	Power	0.083	0.067	0.183	0.213	0.147	0.147	0.323	0.370
GA_3	FDR	0.024	0.025	0.029	0.029	0.027	0.030	0.035	0.031
	Power	0.060	0.067	0.153	0.183	0.110	0.120	0.307	0.323
FO_2	FDR	0.026	0.055	0.057	0.062	0.033	0.065	0.047	0.043
	Power	0.080	0.070	0.163	0.163	0.140	0.130	0.263	0.370
FO_3	FDR	0.019	0.038	0.035	0.041	0.026	0.048	0.044	0.030
	Power	0.060	0.060	0.137	0.163	0.113	0.113	0.277	0.333
PL_3	FDR	0.021	0.022	0.026	0.025	0.024	0.031	0.041	0.031
	Power	0.067	0.060	0.147	0.157	0.113	0.100	0.307	0.327
Oracle	FDR	0.025	0.028	0.036	0.037	0.031	0.034	0.050	0.042
	Power	0.070	0.067	0.187	0.193	0.143	0.140	0.333	0.367
True	FDR	0.034	0.029	0.040	0.038	0.034	0.035	0.044	0.040
	Power	0.097	0.087	0.203	0.210	0.153	0.140	0.327	0.360
`edgeR`	FDR	0.096	0.105	0.044	0.047	0.047	0.048	0.041	0.040
	Power	0.013	0.013	0.103	0.102	0.069	0.066	0.243	0.245
`DESeq2`	FDR	0.020	0.019	0.033	0.034	0.031	0.027	0.039	0.037
	Power	0.013	0.011	0.103	0.098	0.071	0.067	0.237	0.235
					$\mu_1 = 4$				
GA_2	FDR	0.028	0.026	0.038	0.034	0.035	0.033	0.047	0.052
	Power	0.090	0.093	0.257	0.240	0.143	0.177	0.337	0.377
GA_3	FDR	0.021	0.025	0.027	0.026	0.026	0.028	0.034	0.038
	Power	0.080	0.077	0.207	0.190	0.113	0.170	0.297	0.313
FO_2	FDR	0.059	0.055	0.066	0.060	0.067	0.067	0.058	0.060
	Power	0.107	0.093	0.213	0.190	0.133	0.133	0.287	0.310
FO_3	FDR	0.043	0.040	0.041	0.039	0.047	0.050	0.046	0.047
	Power	0.097	0.093	0.213	0.187	0.117	0.147	0.280	0.310
PL_3	FDR	0.022	0.020	0.024	0.023	0.024	0.026	0.033	0.035
	Power	0.077	0.093	0.223	0.190	0.103	0.130	0.300	0.300
Oracle	FDR	0.029	0.024	0.034	0.034	0.032	0.032	0.044	0.048
	Power	0.097	0.090	0.270	0.240	0.140	0.167	0.333	0.353
True	FDR	0.030	0.028	0.040	0.039	0.039	0.032	0.043	0.044
	Power	0.110	0.120	0.270	0.247	0.147	0.173	0.337	0.353
`edgeR`	FDR	0.091	0.080	0.045	0.047	0.046	0.055	0.039	0.042
	Power	0.016	0.014	0.110	0.109	0.076	0.076	0.249	0.249
`DESeq2`	FDR	0.032	0.045	0.032	0.034	0.029	0.034	0.036	0.039
	Power	0.017	0.014	0.109	0.109	0.075	0.075	0.243	0.245

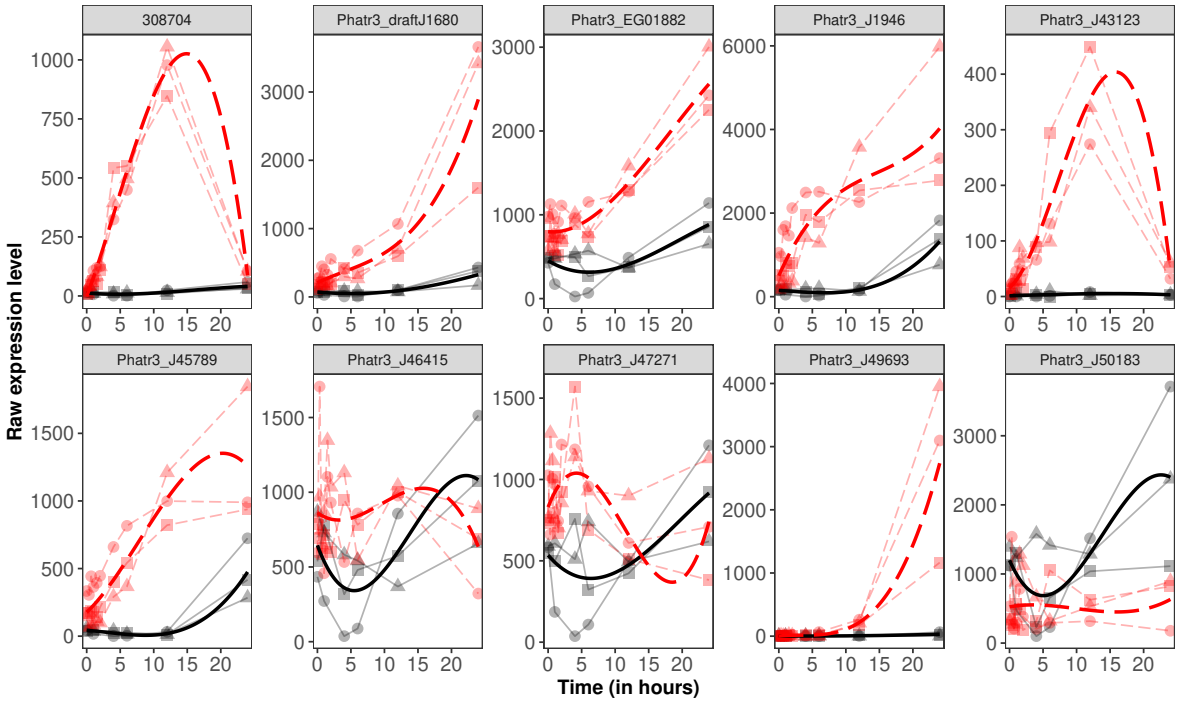


Figure 3.4: Top 10 genes identified by the proposed method with both relative mean shift and NPDE. The red dashed curves represent data from the low light group and the black solid curves represent data from high light group. Dots represent the real data points while the bold smooth curves display smooth estimations using orthogonal polynomials. Captions are the gene tags from [2].

Chapter 4

Group Structured Model with Application to Small RNA Analysis: Normalization and Differential Expression Analysis

4.1 Introduction

Studying transcriptomes through next generation sequencing (NGS) of RNA (RNA-seq) has enabled researchers to better understand the underlying mechanism of biological processes. An important task is analyzing RNA-seq data and detecting genes that are differentially expressed (DE) between experimental conditions [3,31]. Deep sequencing (an advanced type of NGS) refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times. Deep sequencing has been widely used in profiling of small noncoding RNAs [72]. Small RNAs (sRNAs) are a large class of RNAs consisting of many groups such as microRNA (miRNA), short interfering RNA (siRNA), piwi-interacting RNA (piRNA), and more [73–75]. Many miRNAs have been detected in animals and plants [76] and some work has suggested that dedicated pathways generate each class of sRNAs [73, 77]. Thus, analyzing sRNAs allows us to identify novel biomarkers [78].

Similar to RNA-seq analysis, one of the major goals of sRNA analysis is to detect DE genes across different biological conditions. For the purposes of evaluating differential expression among conditions, read counts are usually modeled at the genomic level in the traditional RNA-seq analysis. However, multiple classes of sRNAs introduce a natural grouping structure of genes, which not only generate a variety of interesting group-wise differential expression hypotheses but also introduce statistical challenges in modeling. For example, Hackenberg et al. [8] show that miRNA and other sRNAs in barley are regulated by drought. Specifically, low-expressed miRNAs and repeat-

associated siRNAs are down-regulated, while tsRNAs are up-regulated under drought conditions. This suggests that different classes of sRNAs respond differently to drought as a group.

Raw data that are directly observed from sequencing must be preprocessed and normalized before analysis. The normalization removes technical artifacts or unintended variation due to the method, while retaining the true biological signals across samples [79]. Many normalization methods have been proposed to correct biases between and within samples, such as upper quartile (UQ), trimmed mean of M-values (TMM) and relative log expression (RLE). UQ and other quantile (e.g. median) methods, which involve matching empirical distributions across samples, are widely used [9]. TMM equates the overall expression levels of genes between samples by estimating the relative RNA production levels [10]. Anders and Huber [31] proposed a relative log expression (RLE) method based on the negative binomial distribution; RLE has been used in the well-known packages `edgeR` and `DESeq2`. Both TMM and RLE assume that the majority of genes are not differentially expressed, which is not always true in the sRNA analysis due to the group effects. Dillies et al. [80] and Li et al. [81] provide a comprehensive comparison among these well known normalization procedures using differential expression analysis on RNA-seq data. However, none of these methods dominate others uniformly and none of these studies provide a way to perform normalization on group-structured data. As shown in Section 4.3, if a class of sRNAs reacts to the treatment, traditional normalization methods may fail to control the false discovery rate (FDR) or suffer from low power in differential expression analysis.

In this paper, we focus on analyzing group-structured sRNA data. We develop a unified approach for both a novel normalization method and a powerful testing procedure for DE analysis.

Generalized linear models, such as the negative binomial, have been widely used in differential expression analysis for RNA-Seq data [31,32,82,83]. For example, `edgeR` [32] and `DESeq2` [83] in the Bioconductor program can be directly applied to detecting the DE sRNAs. However, these methods may have difficulties in analyzing group-structured sRNA data.

First, these methods can suffer from low power. Previous studies indicate that `DESeq2` is more conservative than `edgeR` [7,43,84]. Thus, in Section 4.3 we will only compare the performance of

our method with edgeR. Though edgeR adopts the idea of borrowing information across genes, it does not use a group-structured model and its power is not satisfactory compared with our method when dealing with data containing group effects. When data has no group effect, we observe that edgeR and our method perform similarly. Furthermore, most existing methods, such as edgeR and DESeq2, are mainly designed for detecting gene specific overall differential patterns among treatment conditions and cannot make inference on group effects.

Some existing methods for analyzing group-structured data, including the group knockoff filter [85] and p-filter [86], consider the set of true and false discoveries at the group level and estimate the group FDR. The group knockoff filter is directly modified from the Benjamini-Hochberg (BH) procedure by estimating the expected proportion of incorrectly selected features among selected groups [85]. This method assumes that the majority of groups are not DE, which may not be the case for sRNA analysis. The p-filter combines the Simes test [87] and BH procedure to control the FDR at both the individual and the group level [86]. Both of these are two-stage methods that rely critically on a given set of correct p-values, which are then filtered for the group analysis. By contrast, we develop a unified procedure for simultaneous normalization and DE analysis on sRNA.

Specifically, we develop a group-based negative binomial model for sRNA data to address the above challenges, including normalization and DE analysis. In our model, read counts of an sRNA in each sample have negative binomial distributions, with the mean profile generated by both group and gene-specific effects. We then develop an efficient algorithm to estimate the model parameters, using a weighted generalized linear model, which provides the estimation of group-based normalization factors. We introduce a testing framework based on the proposed model. The proposed model and test allow not only identification of the traditional DE genes but also inference on group effects. To demonstrate the advantages of our group-structured model in comparison to other well-known methods, we perform comprehensive simulation studies. As an example, we consider an experiment to explore the roles of piRNAs and WAGO-class 22G-RNAs in regulating gene expression in the roundworm (C. elegans). In the experiment, we collect independent repli-

cates of sRNAs from three treatment conditions: adult wild type (wt) animal, prg-1(n4357) and mut-16(pk710) mutants. Using our proposal, we are able to make inference on group effects. For example, we observe that piRNAs are down-regulated in prg-1 mutants, which agrees with the previous biology conjecture ("prg-1 is the known binding partner of piRNAs in C. elegans") [88, 89]. On the other hand, by analyzing individual gene effects, we discover a list of interesting DE sRNAs that may be mis-annotated.

4.1.1 Outline

In Section 4.2, we present the group-structured model, show the details of the estimation and testing procedure for performing the DE analysis. Simulation studies are shown in Section 4.3. In Section 4.4, we analyze the C. elegans germline data using our method. We conclude with a discussion of our findings in Section 4.5.

4.2 Methods

4.2.1 Model

Let $Y_{j_g,ijk}$ denote the read count from an sRNA sequencing experiment for gene j_g ($j_g = 1, \ldots, j_G$) within group j ($j = 1, \ldots, J$), and under treatment condition i ($i = 1, 2, \ldots, T$), with $i = 1$ denoting control. The subscript k ($k = 1, \ldots, Tp$) is the combined index for p replicates across each of the T treatments. Let $\mathbf{Y}_{j_g}$ denote the vector of read counts for gene j_g across all Tp treatments and replicates. Assume that $\mathbf{Y}_{j_g}$ follows a distribution parameterized by group effect α_j, group by treatment effect β_{ij}, gene effect θ_{j_g}, and gene by treatment effect η_{ij_g}, as well as the normalization factor S_k. A widely applicable model for sRNA-seq analysis is the negative binomial (NB), due to the over-dispersed property of count data [90]. We assume $Y_{j_g,ijk} \sim \text{NB}(m_{j_g,ijk}, r_{j_g})$, where r_{j_g} is the dispersion parameter and $\mathbb{E}\left[Y_{j_g,ijk}\right] = m_{j_g,ijk}$ is further modeled as

$$m_{j_g,ijk} = \mathbb{E}[Y_{j_g,ijk}] = S_k \exp\left[\alpha_j + \beta_{ij}\mathbb{1}_{\{i\neq 1\}} + \theta_{j_g} + \eta_{ij_g}\mathbb{1}_{\{i\neq 1\}}\right], \tag{4.1}$$

for any i, where $\mathbb{1}$ is the indicator function, and

$$\sum_{k=(i-1)p+1}^{ip} \log(S_k) = 0 \text{ for all } i, \ \sum_{j_g=1}^{j_G} \theta_{j_g} = 0, \tag{4.2}$$

and

$$\sum_{j_g=1}^{j_G} \eta_{ij_g} = 0 \text{ for all } i \neq 1. \tag{4.3}$$

Constraints (4.2) and (4.3) ensure identifiability of the model by making the design matrix full rank; other constraints could be employed. By construction, $\text{var}(Y_{j_g,ijk}) = m_{j_g,ijk} + m^2_{j_g,ijk}/r_{j_g}$.

In (4.1), α_j models the mean group intercept and β_{ij} is the relative mean shift for group j under treatment i, θ_{j_g} measures the mean expression level for gene j_g and η_{ij_g} characterizes the potential interactions between the gene j_g and treatment i.

4.2.2 Hypothesis for DE Analysis

For model (4.1), we conduct the DE analysis by testing

$$H_{0,j_g} : \beta_{ij} + \eta_{ij_g} \in \Delta_0 \text{ versus } H_{1,j_g} : \beta_{ij} + \eta_{ij_g} \in \Delta_1. \tag{4.4}$$

The null space Δ_0 can be defined in different ways depending on the problems of interest. For example, if we are interested in knowing whether the mean expression levels of the sRNA in the two treatments differ, we set $\Delta_0 = 0$; if we are interested in detecting up-regulated or down-regulated sRNA, we set $\Delta_0 = (-\infty, 0]$ or $\Delta_0 = [0, \infty)$; if we are interested in detecting sRNAs whose expression changes are large enough, we set $\Delta_0 = \{d : |d| \leq v\}$, where v is a threshold.

Most methods for estimating GLM are based on weighted least squares, but most existing algorithms do not allow modeling genes together with different dispersion parameters for distinct genes, unless the estimation is computed gene by gene. However, in this case, we are not able to do inference on the group effect and gene specific effect at the same time. Therefore, in order to

not only allow different dispersion parameters for distinct genes but also model group treatment effects, we have developed the following procedure.

Pre-Step: Initialization.

First, we estimate r_{j_g} for gene j_g and treat $\hat{r}_{j_g}$ as constant. There are many dispersion parameter estimation methods available, such as methods included in `edgeR` and `DESeq2`, and quasi-likelihood methods [91, 92]. In general, any of these methods could be applied. Then, we calculate $\hat{S}_k^0$ for $k = (i-1)p+1, \ldots, ip-1$, with $i = 1, \ldots, T$, $\hat{\alpha}_j^0$, $\hat{\beta}_{ij}^0$ by the reduced model in (4.5) below,

$$\mathbb{E}[Y_{j_g,ijk}] = S_k \exp\{\alpha_j + \beta_{ij} \mathbb{1}_{\{i \neq 1\}}\}. \tag{4.5}$$

Next, we estimate θ_{j_g} and η_{ij_g} by $\hat{\theta}_{j_g}^0$ and $\hat{\eta}_{ij_g}^0$ for each gene j_g, where $j_g = 1, \ldots, j_G - 1$ by substituting the estimation $\hat{S}_k^0$, $\hat{\alpha}_j^0$ and $\hat{\beta}_{ij}^0$ in the full model (4.1). Let $\hat{\theta}_{j_G}^0 = -\sum_{j_g=1}^{j_G-1} \hat{\theta}_{j_g}^0$ and $\hat{\eta}_{ij_G}^0 = -\sum_{ij_g=1}^{j_G-1} \hat{\eta}_{ij_g}^0$ by constraints (4.2) and (4.3).

Step 1.

We update $\hat{S}_k^1$, $k = (i-1)p+1, \ldots, ip-1$, for $i = 1, \ldots, T$, $\hat{\alpha}_j^1$, $\hat{\beta}_{ij}^1$ in the full model in (4.1) with θ_{j_g} and η_{ij_g} plugged in as $\hat{\theta}_{j_g}^0$ and $\hat{\eta}_{ij_g}^0$ in previous step by using $\hat{r}_{j_g} \hat{m}_{j_g,ijk}^0 / (\hat{r}_{j_g} + \hat{m}_{j_g,ijk}^0)$ as weights in the weighted least square GLM, where $m_{j_g,ijk}^0 = \hat{S}_k^0 \exp\{\hat{\alpha}_j^0 + \hat{\beta}_{ij}^0 \mathbb{1}_{\{i \neq 1\}} + \hat{\theta}_{j_g}^0 + \hat{\eta}_{ij_g}^0 \mathbb{1}_{\{i \neq 1\}}\}$. Details of the weights are given in the appendix.

Step 2.

We update $\hat{\theta}_{j_g}^1$ and $\hat{\eta}_{ij_g}^1$ for $j_g \in J_u$ where J_u is any subset of genes within group j (in practice we choose $|J_u| = 10$), by maximizing the likelihood (4.6) with S_k, α_j, β_{ij} replaced by $\hat{S}_k^1$, $\hat{\alpha}_j^1$ and $\hat{\beta}_{ij}^1$,

$$\ell_{J_u \cup j_G} = \sum_{j_g \in J_u \cup j_G} \sum_{i=1}^{T} \sum_{k=1}^{Tp} \left[C + (y_{j_g,ijk} + \hat{r}_{j_g}) \log(\hat{r}_{j_g} + m_{j_g,ijk}) + y_{j_g,ijk} \log(m_{j_g,ijk}) \right], \tag{4.6}$$

where C is a constant, using weighted least squares (detailed in Section C.0.1). The weight is $\hat{r}_{j_g}\hat{m}^{0*}_{j_g,ijk}/(\hat{r}_{j_g}+\hat{m}^{0*}_{j_g,ijk})$, where $m^{0*}_{j_g,ijk}=\hat{S}^1_k\exp\{\hat{\alpha}^1_j+\hat{\beta}^1_{ij}\mathbb{1}_{\{i\neq 1\}}+\hat{\theta}^0_{j_g}+\hat{\eta}^0_{ij_g}\mathbb{1}_{\{i\neq 1\}}\}$.

Step 3.

Steps 1 and 2 are repeated until the relative change of likelihoods is sufficiently small. In practice, we terminate the algorithm when the relative change of the likelihoods is smaller than 10^{-4}.

4.2.3 Testing Procedure with Bootstrap

Consider the parametrized model (4.1). To test the null hypothesis $H_0:\beta_{ij}+\eta_{ij_g}=0$ in (4.4) using model (4.1), we implement the widely used Wald test. In a given experiment, there are tens of thousands of sRNAs and tens of thousands of parameters η_{ij_g}. We found it difficult to derive the asymptotic variance-covariance matrix of the regression parameter estimates and we leave this as future work. In order to estimate the asymptotic covariance matrix, we bootstrap the Wald test statistic ($\hat{\beta}_{ij}$+ $\hat{\eta}_{ij_g}$). Our proposed procedure is to compute the bootstrap p-value corresponding to the observed value of a test statistic and control the FDR using the BH procedure ([27]). The bootstrap procedure is as follows:

(a) We first compute a vector of parameter estimates $\hat{\Theta}$, where $\Theta=(S_k,\alpha_j,\beta_{ij},\theta_{j_g},\eta_{ij_g})$, and test statistic $\hat{\tau}_{j_g}$, where $\tau_{j_g}=\beta_{ij}+\eta_{ij_g}$.

(b) Using model (4.1) based on the parameter vector $\hat{\Theta}$, we generate B bootstrap samples, each of size equal to that of the original data set. Then we use each bootstrap sample to compute $\hat{\tau}^b_{j_g}$ for $b=1,\ldots,B$ with the procedure shown in Section 4.2.2.

(c) Finally, we calculate the estimated variance for $\hat{\tau}_i$ using the bootstrap samples by finding the empirical variance of $\hat{\tau}^b_{j_g}$, $\hat{V}_{j_gB}=\text{var}(\hat{\tau}^b_{j_g})$. The Wald test statistic is estimated by $w_{j_g}=\hat{\tau}_{j_g}/\sqrt{\hat{V}_{j_gB}}$ and the approximate p-value under null hypothesis can be calculated using standard normal distribution.

We are also able to compare differences between any two treatment conditions by changing the τ_i from $\beta_{ij} + \eta_{ij_g}$ to $\beta_{ij} + \eta_{ij_g} - (\beta_{lj} + \eta_{lj_g})$, for $l \neq i$.

4.3 Monte Carlo Evidence

In this section, we examine the performance of the proposed method with other primary existing methods from two aspects, normalization and differential expression analysis, using simulation studies. We only focus on the pairwise comparison with $T = 2$.

4.3.1 Simulation Setting

We generate $1,000$ sRNAs with J distinct groups and the first and Jth group of gene have nonzero gene-specific effect and non-zero gene treatment effect, which is denoted by θ_{j_g} and η_{2j_g} respectively, where

$$\theta_{j_g} = -1.2\mathbb{1}(j_g = 1, \ldots, 60) + 1.8\mathbb{1}(j_g = 61, \ldots, 100),$$

and

$$\eta_{2j_g} = -2\mathbb{1}(1, \ldots, 20) + \mathbb{1}(j_g = 21, \ldots, 80) - \mathbb{1}(j_g = 81, \ldots, 100).$$

The replicate effects are set to be $\log(S_k) = u_k - \sum_k u_k/(2p)$, where p is the number of replicates (detailed in later section) and u_k are i.i.d unif$(0.5, 1.2)$. The dispersion parameters are simulated as r_{j_g} i.i.d. unif$(1, 3)$.

The experimental data in our case study (Section 4.4) has a known group structure. In practice, however, the sRNA classes may not be available, so that the group structure is unknown and may even be non-existent. Hence, we generate data with both known and unknown group structure to examine the performance of our method.

4.3.2 Known Group Structure

In this section, we examine the performance of the proposed method with other well-known methods when the group structure is known. The known group structure data means that the group information is available and we do not need to reconstruct the group index. In particular, α_j's are the group baseline and β_{2j}'s measure the group treatment effect, where β_{2j} equals to 0 (no group effect) or the group treatment differences $d_{\mathbf{gt}}$ and $d_{\mathbf{gt}} \in \{0, 0.5, 1, 2\}$. Because sRNAs always belong to a small number of distinct classes in practice, we use $J = 3$ and 6 as examples.

Setting A:

In total, we have $J = 6$ different groups and $p = 10$ replicates. Table 4.1 shows detailed information on the settings in each group.

Setting B:

In total, we have $J = 3$ different groups and $p = 6$ or 10 replicates. Table 4.1 shows detailed information on the settings in each group.

Table 4.1: Setting A and B: with $d_{\mathbf{gt}} = 0, 0.5, 1$ or 2.

	Setting A			Setting B		
j	j_G	α_j	β_{2j}	j_G	α_j	β_{2j}
1	500	3	0	800	3	0
2	100	4	0	100	4	$d_{\mathbf{gt}}$
3	100	5	0	100	7	$d_{\mathbf{gt}}$
4	100	6	0			
5	100	7	0 or $d_{\mathbf{gt}}$			
6	100	8	$d_{\mathbf{gt}}$			

The *true test* with the true S_k and r_{j_g} is also conducted for comparison. The empirical FDRs and powers averaged over 100 replications are used for comparison.

Normalization Methods Comparison.

For comparison of known group structure data, we apply the normalization methods shown below and two types of testing procedures for each simulated dataset. In order to derive gene expression measures for subsequent DE analysis, we first need to normalize read counts to adjust for varying replicate sequencing and other technical effects. We compare our Group Based Normalization method, in which each S_k is estimated under model (4.1) using the method in Section 4.2.2, with three other methods from the literature:

- Upper Quartile (UQ): Gene counts are divided by the upper quartile of counts different from 0 associated with their replicates and multiplied by the mean upper quartile across all samples.

- Relative Log Expression (RLE): This is the default normalization method that is included in `DESeq2` package ([31]). It computes

$$\hat{S}_k = \text{median} \frac{Y_{j_g,ijk}}{\left\{\prod_{k=1}^{2p} Y_{j_g,ijk}\right\}^{1/(2p)}},$$

  where p is number of replicates. Each size factor is calculated as the median of the ratio of k-th sample's counts to those of the pseudo-reference.

- Trimmed Mean of M-values (TMM): This normalization method ([10]) is implemented in `edgeR` package. Each size factor is calculated based on a reference sample of pre-normalized count by library size. After calculating trimmed means of the relative sizes of transcriptomes, the relative scaling factors are adjusted to multiply to 1.

Both TMM and RLE assume that most genes are not DE genes. TMM computes the scaling factor based on the trimmed mean between each sample with the reference. RLE takes all samples into account. UQ assumes the similarity of read counts' upper-quartile across samples. Besides these three normalization techniques, other approaches focus on using the housekeeping genes or correcting biases related to GC-content (or guanine-cytosine content) [9, 93, 94]. However, in

practice, we do not have the housekeeping or GC-content in our sRNA experiment, so we do not include these for comparison.

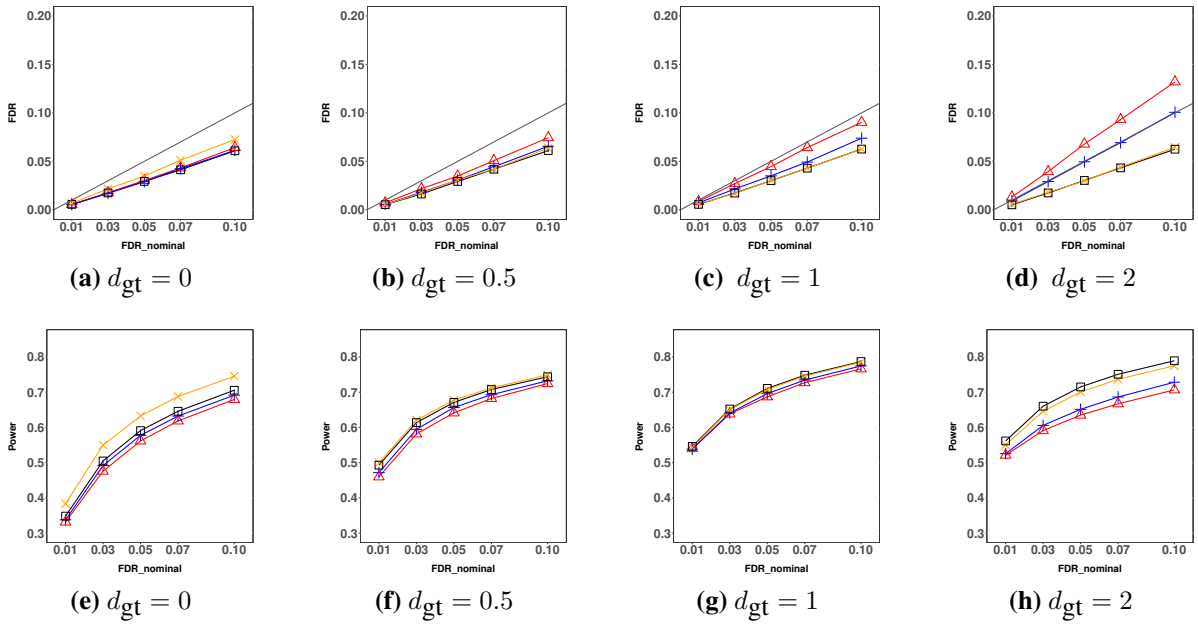
In order to compare the performance of the aforementioned normalization methods, we consider the null hypothesis in (4.4). We compare the performance of our method to others using `edgeR` as the testing method to identify the DE genes. Figures 4.1–4.4 display results for the proposed comparison. Each point on the figures displays the empirical FDR (subfigures (a)–(d)) or power (subfigures (e)–(h)) of the corresponding method at a given nominal FDR level, which is shown on the x-axis.

Overall, all methods have their empirical FDRs controlled with small d_{gt}, $d_{\text{gt}} = 0$ and 0.5. Compared with our method, TMM, RLE and UQ are less powerful. For Setting A with $\beta_5 = 0$, all methods except RLE have the nominal level FDRs controlled, as shown in Figure 4.1 subfigures (a)–(d). When $\beta_5 \neq 0$, i.e. the proportion of genes that have non zero group effect increases, only UQ and the proposed method control the FDRs. For Setting B with $J = 3$, when d_{gt} increases, as shown in Figure 4.3, TMM, RLE and UQ fail to control the FDR at all levels. Even with the number of replicates p increases from 6 to 10, as shown in Figure 4.4 subfigures (c)–(d), the traditional methods fail to control FDR. In addition, as the group treatment effect d_{gt} increases, the proposed normalization method is more superior than existing methods in terms of power.

Differential Expression Analysis.

Results for the DE analysis focus on the hypothesis in (4.4) of the proposed test procedure under settings A and B are displayed in Figures 4.5–4.8. In this section, we compare the following four methods: a) the proposed testing procedure with r_{j_g} estimated using `edgeR`, b) `edgeR` with the proposed normalization in conjunction with `edgeR` (which has the best performance in Section 4.3.2), c) the proposed testing procedure with true r_{j_g} and d) the *true test* which is the proposed testing procedure with true S_k and r_{j_g}.

Under Setting A, Figures 4.5 and 4.6 show that the proposed method with true r_{j_g} performs the best: it has the FDRs controlled and it is the most powerful among all methods. The empirical FDRs of the proposed method with estimated r_{j_g} are slightly inflated due to the small proportion



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

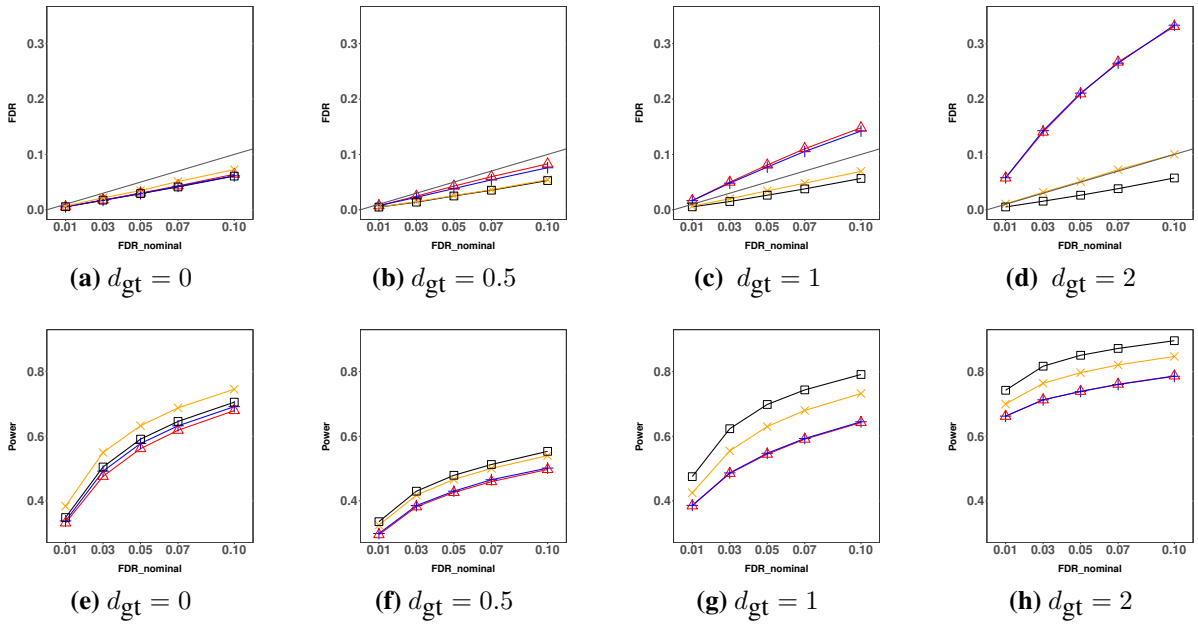
(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

Figure 4.1: Empirical FDRs and powers for testing the overall DE genes by `edgeR` method with simulation setting A, with $\beta_5 = 0$, that is the proportion of group DE is 10%, using the proposed normalization method ($\square$), compared to those of the TMM ($+$), the RLE (Δ) and UQ ($\times$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

of DE genes. When the proportion of DE genes increases (in Figure 4.6) and when the group treatment effect increases, our proposed method with estimated r_{j_g} performs better. The empirical powers of the proposed method with estimated r_{j_g} are very close to the empirical powers of the proposed method with true r_{j_g}, especially with large d_{gt}, and they are much higher than the empirical powers of `edgeR` with our proposed normalization. When the number of groups J decreases from 6 to 3 under Setting B, the proposed method with true r_{j_g} still performs the best, and the FDRs of the proposed method using estimated r_{j_g} are better with increasing d_{gt}; see Figures 4.7 and 4.8. Also, as the number of replicates p increases, the deviations between the proposed method with estimated r_{j_g} and the proposed method with true r_{j_g} quickly vanish.

4.3.3 Unknown Group Structure

In practice, we may not have biological information on group structures. In this section, we examine the performance of the proposed method by comparing the following tests: a) the proposed



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

Figure 4.2: Empirical FDRs and powers for testing the overall DE genes by `edgeR` method with simulation setting A, with $\beta_5 = d_{\text{gt}}$, that is the proportion of group DE is 20%, using the proposed normalization method ($\square$), compared to those of the TMM ($+$), the RLE (Δ) and UQ ($\times$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

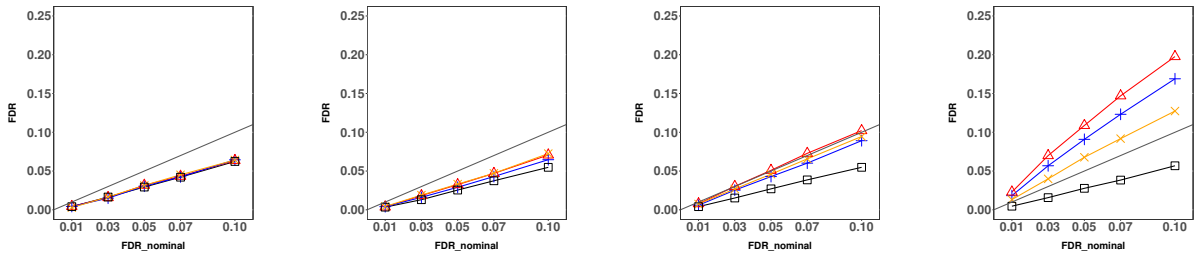
test with r_{j_g} estimated using `edgeR`, b) `edgeR` with the proposed normalization in conjunction with `edgeR`, and c) the proposed testing procedure with true r_{j_g}.

We use Setting C, displayed below, to perform simulations.

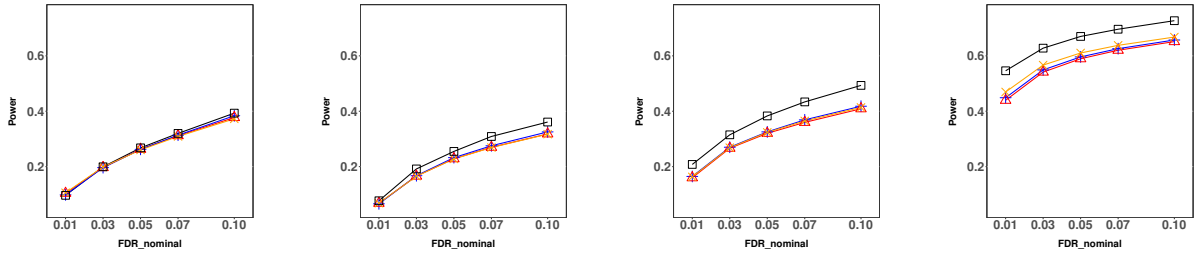
Setting C:

We follow the general simulation setting on genes included in Section 4.3.1. In particular, α_j's are the group baseline, where α_j is a linear function of the group base differences d_{gb}, and $d_{\text{gb}} \in \{0, 0.5, 1, 1.5, 2\}$. β_{2j}'s measure the group treatment effects and equal to 0 or the group treatment differences d_{gt}, where $d_{\text{gt}} \in \{0.2, 1\}$. Table 4.2 includes detailed group-wise settings. In total, we have $J = 3$ different groups, and $1,000$ sRNAs are simulated.

In order to apply the proposed method, we use k-means on the control group to determine the group structure, and use the estimated group index for further analysis. As group based difference d_{gb} measures the relative mean differences between different groups, the smaller the d_{gb} is, the harder it is to accurately recover the group index.



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$



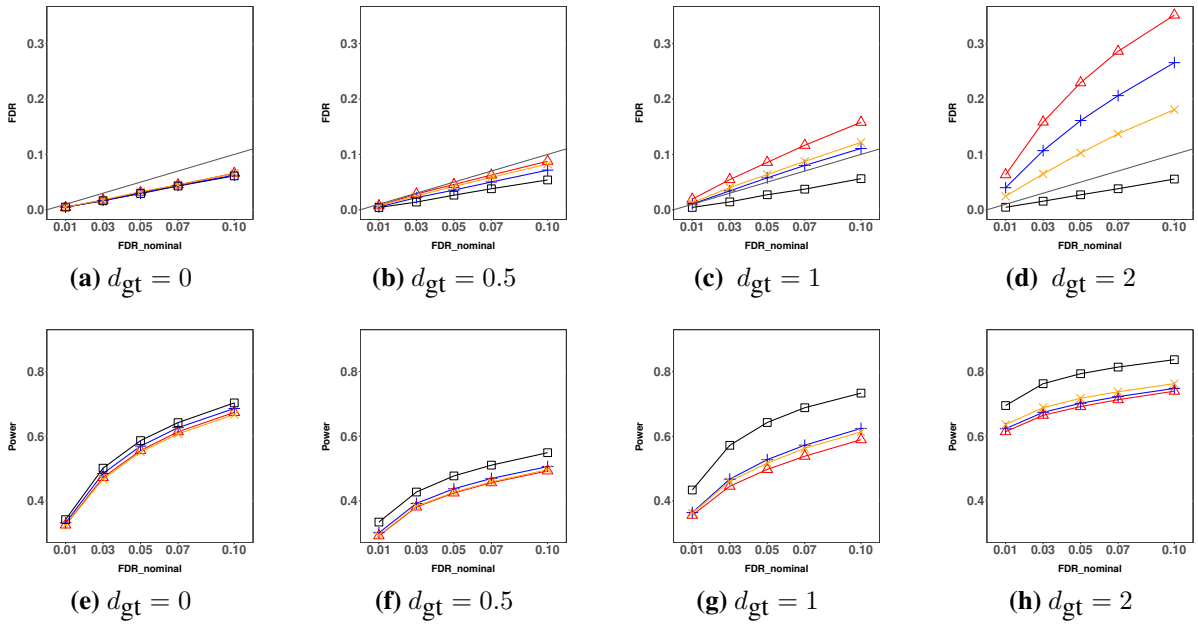
(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

Figure 4.3: Empirical FDRs and powers for testing the overall DE genes by `edgeR` method with simulation setting B, with $p = 6$, using the proposed normalization method (□), compared to those of the TMM (+), the RLE (△) and UQ (×). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

Table 4.2: Setting C: with $d_{\text{gb}} = 0, 0.5, 1, 1.5$ or 2, $d_{\text{gt}} = 0.2$ or 1.

j	j_G	α_j	β_j
1	800	3	0
2	100	$3 + d_{\text{gb}}$	d_{gt}
3	100	$3 + 4\, d_{\text{gb}}$	d_{gt}

Figure 4.9 displays the empirical FDRs (subfigures in the first column) and powers (subfigures in the second column) of the corresponding methods at a given d_{gb} when the nominal level of FDR is 0.1. For Setting C, with small d_{gt}, all methods have the nominal FDRs controlled, as shown in Figure 4.9 subfigures (a) and (e). Compared to our proposed method, `edgeR` is less powerful. The proposed test with true r_{j_g} performs the best overall. The empirical powers of the proposed test with $\hat{r}_{j_g}$ is almost indistinguishable from those of the proposed test with true r_{j_g} when the number of replicates is 10. The empirical FDRs of the proposed method become less stable with increasing d_{gb}, but are still under control. Though unknown group structures may be hard to recover with low d_{gb}, our proposed method still provides satisfactory results, which shows



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

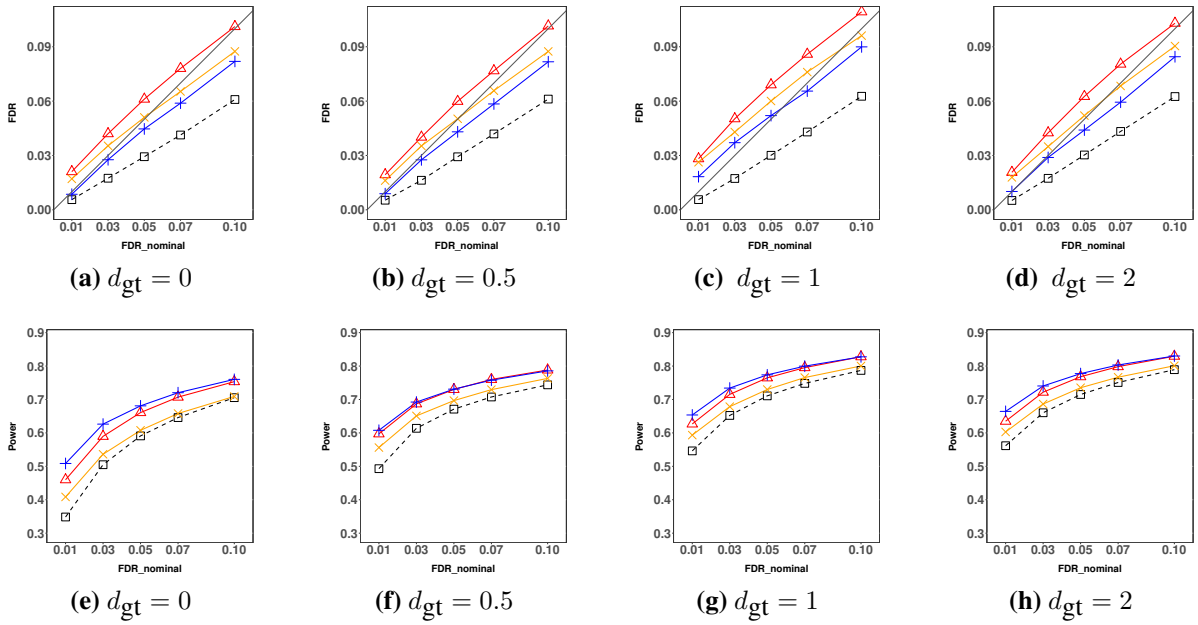
Figure 4.4: Empirical FDRs and powers for testing the overall DE genes by `edgeR` method with simulation setting B, with $p = 10$, using the proposed normalization method ($\square$), compared to those of the TMM ($+$), the RLE (Δ) and UQ ($\times$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

the flexibility and robustness of the method. In addition, as the number of replicates p increases, the deviations between the proposed method and `edgeR` become more substantial.

These numerical results suggest the superiority of the proposed method for analyzing the group-structured data when the group effect is significant, as well as its flexibility in detecting DE genes with unknown group-structured data.

4.4 Application to C. elegans Data

In this section, we apply our method to the sRNAs data from the experiment mentioned in Section 4.1. There are four groups of sRNAs: piRNA, miRNA, CSR-1 class 22G-RNAs and WAGO-class 22G-RNA locus. There are $T = 3$ treatment conditions: wild type (wt), mut16 and prg-1 mutants. Hypotheses defined in (4.4) are tested. Based on the results from the differential expression analysis, we identify the up and down-regulated sRNAs using the estimated gene-specific differences $\hat{\beta}_{ij} + \hat{\eta}_{ij_g}$ for $i = 1, 2$.



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

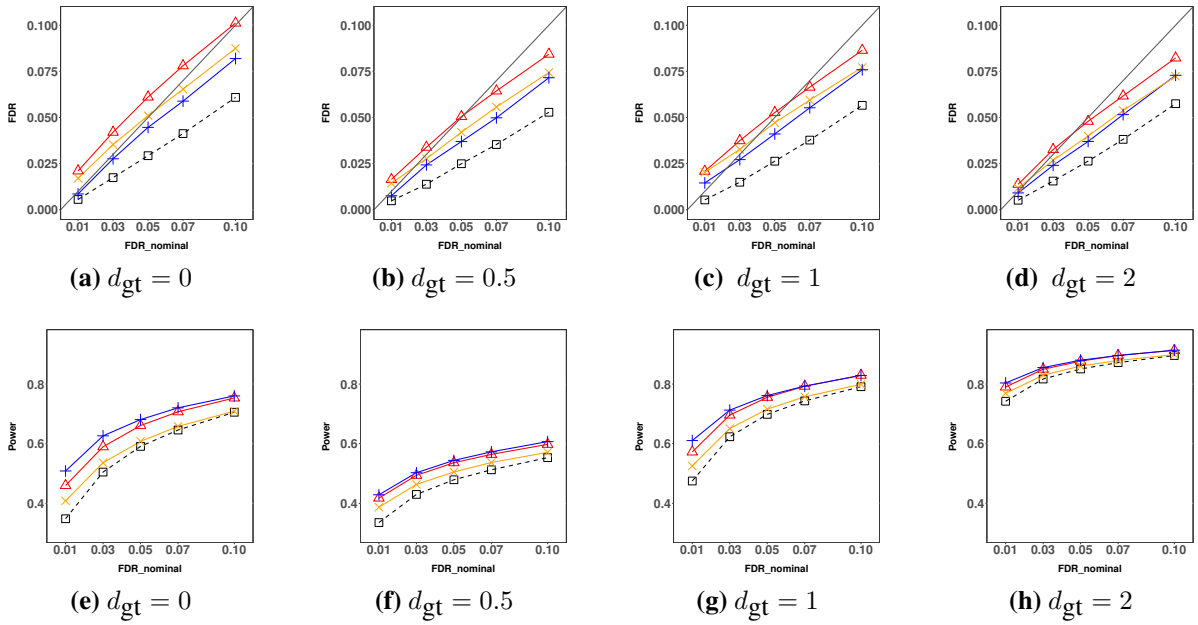
Figure 4.5: Empirical FDRs and powers for testing the overall DE genes under simulation setting A, with $\beta_5 = 0$ by `edgeR` method using the proposed normalization method ($--\square--$), compared to the proposed test procedure with the proposed normalization method ($—\Delta—$), the proposed test procedure with the True r_{jg} ($—+—$), the *true test* with the true S_k and true r_{jg} ($—\times—$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

4.4.1 Analysis and Results

As suggested by many studies [95, 96], we first filter out genes with all zero expression within each treatment trial. If a gene does not display any expression, there is very limited information from a biological point of view. For this analysis, we focus on those genes with some expression. After filtering, $7,146$ genes remain for further analysis. We follow the estimation and testing steps in Sections 4.2.2 and 4.2.3 to perform the DE analysis.

For DE analysis, the nominal FDR level is set at 0.05 and we compare the results using the following seven methods: a) the proposed test with r_{jg} estimated using `edgeR`, b) `edgeR` with TMM, c) `edgeR` with RLE, d) `edgeR` with UQ, e) `edgeR` with our normalization, f) `DESeq2` with RLE and g) `DESeq2` with our normalization.

We focus on the analysis of two pairwise comparisons, wt-versus-mut16 and wt-versus-prg-1. We identified $3,928$ DE sRNAs under the wt and mut16 comparison, and $5,676$ DE sRNAs under wt and prg-1 comparison. The majority of the DE genes we identified were also identified by other



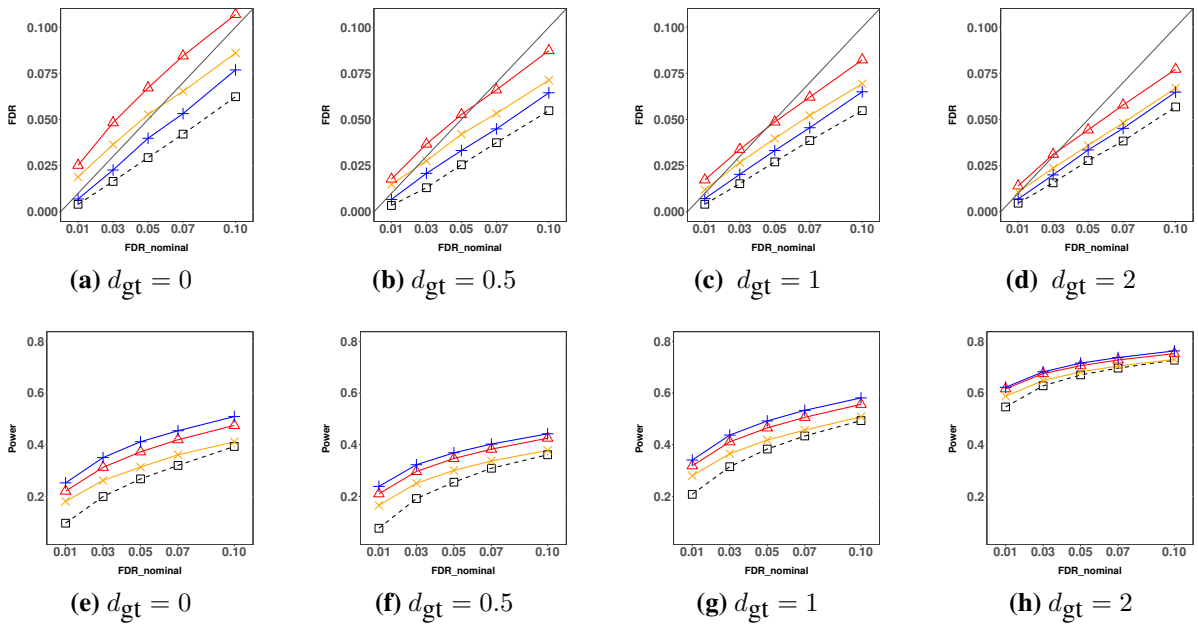
(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

Figure 4.6: Empirical FDRs and powers for testing the overall DE genes under simulation Setting A, with $\beta_5 = d_{\text{gt}}$, by `edgeR` method using the proposed normalization method (−−□−), compared to the proposed test procedure with the proposed normalization method (—Δ–), the proposed test procedure with the True r_{jg} (—+–), the *true test* with the true S_k and true r_{jg} (—×—). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

methods. As we discussed in Section 4.3, non-group based normalization could not control the FDR if the group effect is significant. `DESeq2` combine RLE and `edgeR` with other normalization methods identify more DE genes may be due to the technical artifacts or unintended variation that failed to be removed. Figure 4.10 displays the comparison of our proposal with `DESeq2` using RLE (DESeq2) and `DESeq2` using our proposed normalization (DESeq2_our). More comparisons of the proposed method with `edgeR` are displayed in Figure S.1.

It is known that piRNAs are expected to be down-regulated in prg-1 mutants. Our estimated group effects of piRNA in the wt-versus-prg-1 experiment support this down-regulation. Among the DE sRNAs we find in the wt-versus-prg-1 experiment, four piRNAs are up-regulated, as shown in Table S.1. Those piRNAs may be misannotated. Also, miRNAs are expected to be up-regulated in both prg-1 and mut16 mutants. Our estimated group effects of piRNA in both wt-versus-prg-1 and wt-versus-mut16 experiments also support this up-regulation. Among the DE sRNAs we detected, four miRNAs are down-regulated in the wt-versus-mut16 experiment and 12 in the wt-



(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 0.5$ **(c)** $d_{\text{gt}} = 1$ **(d)** $d_{\text{gt}} = 2$

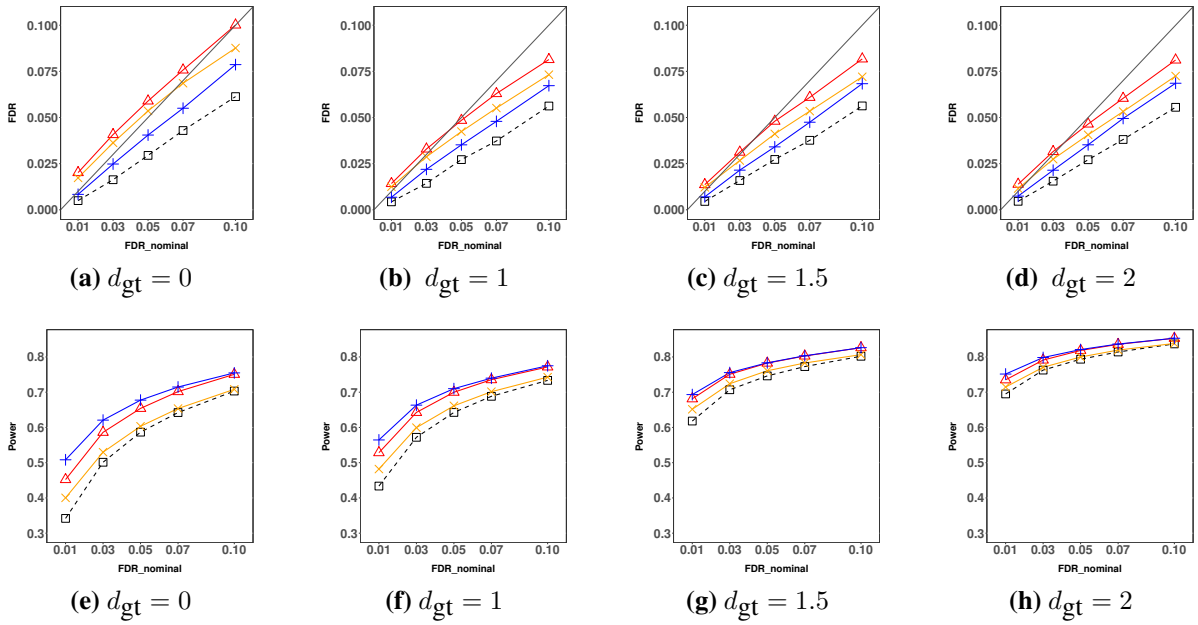
(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 0.5$ **(g)** $d_{\text{gt}} = 1$ **(h)** $d_{\text{gt}} = 2$

Figure 4.7: Empirical FDRs and powers for testing the overall DE genes under simulation Setting B, with $p = 6$, by `edgeR` method using the proposed normalization method ($--\square--$), compared to the proposed test procedure with the proposed normalization method ($—\Delta—$), the proposed test procedure with the True r_{jg} ($—+—$), the *true test* with the true S_k and true r_{jg} ($—\times—$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

versus-prg-1 experiment, which means those miRNAs may be misannotated. Further biological studies will be conducted. As our proposal is able to estimate the group effects of different types of the sRNAs, the results in the DE analysis are more sophisticated than simple individual comparisons. The new test reveals pathways for more detailed analysis in general.

4.5 Discussion

We proposed an inferential procedure for analyzing multiple treatment levels in sRNA data. Our approach includes both group and gene specific treatment effects in the negative binomial model. The proposed method, while maintaining control of the FDR, leads to higher power than the traditional pairwise comparison methods whenever the group effect is significant. The proposed method performs similarly to existing methods from the literature in the absence of group effects. The results in this paper cover both the pairwise and multiple treatments settings.



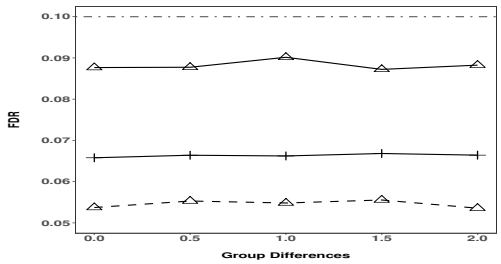
(a) $d_{\text{gt}} = 0$ **(b)** $d_{\text{gt}} = 1$ **(c)** $d_{\text{gt}} = 1.5$ **(d)** $d_{\text{gt}} = 2$

(e) $d_{\text{gt}} = 0$ **(f)** $d_{\text{gt}} = 1$ **(g)** $d_{\text{gt}} = 1.5$ **(h)** $d_{\text{gt}} = 2$

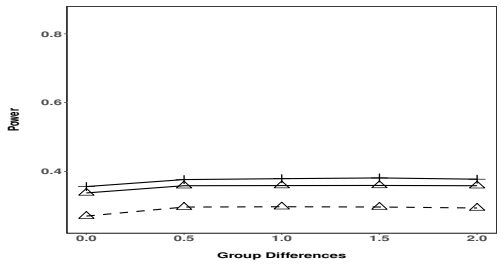
Figure 4.8: Empirical FDRs and powers for testing the overall DE genes under simulation setting B, with $p = 10$, by `edgeR` method using the proposed normalization method ($--\square--$), compared to the proposed test procedure with the proposed normalization method ($—\Delta–$), the proposed test procedure with the True r_{jg} ($—+–$), the *true test* with the true S_k and true r_{jg} ($—\times—$). Each point on figures displays the empirical FDR or power of the corresponding method at a given nominal FDR level.

This approach allows not only traditional differential expression analysis, but also inference for each sRNA group. By contrast, existing pipelines such as `edgeR` and `DESeq2` are not able to make inference for the group effect. To analyze the group effect, one can use the same procedure in Section 4.2.3 by letting $\tau_i = \beta_{ij}$ or $\tau_i = \beta_{ij} - \beta_{lj}$, for $i \neq l$ and use Bonferroni correction to control the multiple testing error. We leave this for future investigation.

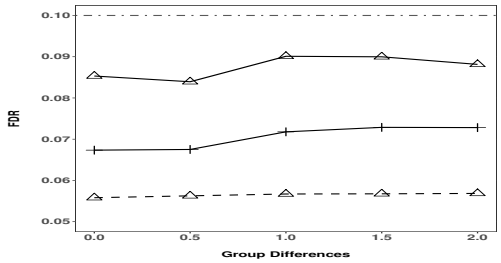
Our method incurs some computational cost due to bootstrapping of the test statistics, which can be partly offset in practice by parallel implementation. A possible alternative to the bootstrap would be to derive the Fisher information matrix in this setting and use it to obtain the asymptotic variance-covariance matrix of the parameter estimates. However, the very large number of parameters leads to more involved questions that we leave for future study.



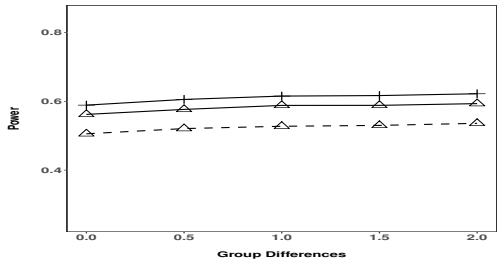
(a) $p = 6$, $d_{\text{gt}} = 0.2$



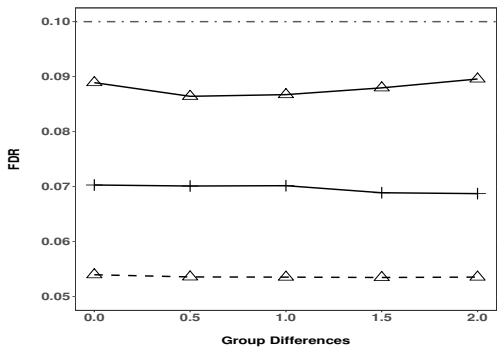
(b) $p = 6$, $d_{\text{gt}} = 0.2$



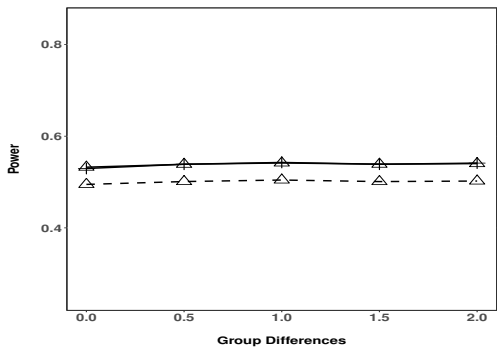
(c) $p = 6$, $d_{\text{gt}} = 1$



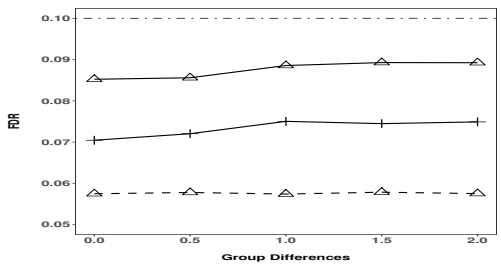
(d) $p = 6$, $d_{\text{gt}} = 1$



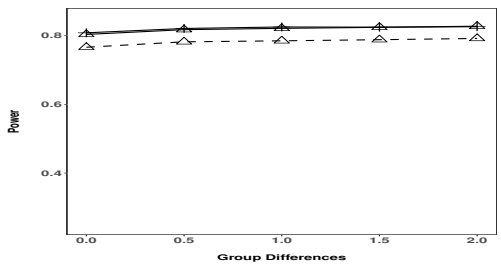
(e) $p = 10$, $d_{\text{gt}} = 0.2$



(f) $p = 10$, $d_{\text{gt}} = 0.2$

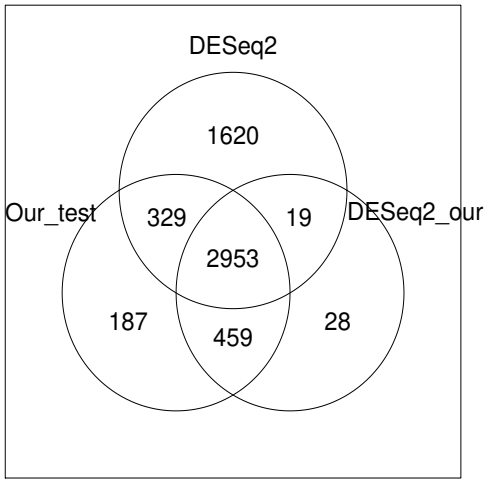


(g) $p = 10$, $d_{\text{gt}} = 1$

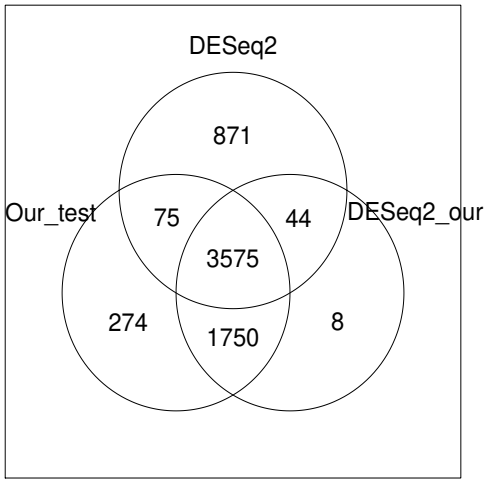


(h) $p = 10$, $d_{\text{gt}} = 1$

Figure 4.9: Empirical FDRs and powers for testing the overall DE genes by the proposed method with estimated r_{j_g} (—Δ—), with true r_{j_g} (— + —) compared with `edgeR` using the proposed normalization method (− − Δ−) with simulation setting C. Each point on figures displays the empirical FDR or power of the corresponding method at different group differences d_{gb} based on given nominal FDR level $\alpha = 0.1$.



(a) wt-vs-mut16



(b) wt-vs-prg-1

Figure 4.10: differential expression analysis results comparing the proposed method with `DESeq2`.

Chapter 5

Understanding the Drivers of Sensitive Behavior Using Poisson Regression from Quantitative Randomized Response Technique Data

5.1 Introduction

5.1.1 Background and Related Work

Sensitive behaviors are those that are non-compliant with rules or regulations or are socially unacceptable. Sensitive behaviors are relevant to a variety of fields, including health sciences (e.g., abortion, illicit drug use, sexual activity), natural resource management (e.g., poaching of flora and fauna), business (e.g., tax evasion, insider trading), and education (e.g., cheating on exams). Although widespread, such behaviors are typically challenging to research, but understanding the behavior is paramount to creating effective interventions for the benefit of society at large. Successful interventions often require knowledge of who is engaged in the sensitive behavior, what the individuals are doing, where the sensitive behaviors take place, and why the individuals are engaged in the sensitive behaviors [97]. However, methodological constraints hamper collection of accurate data on such behaviors because participants are unlikely to disclose sensitive behaviors for fear of retribution or due to social undesirability.

Indirect survey methods allow researchers to gather information on sensitive behavior without the threat of implicating respondents [97]. Indirect methods for studying sensitive behavior include the randomized response technique (RRT; [98]), which provides anonymity to interviewees who answer sensitive questions. The original RRT has been modified by researchers (e.g., [99–101]) and applied in many contexts to help understand sensitive behaviors. See Fox and Tracy [102] or Chaudhuri and Mukerjee [103] for overviews of such methods and [104] for validation via a meta-

analysis of randomized response studies. Use of RRT in surveys has been shown to increase a respondent's proclivity to respond to questions about the sensitive behavior, as well as to increase the likelihood that a respondent provides accurate responses [104–107]. This method has shed light on sensitive behaviors in the fields of health sciences, natural resource management, business, education and political sciences [104]

The standard RRT approach uses a randomizing device, such as a coin or die, to determine the question a respondent answers. One or more questions are innocuous while another focuses on the sensitive behavior. The interviewer has no way of knowing which question the respondent is answering, thereby ensuring anonymity and increasing response rates and accuracy of responses provided. In this paper, we focus on nonnegative count data obtained via a modification of the technique referred to as the quantitative randomized response technique (QRRT) [101], which allows researchers to understand prevalence of a sensitive behavior in a community or society (e.g., [106]), as well as estimates of the frequency or quantity of the sensitive behavior (e.g., [100, 101]).

A major gap with the use of RRT has been in answering questions concerning drivers of non-compliance—the "why" question [97]. This is an essential question to investigators as it is typically critical when designing effective interventions to address non-compliance. Statistically, this corresponds to building and testing regression models for randomized response data. Logistic regression models for binary randomized response data are treated in [108] by recognizing the structure as a generalized linear model with a particular link function. Regression models are also developed in [108] for multi-category randomized response data, when the vector-valued observation comes from multiple randomized response questions. Another approach to inference with multiple sensitive questions is to sum the randomized responses; [109] and [110] develop regression models for such sum scores, including one based on zero-inflated Poisson regression. Some `R` packages [111, 112] have been developed to support this regression methodology.

However, to the best of our knowledge, regression methodology has not been developed for count data from QRRT [101]. We develop a methodology for Poisson regression with QRRT data, based on maximum likelihood implemented via the EM algorithm [113]. We implement

the methodology in a freely-available `R` package, by adapting existing software for generalized linear models. Further, we provide an asymptotic theory to support estimation and testing of models. In particular, we derive the Fisher information matrix in this setting and use it to obtain the asymptotic variance-covariance matrix of the regression parameter estimates. Simulation results illustrate the quality of the asymptotic approximations. Using a case study of noncompliance with natural resource regulations [114], we demonstrate our new statistical approach to examine drivers of sensitive behavior.

5.1.2 Case Study: Non-compliance with Hunting Regulations in Sierra Leone

To demonstrate the utility of this new analytical approach, we examine the relative effects of different hypothesized drivers of non-compliant resource use activities inside the Western Area Peninsula Forest Reserve (WAPFR) in Sierra Leone. WAPFR comprises 175 km^2 located between the Atlantic Ocean to the west and south, the capital city of Freetown 5km to the north, and a low-lying plain to the east. WAPFR is an important site for conservation in Sierra Leone because of the biodiversity it protects, including numerous endemic and highly threatened species, and also due to ecosystem services the reserve provides to 50 surrounding communities, including the main water source for Freetown's 1.5 million residents. Communities neighboring WAPFR are home to all 17 of the country's ethnic groups, which rely on gardening, small-scale businesses, sand extraction, fishing, and hunting for subsistence. Resource extraction is strictly prohibited inside WAPFR, but illegal hunting is a major threat [114–116].

The case study reported here was part of a larger examination of non-compliance in WAPFR (see [114–116]). We randomly selected 842 households (sampling every other household on a street) in eight communities that had similar numbers of households (100–500 households each). Coauthor Abu Conteh, a citizen of Sierra Leone, carried out the field research. Conteh surveyed heads of households in Krio (the lingua franca of Sierra Leone). Survey questionnaires can be found in Conteh, 2010 [117] (Appendices IV a & b) and are reproduced in [117] for convenient reference. Ninety-eight percent of households answered all questions posed. Before beginning

work in each community, Conteh obtained permission from community leaders, and all respondents gave verbal consent (written consent was not used due to illiteracy rates in some of the communities sampled) to participate in the research. We did not record any information that could be used to identify individual respondents. Ethics approval to conduct the research was obtained from Victoria University of Wellington (Approval No. 15521).

We used the QRRT ([101, 118]) to estimate quantities of illegal hunting (see [114] for additional details). We recorded information on hunting activities over a nine-month period anchored by two widely known dates (New Years Day (January 1st) and Eid Ul Adha (October 1st) to reduce recall bias.

We designed and constructed a sealed, transparent, round bottomed container to serve as the randomizing device for QRRT. The container had a narrow neck that could only house one ball at a time. We placed 25 orange and 25 green balls into the container. Green balls had numbers from a known distribution painted on them [118]. Each respondent first turned their back on the interviewer and shook the container. If green fell into the neck of the container, the respondent read the number off the ball. If an orange ball fell into the neck, the respondent provided a numerical answer to the sensitive question the interviewer had posed prior to initiation of the exercise. The interviewer had no way of knowing whether the number stated by the respondent was innocuous (i.e. the number from a green ball) or was referring to the sensitive question (i.e. how many times per month on average did someone from the household hunt inside the reserve with the use of traps during the nine-month study period?). By ensuring anonymity in this way, QRRT encourages more truthful answers to questions regarding sensitive behavior [118]. However, as we outline below, because the researcher knows the probability of a respondent choosing a green or orange ball, as well as the distribution of numbers written on the green balls, estimates can be made of the quantities of sensitive behavior being conducted by different sectors of the populations.

Compliance with natural resource use regulations may be driven by a wide variety of potential factors [119–123]. To demonstrate the new analytical approach for the analysis of QRRT data, we compare the relative support for different hypothesized drivers of non-compliance with con-

servation regulations in Sierra Leone using an information theoretic approach. Specifically, we construct latent Poisson regression models that describe the effects on the amount of illegal trapping in WAPFR of perceived enforcement of the regulations, perceived resource rarity, access to alternative livelihoods, and other factors. We then fit and test these models using our new QRRT regression methodology.

5.2 Methods

5.2.1 Probability Model

Let T_i denote the true count of the sensitive behavior, let $z_i > 0$ denote a known offset, and let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a p-vector of known covariates for the ith individual, $i = 1, \ldots, n$. Assume that

$$\begin{aligned} T_i &\sim \text{independent Poisson } (z_i \lambda_i) \\ \ln(\lambda_i) &= \boldsymbol{x}_i' \boldsymbol{\beta} \end{aligned}$$

where

$$\mathrm{P}\left[T_i = t \mid \lambda_i\right] = \frac{e^{-z_i \lambda_i} (z_i \lambda_i)^t}{t!} = \pi_i(t \mid \boldsymbol{\beta}) \tag{5.1}$$

for $t = 0, 1, 2, \ldots$, and $\boldsymbol{\beta}$ is a p-vector of unknown parameters.

The $\{T_i\}_{i=1}^n$ are not observed directly, but are masked through QRRT [101] as described in the Sierra Leone example. Let m be a known positive integer and let $b(r)$ denote a completely known probability mass function on the integers $0, 1, \ldots, m, m+1$. Let N denote the total number of balls and assume that $Nb(0), Nb(1), \ldots, Nb(m+1)$ are all integers. Then $Nb(0)$ balls are marked 0, $Nb(1)$ balls are marked 1, …, $Nb(m)$ balls are marked m, and $Nb(m+1)$ balls are blank. The ith interviewee selects a random integer $B_i \sim b(r)$ by selecting one of the balls. If $B_i \leq m$, the ball is numbered and the interviewee's response is the ball number, $R_i = B_i$. If $B_i = m+1$, the ball is

blank and the interviewee's response is the true count, $R_i = T_i$. Since no one but the interviewee knows the value of B_i, only the interviewee knows whether the response is a true value T_i or a randomized response B_i, assuming $T_i \leq m$. This requires some care in the choice of m, to ensure it is sufficiently large: any reported values larger than m are known to be true counts. The higher the ratio of blank balls to marked balls, $b(m+1)/\sum_{r=0}^{m} b(r)$, the higher the expected number of true responses and the more powerful the inference, but the lower the guarantee of anonymity. The lower the ratio of blank to marked, the lower the expected number of true responses, but the higher the guarantee of anonymity; see [101]. While the choice of the distribution $b(r)$ is up to the researcher, it would be very difficult to optimize this choice without detailed information about the unknown distribution of true responses.

5.2.2 Poisson Regression via EM algorithm

If the $\{T_i\}_{i=1}^n$ were observed directly, inference could proceed via Poisson regression fitted by maximum likelihood. Since only the realized values $\{r_i\}_{i=1}^n$ of the random variables $\{R_i\}_{i=1}^n$ are observed, we use the Expectation-Maximization (EM) algorithm [113] to maximize the likelihood, by first augmenting with the unobserved values $\{B_i\}_{i=1}^n$.

If the $\{B_i\}_{i=1}^n$ values were known, we would discard all but the true data values, for which $\mathbf{1}_{\{B_i=m+1\}} = 1$, resulting in the complete-data log-likelihood

$$\sum_{i=1}^{n} \mathbf{1}_{\{B_i=m+1\}} \left\{ -\ln(r_i!) - z_i\lambda_i + r_i \ln z_i + r_i \ln \lambda_i \right\}. \tag{5.2}$$

The incomplete-data log-likelihood is the conditional expectation of (5.2) given the observed data and the current estimate of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}^{(k)}$:

$$\sum_{i=1}^{n} \mathrm{P}\left[B_i = m+1 \mid \{R_i\} = \{r_i\}, \boldsymbol{\beta}^{(k)} \right] \left\{ -\ln(r_i!) - z_i\lambda_i + r_i \ln z_i + r_i \ln \lambda_i \right\}$$

$$= \sum_{i=1}^{n} \omega_i^{(k)} \left\{ -\ln(r_i!) - z_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + r_i \ln z_i + r_i \boldsymbol{x}_i'\boldsymbol{\beta} \right\}, \tag{5.3}$$

where the conditional probabilities $\{\omega_i^{(k)}\}_{i=1}^n$ are computed via Bayes' rule as

$$\omega_i^{(k)} = \frac{\pi_i\left(r_i \mid \boldsymbol{\beta}^{(k)}\right) b(m+1)}{b(r_i)\mathbf{1}_{\{r_i<m+1\}} + \pi_i\left(r_i \mid \boldsymbol{\beta}^{(k)}\right) b(m+1)}. \tag{5.4}$$

The EM algorithm then reduces to iterating the following steps across k to maximize the likelihood and obtain the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}$:

- **E-step:** compute weights from (5.4) under the current maximized model with parameters $\boldsymbol{\beta}^{(k)}$.
- **M-step:** maximize the weighted log-likelihood (5.3) for Poisson regression.

5.2.3 Asymptotic Distribution and Variance Estimation

In derivations not described here, we have verified the regularity conditions in chapter 2 of Fahrmeir and Tutz [124], establishing that the MLE is asymptotically normally distributed as $n \to \infty$. Thus, in large samples,

$$\widehat{\boldsymbol{\beta}} \text{ is approximately } \mathcal{N}\left(\boldsymbol{\beta}, \mathcal{I}^{-1}\left(\boldsymbol{\beta}\right)\right),$$

where $\boldsymbol{\beta}$ is the vector of true regression coefficients and $\mathcal{I}^{-1}\left(\boldsymbol{\beta}\right)$ is the inverse of the Fisher information matrix. We derive the Fisher information matrix in the supplemental material, Section D.1. The asymptotic covariance matrix $\mathcal{I}^{-1}\left(\boldsymbol{\beta}\right)$ is then estimated by plugging in the MLE, $\widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \mathcal{I}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)$.

5.2.4 Hypothesis Testing and Model Selection

The log-likelihood $\ell\left(\boldsymbol{\beta}; \{r_i\}_{i=1}^n\right)$ derived in the supplemental material, Section D.1, can be used in hypothesis testing and model selection. First, let $\boldsymbol{\beta}_{\text{full}}$ be a vector of p parameters for a full model that fits the data well. Let $\boldsymbol{\beta}_{\text{reduced}}$ be a vector of q parameters for a nested (reduced) model within the full model (that is, a model obtained by setting $p - q$ of the parameters in $\boldsymbol{\beta}_{\text{full}}$

equal to zero). To test the null hypothesis that the reduced model fits the data equally as well as the full model, we compute the likelihood ratio test statistic

$$W = -2\ell\left(\widehat{\boldsymbol{\beta}}_{\text{reduced}}; \{r_i\}_{i=1}^n\right) + 2\ell\left(\widehat{\boldsymbol{\beta}}_{\text{full}}; \{r_i\}_{i=1}^n\right), \tag{5.5}$$

where $\widehat{\boldsymbol{\beta}}_{\text{full}}$ and $\widehat{\boldsymbol{\beta}}_{\text{reduced}}$ are the MLE's for the full and reduced models, respectively. Standard asymptotic theory shows that for n large, W has an approximate χ^2_{p-q} distribution, the chi-squared distribution with $p-q$ degrees of freedom. We reject the reduced model in favor of the full model if the test statistic is large (e.g., [125]).

The maximized log-likelihood can also be used to compare models that need not be nested, via Akaike's information criterion (AIC, [126]). For a model with p parameters $\boldsymbol{\beta}$,

$$\text{AIC} = -2\ell\left(\widehat{\boldsymbol{\beta}}; \{r_i\}_{i=1}^n\right) + 2p.$$

We use AIC to rank models for comparison, with small AIC being the best. Models are competitive with one another if their AIC values differ by less than two.

5.2.5 Numerical Implementation

Maximization of the weighted log-likelihood (5.3) for Poisson regression can be accomplished with standard software, such as the `R` function `glm`, using case weights (5.4) obtained in the E-step. We developed custom code for fitting of these models, and have made it available as an `R` package called `QRRT`, freely downloadable from GitHub; see the supplemental material Section D.2 for details.

We use multiple starting values and iterate each to convergence, assessed by checking the value of the score vector derived in Section D.1. We then choose the set of converged parameter estimates that yield the highest log-likelihood value. Standard errors (SE's) for each estimated parameter $\widehat{\beta}_j$ are calculated from diagonal elements of the estimated Fisher information matrix (Section D.1). The t-statistic is calculated as $t_j = \widehat{\beta}_j/\text{SE}_j$ and the corresponding p-value is the

probability that the absolute value of a standard normal random variable is greater than or equal to $|t_j|$; that is, the probability under the asymptotic distribution of obtaining a statistic this extreme or more extreme under the null hypothesis that the true β_j coefficient is zero. The code returns AIC and the maximized log-likelihood, so that non-nested models can be compared, and nested models can be tested.

5.3 Results

5.3.1 Monte Carlo Results

We illustrate the methodology and the quality of the asymptotic approximations via a Monte Carlo experiment using our `R` package `QRRT`. Details on reproducing results of this simulation experiment are given in the supplemental material, Section D.2.

We consider a setting in which $n = 400$ true counts are generated independently as $T_i \sim$ Poisson(λ_i) with

$$\ln \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_4 \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_5 x_{i1} x_{i2}, \tag{5.6}$$

where x_{1i} and x_{2i} are continuous predictors and x_{3i} is a categorical predictor with levels "A", "B", and "C", and "A" is the baseline level. We set

$$(\beta_0, \ldots, \beta_5) = (1.5, 1.0, -0.5, 0.4, 0.3, 0.2).$$

Next, we simulate $\{B_i\}_{i=1}^n$ as independent and identically distributed from the same $b(r)$ distribution as in Conteh [114], with $m = 8$ and

$$(b(0), b(1), \ldots, b(8), b(9)) = \frac{1}{50}(6, 7, 4, 2, 2, 1, 1, 1, 1, 25). \tag{5.7}$$

Observations are then $R_i = B_i$ if $B_i \leq m = 8$ and $R_i = T_i$ if $B_i = m + 1 = 9$.

For our Monte Carlo experiment, we fixed $\{(x_{1i}, x_{2i}, x_{3i})\}_{i=1}^{400}$ and, over 1000 independent realizations, simulated $\{T_i\}$ using the model (D.2) and $\{B_i\}$ and $\{R_i\}$ as described. We fitted each of the 1000 simulated data sets both with the true model (D.2), and with the larger-than-necessary model with all two-way interactions,

$$\begin{aligned}&\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_4 \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_5 x_{i1} x_{i2} \\ &+ \beta_6 x_{1i} \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_7 x_{1i} \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_8 x_{2i} \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_9 x_{2i} \mathbf{1}_{\{x_{i3}=\mathrm{C}\}},\end{aligned} \tag{5.8}$$

in which $\beta_6 = \cdots = \beta_9 = 0$.

For each simulated realization and both fits, we recorded the vector of estimated coefficients and the inverse Fisher information evaluated at the estimated parameters. We then compared the average regression coefficient vector over the 1000 Monte Carlo replicates to the vector of true coefficients, to assess the quality of the point estimation, and the empirical covariance matrix over the 1000 Monte Carlo replicates to the asymptotic covariance matrix given by the inverse Fisher information at the true values, to assess the quality of the variance approximation. Further, we compared the average estimated inverse Fisher information to the empirical covariance matrix, to assess the quality of the variance estimators.

Results are given in Table 5.1 and show that the asymptotic approximations are excellent. For both the true additive model and the larger interaction model, the MLE's are approximately unbiased, their variances are well-approximated by diagonal elements of inverse Fisher information, and the estimated variances obtained by plugging MLE's into inverse Fisher information are nearly unbiased for the true variances.

Table 5.1: Simulation Results. True coefficients, estimated parameters, Monte Carlo standard error, inverse Fisher information matrix evaluated at estimated parameters and inverse Fisher information matrix at the true value. All parameters are calculated based on 1000 Monte Carlo replicates with sample size equals to 400.

		True Model				**Interaction Model**			
	True β	$\widehat{\beta}$	Monte Carlo S.E.	Average estimated S.E.	Inverse Fisher at true value	$\widehat{\beta}$	Monte Carlo S.E.	Average estimated S.E.	Inverse Fisher at true value
β_0	1.5	1.4951	0.0815	0.0791	0.0789	1.4899	0.1489	0.1464	0.1456
β_1	1.0	1.0037	0.0690	0.0676	0.0675	1.0067	0.1446	0.1405	0.1400
β_2	-0.5	-0.5005	0.0651	0.0646	0.0645	-0.5008	0.0750	0.0733	0.0728
β_3	0.4	0.3987	0.0514	0.0508	0.0507	0.3998	0.1874	0.1818	0.1808
β_4	0.3	0.3007	0.0523	0.0501	0.0500	0.3031	0.1846	0.1818	0.1808
β_5	0.2	0.2003	0.0626	0.0623	0.0622	0.2006	0.0648	0.0631	0.0632
β_6						0.0000	0.1800	0.1720	0.1715
β_7						-0.0014	0.1776	0.1734	0.1726
β_8						-0.0002	0.0523	0.0512	0.0510
β_9						-0.0008	0.0539	0.0517	0.0520

Finally, for each simulated realization we tested the null hypothesis that the true model (D.2) suffices,

$$H_0 : \ \beta_6 = \cdots = \beta_9 = 0,$$

against the alternative that the larger, two-way model (D.3) is necessary. These hypotheses were compared via a likelihood ratio test, computed as

$$-2\ln(\text{likelihood of true model}) + 2\ln(\text{likelihood of full two-way model}) \tag{5.9}$$

and compared to the χ^2 distribution with 4 degrees of freedom, rejecting H_0 for large values of the test statistic. Since the null hypothesis is true in each simulated realization, the p-values should theoretically follow a uniform distribution. The empirical results (not shown here) are consistent with the uniform distribution. In particular, the empirical proportion of rejections is 0.047 at the 0.05 significance level and 0.092 at the 0.10 significance level.

5.3.2 Application to Poaching in Sierra Leone

We applied our method to responses to the question "how many times per month on average did someone from the household hunt inside the reserve with the use of traps during the nine-month study period?" Instrumental models of compliance [119, 121, 127] posit that compliance is primarily driven by factors external to the individual, including the probability of being caught and convicted. To test for the effects of perceived enforcement we asked respondents if they knew that a protected area existed neighboring their community, if reserve personnel restricted the activities allowed inside the protected area, if reserve personnel patrolled the reserve, if the personnel were efficient in their enforcement duties, if conservation personnel were quick to apprehend those engaged in non-compliant activities in the reserve, and if those caught were punished.

Non-compliance may also be influenced by other perceived costs and benefits of a particular behavior. For example, if resources are rare, the efforts needed to obtain them may outweigh any benefits received. To test the effect of perceived rarity, we asked respondents about the rarity of

targeted species. Similarly, we tested for the effect of household size (the number of people living in the household), as larger households may require more resources, which would increase the likelihood of violating hunting regulations while searching for food.

In addition, alternative livelihoods may reduce the need for subsistence-based hunting practices [128]. We stratified our sample based on access to alternative livelihoods.

Urban centers can both drive more illegal hunting by providing markets for bushmeat, or wage labor in urban areas may reduce illegal hunting by offering alternative livelihoods [128–131]. Therefore, we surveyed communities with both high and low access to the main urban center of Freetown. Similarly, we might predict less illegal hunting in locations with better ocean access, due to the presence of alternative marine-based livelihoods [132]; and therefore we surveyed communities with both direct and no access to the ocean. Sierra Leone's civil war (1992–2002) displaced millions of people. Many of the displaced settled in communities near Freetown. Communities surrounding WAPFR vary widely in terms of the proportion of residents that arrived as internally displaced people during Sierra Leone's civil war. Many of the internally displaced do not have access to suitable land for agriculture or other alternative livelihoods to meet basic needs, which can lead to increases in resource extraction rates from the reserve. We surveyed communities with either no internally displaced people or substantial populations of internally displaced people. We then included community dummy variables in our models to examine the effect of context, including access to alternative livelihoods.

The normative view argues that compliance is more internally driven by perceived behavioral norms [119, 132–134]. Here we explore the effects of descriptive norms, which involve a person's perceptions of the prevalence of a behavior [135]. Based on descriptive norms, we would hypothesize that people will be more likely to violate regulations if they believe many of their peers are also non-compliant. To test for these normative effects we asked respondents if people from their community hunted inside the reserve, and if they thought people from outside the community hunted in the reserve.

Finally, hunting requires specialized knowledge of the local ecosystem and of target species. Ecological knowledge can accumulate over time as hunters compile more first-hand experience, and several ethnobiological studies have found residence time to be positively correlated with increased natural resource use [136, 137]. To the contrary, formal education has often been significantly linked to lower levels of ecological knowledge and subsistence resource use [138–141]. Based on these prior findings, we tested for the effects of both formal education level and local residence time in our models.

Summarizing, we then have the following set of hypothesized drivers and corresponding covariates:

Table 5.2: Drivers and covariates. Hypothesized drivers of non-compliant behavior and corresponding measured covariates in the Sierra Leone dataset.

Driver	Covariates
perceived enforcement	knowledge of protected area, no perceived restriction on extraction, perceived efficiency of conservation personnel, perceived patrols, perceived rapid detention, perceived punishment
perceived rarity	
household size	
alternative livelihoods	rural, seaside, displaced
descriptive norms	residents hunted, outsiders hunted
residence time	
formal education	

Among these covariates, all of the Yes-No-Don't Know variables were converted to Yes indicators, and all agreement scales ($1 =$ Strongly Agree, $\ldots$, $5 =$ Strongly Disagree) were converted to Agreement (Agree or Strongly Agree) indicators. The data set was then restricted to records with non-missing values for all of the above variables, to ensure comparability across fitted models. There are $n = 662$ complete records in this data set. These data are in the supplemental material of [142] and available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6161884/bin/pone.0204433.s003.csv.

We fitted a series of models corresponding to these hypothesized drivers. Each of the models included an intercept. All covariates for a hypothesized driver were either simultaneously included or excluded from a model; for example, all six covariates corresponding to perceived enforcement were either in or out of a given model. Hence, with seven drivers there were $2^7 = 128$ possible additive models for consideration, with the largest model including the intercept and all seven drivers, and the smallest (null) model including only the intercept.

We used our code to fit all of these models, plus three sets of additional models, each with 128 subset models: (1) all subsets of the seven drivers, with alternative livelihoods replaced by (alternative livelihoods)^2, meaning the three community variables plus all three of their two-way interactions; (2) all subsets of the seven drivers, but with the six variables of perceived enforcement replaced by the single variable "Efficient Conservation: perceived efficiency of conservation personnel"; (3) all subsets of the seven drivers, but with both alternative livelihoods replaced by (alternative livelihoods)^2 and perceived enforcement replaced by Efficient Conservation. We computed AIC for all of these subset models and determined minimum AIC within each model class (see Table 5.3). Based on these computations, we restricted attention to the model class Efficient Conservation + $(\text{alternative livelihoods})^2$.

Table 5.3: Minimum AIC for four different model classes. Minimum AIC over all 128 subset models in each model class. All models are fitted to randomized responses based on the EM algorithm with 20 different random starting values to avoid convergence to local modes.

Model Class	Minimum AIC
Efficient Conservation + alternative livelihoods	2728.699
Efficient Conservation + $(\text{alternative livelihoods})^2$	2680.613
perceived enforcement + alternative livelihoods	2729.034
perceived enforcement + $(\text{alternative livelihoods})^2$	2688.773

Table 5.4: Top models with ΔAIC less than 2 for Efficient Conservation + $(\text{alternative livelihoods})^2$. ΔAIC, maximum likelihood estimates for models fitted to randomized responses. All model fits are based on the EM algorithm with 20 different random starting values to avoid convergence to local modes.

	1	2	3	4	5	6	7	8	9
(Intercept)	-0.469	-0.338	-0.460	-0.338	-0.589	-0.319	-0.342	-0.454	-0.451
EfficientConservation	-3.261	-3.316	-3.127	-2.716	-3.213	-2.879	-3.200	-3.248	-3.264
AnimalsRare		-0.218					-0.201		-0.212
HouseholdSize								-0.002	
HighDisplace	0.691	0.671	0.670	0.668	0.713	0.685	0.656	0.688	0.690
Rural	1.404	1.489	1.500	1.664	1.414	1.559	1.559	1.395	1.505
Seaside	1.814	1.762	1.842	1.915	1.790	1.900	1.802	1.810	1.743
Rural:HighDisplace	-0.288	-0.308	-0.354	-0.493	-0.315	-0.397	-0.355	-0.275	-0.340
Seaside:HighDisplace	-0.887	-0.781	-0.886	-0.909	-0.854	-0.909	-0.800	-0.877	-0.751
Seaside:Rural	-2.361	-2.408	-2.382	-2.434	-2.329	-2.457	-2.424	-2.357	-2.381
OutsidersHunted			-0.283	-0.366			-0.273		
ResidentsHunted			0.282	0.342			0.280		
Residencetime	0.007	0.008	0.006		0.007		0.007	0.007	0.008
Education.level					0.038				0.036
AIC	2680.613	2681.220	2681.764	2681.868	2681.916	2682.067	2682.500	2682.591	2682.601
ΔAIC	0	0.607	1.151	1.255	1.303	1.454	1.887	1.978	1.988

Within this model class, we computed ΔAIC as AIC minus minimum AIC, and focused on the nine optimal models with $\Delta\text{AIC} < 2$ (see Table 5.4). We found support for some hypotheses we tested and not for others in this set. In addition, none of the individual hypothesized factors alone explains the variation in frequencies of illegal hunting. The ΔAIC value of models containing just individual factors are between 31.231 and 118.455. Instead, all optimal models contained a combination of different factors.

Other likelihood-based criteria could be applied, such as the Bayesian Information Criterion (BIC) [143]. AIC and BIC both allow for model selection in large model spaces, but using different approaches: AIC efficiently selects a good approximating model in the model space, while BIC consistently estimates the true model if a true model is in fact in the model space. We computed BIC for all 128 models in the same model class as considered for Table 5.4. As expected, BIC tends to prefer smaller models, but model 6 and model 1 in Table 5.4 are the first and second model selected based on BIC.

All optimal models included a large, negative coefficient for perceived enforcement, indicating that higher levels of enforcement may serve as a critical deterrent against illegal hunting in WAPFR (Table 5.4), as has been found in a wide variety of other protected areas. This outcome has clear policy and management implications; however, the potential to increase enforcement may be limited in Sierra Leone. The country faces many fiscal challenges, and conservation capacity in WAPFR has yet to recover to levels seen prior to the civil war [116].

Three of the nine optimal models also point to the importance of normative influences on the amount of non-compliance (Table 5.4). We found that community members were more likely to engage in illegal hunting when they believed their neighbors in the same community were also doing so (positive coefficients in Table 5.4). Norms have been shown to influence compliance with conservation regulations and to shape natural resource use patterns across a broad range of contexts from recreational fishing in New Zealand [119] to rangeland management in Mongolia [144]. Management interventions can influence norms, but care must be taken as the introduction of new rules and regulations can undermine long-standing norms and drive greater non-compliance

[145]. One promising approach is community-based social marketing, which can use social norms as the center piece of persuasive behavior-change communication campaigns [146].

We also found that, contrary to initial hypotheses, respondents were less likely to hunt illegally if they perceived outsiders were hunting in the reserve (negative coefficients in Table 5.4). One possible reason for this apparent contradiction is that the effects of descriptive norms are moderated by group identity. Specifically, when an individual perceives a group to be more similar to themselves, the individual may identify more closely with the group, and this may increase the influence of descriptive norms on the individual's behavior [147]. In other words, individuals should be more likely to participate in a behavior that is common among a group they identify with (in this case their home community) than a behavior common in a less similar group (in this case outsiders). This could explain why the perceived behaviors of outsiders would have less effect on the amount of non-compliance than the behaviors of community members. However, we found that the effect of outsiders was as strong as that of community members, but in the opposite direction: perceptions of hunting by outsiders correlates with less hunting by respondents and perceptions of hunting by community members correlates with more hunting. The effect of outsider's behaviors may instead be explained by the history of the region. During the war, combatants frequented the forest inside the reserve, and local people may still harbor memories that associate the forest with zones of active combat [115]. Therefore, increased activity of outsiders in the reserve may provide local people with ample reason to avoid the area.

All nine optimal models also included community variables (Table 5.4). As described above, we had included community as a variable in our models as a proxy for access to alternative livelihoods. Some of the results support the idea that increased availability of alternative livelihoods can reduce resource use and non-compliance with conservation regulations. For example, rural communities, with less access to labor markets in urban centers, tended to hunt more in the reserve (positive coefficients in Table 5.4). In addition, communities with a greater proportion of displaced people also hunted more. However, contrary to our hypotheses, we found more hunting to occur in seaside communities, despite their access to additional marine resources. Also, examining in-

teraction effects among community types, further confounds the relationship between access to alternative livelihoods and frequency of hunting. For example, we would expect rural communities with many displaced people to have high rates of hunting, however, all of our optimal models found that these communities had lower rates of hunting (negative coefficient in Table 5.4 for interaction between rural and displaced). Overall, we see a significant difference in hunting rates among communities, but these differences cannot be explained by access to alternative livelihoods. Instead, other aspects of the local context not measured here must be driving these differences in hunting rates among communities.

Seven of the nine optimal models also contained residence time. The small coefficients (Table 5.4) indicate the smaller effect the variable had on the outcome. In all cases the longer a household had lived in a community, the greater the likelihood they had participated in illegal hunting. This corroborates prior findings of increases in the use of forest resources with longer residence times, which may be linked to the accumulation of ecological knowledge over time [136]. Three of the optimal models included perceived rarity, and supported the prediction that residents were less likely to participate in illegal hunting when they perceived animals to be rare in the reserve. Only two models included education, but contrary to prior studies [138–141] our results indicate that increases in formal education are associated with greater amounts of illegal hunting. However, some studies in Africa have found similar results using indirect questioning methods [148, 149]. In addition to the value of using an indirect questioning method such as QRRT, our finding might be explained by the links between hunting and bushmeat markets in the nearby capital of Freetown. Higher levels of education may assist some families in integrating with these markets, but further research is needed to confirm this possible link between education and hunting. Finally, only one model in the optimal set contained household size. The coefficient for the variable was small and surprisingly indicated that larger households would be slightly less likely to hunt illegally in the reserve. Although this finding is in contrast to theory, similar results have been recorded in Gabon in the case of hunting for bushmeat [150].

Although our models allow us to compare the relative importance of possible drivers of illegal hunting, the models still only explained a relatively small proportion of the variance in hunting rates. This is not surprising given that our aim was to use this case to demonstrate a new methodological approach and we did not attempt to measure all possible determinants of non-compliance. For this case, future research might include variables or models not tested here, but for which strong theoretical foundations exist. Possibilities for additional theories to test that have been found to be good predictions of conservation-related behavior in past studies include the theory of planned behavior, which posits that attitudes and perceived behavioral control, along with social norms all influence behavioral intentions [151], Bamberg and Moser's [152] framework of pro-environmental behavior, and models of legitimacy, which include both measures of participation in decision-making as well as perceptions of the fairness of rules and enforcement outcomes [153, 154].

5.4 Conclusion

The methods we present here provide a methodological blueprint for examining possible drivers of sensitive behaviors. Researchers across multiple disciplines are interested in understanding sensitive behaviors, and policy makers and program managers seek more effective means to reduce the frequency of a wide variety of sensitive behaviors. QRRT provides a means for gathering data on the frequency of sensitive behaviors while protecting respondent anonymity. The new analytical approach and tools we present here will allow researchers to explore drivers of a wide variety of sensitive behaviors using QRRT data.

Chapter 6

Conclusion and Future Work

In this dissertation, we developed new statistical models, testing frameworks and inferential procedures based on classic generalized linear models including Gaussian, Poisson and negative binomial regressions. These methods are designed specifically for three classes of important problems arising from genomics and sociological contexts. The good performance of these methods has been demonstrated using extensive numerical simulations, and the methods have been applied to the real-world data that motivated these studies.

The first class of problems involves count data arising from longitudinal RNA-seq experiments. In particular, we considered two questions: 1) whether the treatment affects the geometric attributes of the temporal profiles and 2) whether any treatment effect varies over time. To answer the first question, in Chapter 2 we modeled the transformed count data for genes at each time point using a Gaussian distribution and developed a testing framework based on a permutation procedure. We show that it achieves good power and has its FDR controlled via simulation studies and we applied it to the data collected from a light physiology experiment on maize. In Chapter 3, we focused on solving question 2. We propose an inferential procedure that maximizes average power and controls FDR. Conditional on a latent Gaussian mixture, the time-course RNA-seq data is modeled by negative binomial distributions. This latent Gaussian-Negative Binomial model allows feasible estimation of unknown model parameters and testing a variety of general composite null hypotheses of great biological interest. The simulation studies in Chapter 3 show the advantages of our proposal over existing methods.

The second class of problems involves analyzing group-structured sRNA data that consist of independent replicates of counts for each sRNA across experimental conditions. In Chapter 4, we introduce an inferential procedure based on a group-based negative binomial model and conduct a testing framework via bootstrap. This procedure not only provides a group-based normalization

factor, but also enables the group-based DE analysis. Our method shows good performance in both simulation studies and real data analysis of experiment on roundworm.

The third class of problems concerns count data distributed as a mixture that arises due to the randomization mechanism used in QRRT to guarantee respondent anonymity. This guaranteed anonymity allows sociological researchers to investigate sensitive behaviors. In Chapter 5, we propose a Poisson regression method that can be used to identify potential drivers of non-compliant behaviors. This method is based on maximum likelihood estimation and the model parameters are estimated via the EM algorithm. As a case study, we use this approach to compare the relative importance of possible drivers of illegal hunting in Sierra Leone.

There are many possible extensions and future works for the topics covered by this dissertation. These future research directions in general can be categorized into three subjects: seeking alternative base models, increasing current model complexity and improving performance by further investigating model's theoretical properties. In Chapter 2, we modeled the transformed count data using a Gaussian distribution. Alternatively, without losing any information, one can model the gene's temporal profiles by Poisson or negative binomial distributions. In Chapter 3, one future direction is to generalize from the two-sample problem to three or more treatments. This extension would provide more sophisticated ways to identify genes' reaction to multi-level treatments. In Chapter 4, to reduce the high computational cost on the testing procedure due to the bootstrap, one possible improvement is to conduct the testing procedure based on the asymptotic variance-covariance matrix of the regression parameter estimates.

Bibliography

[1] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[2] Graham Peers, Nathan Sindt, Wen Zhou, Jessica Prenni, Chris L. Dupont, and Andrew E Allen. A systems level investigation of low-light acclimation in the marine diatom phaeodactylum tricornutum. *preprint for New Phytologist*, 2016.

[3] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[4] Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, 99(2):248–256, 2012.

[5] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14(1):91, 2013.

[6] Yang Xie, Wei Pan, and Arkady B Khodursky. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280–4288, 2005.

[7] Xiaoxiao Sun, David Dalpiaz, Di Wu, Jun S Liu, Wenxuan Zhong, and Ping Ma. Statistical inference for time course rna-seq data using a negative binomial mixed-effect model. *BMC bioinformatics*, 17(1):324, 2016.

[8] Michael Hackenberg, Perry Gustafson, Peter Langridge, and Bu-Jun Shi. Differential expression of micro RNA s and other small RNA s in barley between water and drought conditions. *Plant Biotechnology Journal*, 13(1):2–13, 2015.

[9] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

[10] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[11] Martin J Aryee, José A Gutiérrez-Pabello, Igor Kramnik, Tapabrata Maiti, and John Quackenbush. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC bioinformatics*, 10(1):409, 2009.

[12] John D Storey, Wenzhong Xiao, Jeffrey T Leek, Ronald G Tompkins, and Ronald W Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[13] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[14] JE Eckel, C Gennings, VM Chinchilli, LD Burgoon, and TR Zacharewski. Empirical bayes gene screening tool for time-course or dose–response microarray data. *Journal of biopharmaceutical statistics*, 14(3):647–670, 2004.

[15] Yu Chuan Tai, Terence P Speed, et al. A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics*, 34(5):2387–2412, 2006.

[16] Jiang Qian, Marisa Dolled-Filhart, Jimmy Lin, Haiyuan Yu, and Mark Gerstein. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression

profiles identifies new, biologically relevant interactions. *Journal of molecular biology*, 314(5):1053–1066, 2001.

[17] Marco F Ramoni, Paola Sebastiani, and Isaac S Kohane. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, 99(14):9121–9126, 2002.

[18] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K Gifford, and Tommi S Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences*, 100(18):10146–10151, 2003.

[19] Xie L Xu, James M Olson, and Lue Ping Zhao. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible huntington's disease transgenic model. *Human Molecular Genetics*, 11(17):1977–1985, 2002.

[20] Yuedong Wang. Mixed effects smoothing spline analysis of variance. *Journal of the royal statistical society: Series b (statistical methodology)*, 60(1):159–174, 1998.

[21] Gareth M James and Trevor J Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.

[22] John A Rice and Colin O Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1):253–259, 2001.

[23] Sunghee Oh, Seongho Song, Gregory Grabowski, Hongyu Zhao, and James P Noonan. Time series expression analyses using rna-seq: a statistical approach. *BioMed research international*, 2013.

[24] Olga A Vsevolozhskaya, Dmitri V Zaykin, Mark C Greenwood, Changshuai Wei, and Qing Lu. Functional analysis of variance for association studies. *PloS one*, 9(9):e105074, 2014.

[25] Piotr Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[26] Guanghui Wang, Changliang Zou, and Guosheng Yin. Change-point detection in multinomial data with a large number of categories. *Annals of Statistics*, 2017.

[27] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[28] John D Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[29] Kai-Florian Storch, Ovidiu Lipan, Igor Leykin, N Viswanathan, Fred C Davis, Wing H Wong, and Charles J Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78–83, 2002.

[30] Robert T Furbank and William C Taylor. Regulation of photosynthesis in C3 and C4 plants: a molecular approach. *The Plant Cell*, 7(7):797, 1995.

[31] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[32] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[33] María José Nueda, Sonia Tarazona, and Ana Conesa. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, 30(18):2598–2602, 2014.

[34] Agata Michna, Herbert Braselmann, Martin Selmansberger, Anne Dietz, Julia Hess, Maria Gomolka, Sabine Hornhardt, Nils Blüthgen, Horst Zitzelsberger, and Kristian Unger. Natural cubic spline regression modeling followed by dynamic network reconstruction for the

identification of radiation-sensitivity gene association networks from time-course transcriptome data. *PloS one*, 11(8):e0160791, 2016.

[35] Tarmo Äijö, Vincent Butty, Zhi Chen, Verna Salo, Subhash Tripathi, Christopher B Burge, Riitta Lahesmaa, and Harri Lähdesmäki. Methods for time series analysis of rna-seq data with application to human th17 cell differentiation. *Bioinformatics*, 30(12):i113–i120, 2014.

[36] Ning Leng, Yuan Li, Brian E McIntosh, Bao Kim Nguyen, Bret Duffin, Shulan Tian, James A Thomson, Colin N Dewey, Ron Stewart, and Christina Kendziorski. Ebseq-hmm: a bayesian approach for identifying gene-expression changes in ordered rna-seq experiments. *Bioinformatics*, 31(16):2614–2622, 2015.

[37] Denis Agniel and Boris P Hejblum. Variance component score test for time-course gene set analysis of longitudinal rna-seq data. *Biostatistics*, 18(4):589–604, 2017.

[38] David S Fischer, Fabian J Theis, and Nir Yosef. Impulse model-based differential expression analysis of time course sequencing data. *Nucleic acids research*, 46(20):e119–e119, 2018.

[39] Markus Heinonen, Olivier Guipaud, Fabien Milliat, Valérie Buard, Béatrice Micheau, Georges Tarlet, Marc Benderitter, Farida Zehraoui, and Florence d'Alché Buc. Detecting time periods of differential gene expression using gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics*, 31(5):728–735, 2014.

[40] Dan Luo, Sara Ziebell, and Lingling An. An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics*, 33(9):1286–1292, 2017.

[41] Hande Topa and Antti Honkela. GPrank: an R package for detecting dynamic elements from genome-wide time series. *BMC bioinformatics*, 19(1):367, 2018.

[42] Daniel Spies, Peter F Renz, Tobias A Beyer, and Constance Ciaudo. Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Briefings in bioinformatics*, 20(1):288–298, 2017.

[43] Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany*, 99(2):248–256, 2012.

[44] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[45] Chen Chu, Zhaoben Fang, Xing Hua, Yaning Yang, Enguo Chen, Allen W Cowley, Mingyu Liang, Pengyuan Liu, and Yan Lu. degps is a powerful tool for detecting differential expression in rna-sequencing studies. *BMC genomics*, 16(1):455, 2015.

[46] Koen Van den Berge, Charlotte Soneson, Mark D Robinson, and Lieven Clement. stager: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome biology*, 18(1):151, 2017.

[47] Yet Nguyen. Multiple hypothesis testing and rna-seq differential expression analysis accounting for dependence and relevant covariates. 2018.

[48] JT Gene Hwang and Peng Liu. Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[49] John D Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007.

[50] Yaqing Si and Peng Liu. An optimal test with maximum average power while controlling fdr with application to rna-seq data. *Biometrics*, 69(3):594–605, 2013.

[51] Xiangqin Cui, JT Gene Hwang, Jing Qiu, Natalie J Blades, and Gary A Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 2005.

[52] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2007.

[53] Richard A Davis and Rongning Wu. A negative binomial model for time series of counts. *Biometrika*, 96(3):735–749, 2009.

[54] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

[55] Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

[56] Sijian Wang and Ji Zhu. Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448, 2008.

[57] Henry Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

[58] Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

[59] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233, 1995.

[60] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

[61] Christian Hennig. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

[62] Jogi Henna. Examples of identifiable mixture. *Journal of the Japan Statistical Society, Japanese Issue*, 24(2):193–200, 1994.

[63] Philippe Heinrich, Jonas Kahn, et al. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

[64] Peiming Wang, Martin L Puterman, Iain Cockburn, and Nhu Le. Mixed poisson regression models with covariate dependent rates. *Biometrics*, pages 381–400, 1996.

[65] Nhat Ho, XuanLong Nguyen, et al. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016.

[66] Aaron TL Lun, Yunshun Chen, and Gordon K Smyth. ItâĂŹs de-licious: a recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edger. In *Statistical Genomics*, pages 391–416. Springer, 2016.

[67] Chong Gu. Model diagnostics for smoothing spline anova models. *Canadian Journal of Statistics*, 32(4):347–358, 2004.

[68] Paul G Falkowski and John A Raven. *Aquatic photosynthesis*. Princeton University Press, 2013.

[69] Chris Bowler, Andrew E Allen, Jonathan H Badger, Jane Grimwood, Kamel Jabbari, Alan Kuo, Uma Maheswari, Cindy Martens, Florian Maumus, Robert P Otillar, et al. The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239, 2008.

[70] Yunshun Chen, Aaron TL Lun, and Gordon K Smyth. Differential expression analysis of complex rna-seq experiments using edger. In *Statistical analysis of next generation sequencing data*, pages 51–74. Springer, 2014.

[71] Belinda Phipson, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics*, 10(2):946, 2016.

[72] Shirley Tam, Ming-Sound Tsao, and John D McPherson. Optimization of miRNA-seq data preprocessing. *Briefings in Bioinformatics*, 16(6):950–963, 2015.

[73] Elisabeth J Chapman and James C Carrington. Specialization and evolution of endogenous small RNA pathways. *Nature Reviews Genetics*, 8(11):884, 2007.

[74] James A Birchler and Harsh H Kavi. Slicing and dicing for small RNAs. *Science*, 320(5879):1023–1024, 2008.

[75] Daniela Witten, Robert Tibshirani, Sam Guoping Gu, Andrew Fire, and Weng-Onn Lui. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*, 8(1):58, 2010.

[76] Sam Griffiths-Jones, Russell J Grocock, Stijn Van Dongen, Alex Bateman, and Anton J Enright. mirbase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl_1):140–144, 2006.

[77] Qinghua Liu and Zain Paroo. Biochemical principles of small RNA pathways. *Annual Review of Biochemistry*, 79:295–319, 2010.

[78] illumina. illumina-Small RNA Sequencing. https://www.illumina.com/techniques/sequencing/rna-sequencing/small-rna-seq.html, 2015. [Online; Accessed: 2019-11-25].

[79] Kevin P McCormick, Matthew R Willmann, and Blake C Meyers. Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, 2(1):2, 2011.

[80] Dillies, Marie-Agnès and Rau, Andrea and Aubert, Julie and Hennequet-Antier, Christelle and Jeanmougin, Marine and Servant, Nicolas and Keime, Céline and Marot, Guillemette and Castel, David and Estelle, Jordi and others. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[81] Peipei Li, Yongjun Piao, Ho Sun Shon, and Keun Ho Ryu. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16(1):347, 2015.

[82] Ana Conesa, María José Nueda, Alberto Ferrer, and Manuel Talón. masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9):1096–1102, 2006.

[83] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the DESeq2 package. *Genome Biol*, 15(550):10–1186, 2014.

[84] Meng Cao, Wen Zhou, F Jay Breidt, and Graham Peers. Large scale maximum average power multiple inference on time-course count data with application to RNA-Seq analysis. *Biometrics*, 2019.

[85] Ran Dai and Rina Foygel Barber. The knockoff filter for fdr control in group-sparse and multitask regression. *arXiv preprint arXiv:1602.03589*, 2016.

[86] Rina Foygel Barber and Aaditya Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1247–1268, 2017.

[87] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[88] Guilin Wang and Valerie Reinke. A c. elegans piwi, prg-1, regulates 21u-rnas during spermatogenesis. *Current Biology*, 18(12):861–867, 2008.

[89] Kailee J Reed, Joshua M Svendsen, Kristen C Brown, Brooke E Montgomery, Taylor N Marks, Tarah Vijayasarathy, Dylan M Parker, Erin Osborne Nishimura, Dustin L Updike, and Taiowa A Montgomery. Widespread roles for pirnas and wago-class sirnas in shaping the germline transcriptome of caenorhabditis elegans. *Nucleic Acids Research*, 2019.

[90] Shirley Tam, Ming-Sound Tsao, and John D McPherson. Optimization of mirna-seq data preprocessing. *Briefings in bioinformatics*, 16(6):950–963, 2015.

[91] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

[92] William Michael Landau and Peng Liu. Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods. *PloS one*, 8(12), 2013.

[93] T Strub, S Giuliano, T Ye, C Bonet, C Keime, D Kobi, S Le Gras, M Cormont, R Ballotti, C Bertolotto, et al. Essential role of microphthalmia transcription factor for dna replication, mitosis and genomic stability in melanoma. *Oncogene*, 30(20):2319, 2011.

[94] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

[95] Pedro López-Romero. Pre-processing and differential expression analysis of agilent microrna arrays using the agimicrorna bioconductor library. *BMC genomics*, 12(1):64, 2011.

[96] Andrea Rau, Mélina Gallopin, Gilles Celeux, and Florence Jaffrézic. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, 29(17):2146–2152, 2013.

[97] Michael C Gavin, Jennifer N Solomon, and Sara G Blank. Measuring and monitoring illegal use of natural resources. *Conservation Biology*, 24(1):89–100, 2010.

[98] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[99] Abdel-Latif A Abul-Ela, Gernard G Greenberg, and Daniel G Horvitz. A multi-proportions randomized response model. *Journal of the American Statistical Association*, 62(319):990–1008, 1967.

[100] Bernard G Greenberg, Roy R Kuebler Jr, James R Abernathy, and Daniel G Horvitz. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66(334):243–250, 1971.

[101] PT Liu and LP Chow. A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, 71(353):72–73, 1976.

[102] James Alan Fox and Paul E. Tracy. *Randomized Response: A Method for Sensitive Surveys*. SAGE Publications, 1986.

[103] Arijit Chaudhuri and Rahul Mukerjee. *Randomized Response: Theory and Techniques*. Marcel Dekker, 1988.

[104] Gerty JLM Lensvelt-Mulders, Joop J Hox, Peter GM Van der Heijden, and Cora JM Maas. Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, 33(3):319–348, 2005.

[105] Uchila N Umesh and Robert A Peterson. A critical evaluation of the randomized response method: Applications, validation, and research agenda. *Sociological Methods & Research*, 20(1):104–138, 1991.

[106] Jennifer Solomon, Susan K Jacobson, Kenneth D Wald, and Michael Gavin. Estimating illegal resource use at a ugandan park with the randomized response technique. *Human Dimensions of Wildlife*, 12(2):75–88, 2007.

[107] Alyssa S Thomas, Michael C Gavin, and Taciano L Milfont. Estimating non-compliance among recreational fishers: Insights into factors affecting the usefulness of the randomized response and item count techniques. *Biological Conservation*, 189:24–32, 2015.

[108] Ardo van den Hout, Peter GM van der Heijden, and Robert Gilchrist. The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis*, 51(12):6060–6069, 2007.

[109] Maarten JLF Cruyff, Ardo Van Den Hout, and Peter GM Van Der Heijden. The analysis of randomized response sum score variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):21–30, 2008.

[110] Maarten JLF Cruyff, Ulf Böckenholt, Ardo van den Hout, and Peter GM van der Heijden. Accounting for self-protective responses in randomized response data from a social security survey using the zero-inflated Poisson model. *The Annals of Applied Statistics*, pages 316–331, 2008.

[111] Charlotte Chang and Maarten J.L.F. Cruyff. *zapstRR: ZoologicAl Package for Randomized Response Technique (RRT)*, 2017. R package version 0.0.0.9000.

[112] Daniel W. Heck and Morten Moshagen. *RRreg: Correlation and Regression Analyses for Randomized Response Data*, 2017. R package version 0.6.2).

[113] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[114] Abu Conteh, Michael C Gavin, and Jennifer Solomon. Quantifying illegal hunting: A novel application of the quantitative randomized response technique. *Biological Conservation*, 189:16–23, 2015.

[115] Abu Conteh and Michael C Gavin. Influence of war on hunting patterns and pressure in sierra leone. *Environmental Conservation*, 44(2):131–138, 2017.

[116] Abu Conteh, Michael C Gavin, and Joe McCarter. Assessing the impacts of war on perceived conservation capacity and threats to biodiversity. *Biodiversity and Conservation*, 26(4):983–996, 2017.

[117] Abu Conteh. *Impact of War on Biodiversity Conservation in the Western Area Peninsula Forest Reserve, Sierra Leone: A Thesis Submitted to the Victoria University of Wellington in Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Environmental Studies*. PhD thesis, Victoria University of Wellington, 2010.

[118] Paul E Tracy and James Alan Fox. The validity of randomized response for sensitive measurements. *American Sociological Review*, pages 187–200, 1981.

[119] Alyssa S Thomas, Taciano L Milfont, and Michael C Gavin. A new approach to identifying the drivers of regulation compliance using multivariate behavioural models. *PloS one*, 11(10):e0163868, 2016.

[120] Jennifer N Solomon, Michael C Gavin, and Meredith L Gore. Detecting and understanding non-compliance with conservation rules, 2015.

[121] Gary S Becker. Crime and punishment: An economic approach. In *The Economic Dimensions of Crime*, pages 13–68. Springer, 1968.

[122] Freya AV St John, Gareth Edwards-Jones, and Julia PG Jones. Conservation and human behaviour: lessons from social psychology. *Wildlife Research*, 37(8):658–667, 2011.

[123] Aidan Keane, Julia PG Jones, Gareth Edwards-Jones, and Eleanor J Milner-Gulland. The sleeping policeman: understanding issues of enforcement and compliance in conservation. *Animal conservation*, 11(2):75–82, 2008.

[124] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013.

[125] John Ashworth Nelder and R Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

[126] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.

[127] Dennis M King and Jon G Sutinen. Rational noncompliance and the liquidation of northeast groundfish resources. *Marine Policy*, 34(1):7–21, 2010.

[128] Justin S Brashares, Christopher D Golden, Karen Z Weinbaum, Christopher B Barrett, and Grace V Okello. Economic and geographic drivers of wildlife consumption in rural africa. *Proceedings of the National Academy of Sciences*, 108(34):13931–13936, 2011.

[129] David S Wilkie and Julia F Carpenter. Bushmeat hunting in the congo basin: an assessment of impacts and options for mitigation. *Biodiversity & Conservation*, 8(7):927–955, 1999.

[130] Richard Damania, EJ Milner-Gulland, and DJ Crookes. A bioeconomic analysis of bushmeat hunting. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1560):259–266, 2005.

[131] Emmanuel De Merode, Kes Hillman Smith, Katherine Homewood, Richard Pettifor, Marcus Rowcliffe, and Guy Cowlishaw. The impact of armed conflict on protected-area efficacy in central africa. *Biology Letters*, 3(3):299–301, 2007.

[132] Justin S Brashares, Peter Arcese, Moses K Sam, Peter B Coppolillo, Anthony RE Sinclair, and Andrew Balmford. Bushmeat hunting, wildlife declines, and fish supply in west africa. *Science*, 306(5699):1180–1183, 2004.

[133] César Viteri and Carlos Chávez. Legitimacy, local participation, and compliance in the galápagos marine reserve. *Ocean & Coastal Management*, 50(3):253–274, 2007.

[134] Freya AV St John, Aidan M Keane, Gareth Edwards-Jones, Lauren Jones, Richard W Yarnell, and Julia PG Jones. Identifying indicators of illegal behaviour: carnivore killing in human-managed landscapes. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20111228, 2011.

[135] Robert B Cialdini. Descriptive social norms as underappreciated sources of social control. *Psychometrika*, 72(2):263, 2007.

[136] Michael C Gavin and Gregory J Anderson. Socioeconomic predictors of forest use values in the peruvian amazon: A potential tool for biodiversity conservation. *Ecological Economics*, 60(4):752–762, 2007.

[137] Jessica L'Roe and Lisa Naughton-Treves. Effects of a policy-induced income shock on forest-dependent households in the peruvian amazon. *Ecological Economics*, 97:1–9, 2014.

[138] Ravi Hegde and T Enters. Forest products and household economy: a case study from mudumalai wildlife sanctuary, southern india. *Environmental Conservation*, 27(3):250–259, 2000.

[139] Celeste Lacuna-Richman. The socioeconomic significance of subsistence non-wood forest products in leyte, philippines. *Environmental Conservation*, 29(2):253–262, 2002.

[140] Joe McCarter and Michael C Gavin. Local perceptions of changes in traditional ecological knowledge: a case study from malekula island, vanuatu. *Ambio*, 43(3):288–296, 2014.

[141] Marsha B Quinlan and Robert J Quinlan. Modernization and medicinal plant knowledge in a caribbean horticultural village. *Medical Anthropology Quarterly*, 21(2):169–192, 2007.

[142] Meng Cao, F Jay Breidt, Jennifer N Solomon, Abu Conteh, and Michael C Gavin. Understanding the drivers of sensitive behavior using poisson regression from quantitative randomized response technique data. *PloS one*, 13(9), 2018.

[143] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[144] Maria E Fernandez-Gimenez. The role of mongolian nomadic pastoralists' ecological knowledge in rangeland management. *Ecological Applications*, 10(5):1318–1326, 2000.

[145] Julia PG Jones, Mijasoa M Andriamarovololona, and Neal Hockley. The importance of taboos and social norms to conservation in madagascar. *Conservation Biology*, 22(4):976–986, 2008.

[146] Doug McKenzie-Mohr. *Fostering sustainable behavior: An introduction to community-based social marketing*. New Society Publishers, 2011.

[147] Rajiv N Rimal, Maria K Lapinski, Rachel J Cook, and Kevin Real. Moving toward a theory of normative influences: How perceived benefits and similarity moderate the impact of descriptive norms on behaviors. *Journal of Health Communication*, 10(5):433–450, 2005.

[148] ANA Nuno, Nils Bunnefeld, Loiruck C Naiman, and Eleanor J Milner-Gulland. A novel approach to assessing the prevalence and drivers of illegal bushmeat hunting in the serengeti. *Conservation Biology*, 27(6):1355–1365, 2013.

[149] Robin C Whytock, Bethan J Morgan, Taku Awa, Zacharie Bekokon, Ekwoge A Abwe, Ralph Buij, Munir Virani, Juliet A Vickery, and Nils Bunnefeld. Quantifying the scale and socioeconomic drivers of bird hunting in central african forest communities. *Biological Conservation*, 218:18–25, 2018.

[150] Steffen Foerster, David S Wilkie, Gilda A Morelli, Josefien Demmer, Malcolm Starkey, Paul Telfer, Matthew Steil, and Arthur Lewbel. Correlates of bushmeat hunting among remote rural households in gabon, central africa. *Conservation Biology*, 26(2):335–344, 2012.

[151] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[152] Sebastian Bamberg and Guido Möser. Twenty years after hines, hungerford, and tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. *Journal of Environmental Psychology*, 27(1):14–25, 2007.

[153] Tom R Tyler. *Why people obey the law*. Princeton University Press, 2006.

[154] Jesper Raakjær Nielsen and Christoph Mathiesen. Important factors influencing rule compliance in fisheries lessons from denmark. *Marine Policy*, 27(5):409–416, 2003.

[155] Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

[156] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[157] Art B Owen. Quasi-monte carlo sampling. *Monte Carlo Ray Tracing: Siggraph*, 1:69–88, 2003.

[158] John H Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2(1):84–90, 1960.

[159] Wayne A Fuller. *Sampling statistics*, volume 560. John Wiley & Sons, 2011.

[160] Yoshiya Oda, Takeshi Nagasu, and Brian T Chait. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nature biotechnology*, 19(4):379, 2001.

[161] Laurence Boudière, Morgane Michaud, Dimitris Petroutsos, Fabrice Rébeillé, Denis Falconet, Olivier Bastien, Sylvaine Roy, Giovanni Finazzi, Norbert Rolland, Juliette Jouhet, et al. Glycerolipids in photosynthesis: composition, synthesis and trafficking. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1837(4):470–480, 2014.

[162] Jennifer Levering, Jared Broddrick, Christopher L Dupont, Graham Peers, Karen Beeri, Joshua Mayers, Alessandra A Gallina, Andrew E Allen, Bernhard O Palsson, and Karsten

Zengler. Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One*, 11(5):e0155038, 2016.

[163] Peter McCullagh. *Generalized linear models*. Routledge, 1990.

[164] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

Appendix A

Supplemental materials for Chapter 2

A.1 More Simulation Results

Table A.1: Simulation A: Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (A) with $\sigma = 0.7$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.020	0.076	0.098	0.160	0.044	0.071	0.095	0.155
		Power	0.937	0.958	0.96	0.966	0.962	0.966	0.969	0.972
	T_g based	FDR	0.060	0.087	0.113	0.171	0.06	0.086	0.115	0.173
		Power	0.955	0.959	0.962	0.966	0.965	0.968	0.970	0.973
	Standard	FDR	0.021	0.015	0.014	0.013	0.021	0.016	0.014	0.014
		Power	0.510	0.701	0.763	0.827	0.526	0.707	0.779	0.829
9	FANOVA based	FDR	0.048	0.070	0.095	0.144	0.046	0.066	0.098	0.136
		Power	0.964	0.967	0.969	0.972	0.97	0.973	0.976	0.978
	T_g based	FDR	0.058	0.085	0.103	0.159	0.058	0.085	0.116	0.160
		Power	0.965	0.968	0.970	0.973	0.972	0.975	0.977	0.979
	Standard	FDR	0.010	0.010	0.009	0.009	0.012	0.011	0.011	0.011
		Power	0.765	0.812	0.843	0.889	0.761	0.809	0.832	0.880
15	FANOVA based	FDR	0.048	0.064	0.090	0.132	0.044	0.066	0.082	0.137
		Power	0.973	0.975	0.977	0.979	0.979	0.981	0.982	0.984
	T_g based	FDR	0.055	0.077	0.108	0.157	0.055	0.082	0.104	0.137
		Power	0.974	0.976	0.977	0.980	0.979	0.982	0.983	0.984
	Standard	FDR	0.006	0.006	0.006	0.007	0.008	0.008	0.008	0.008
		Power	0.858	0.881	0.903	0.927	0.858	0.882	0.906	0.930

Table A.2: Simulation B: (Peak) Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (B) with $\sigma = 1$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.048	0.076	0.099	0.156	0.053	0.079	0.105	0.141
		Power	0.991	0.997	0.997	0.998	1.000	1.000	1.000	1.000
	T_g based	FDR	0.079	0.107	0.132	0.198	0.053	0.079	0.105	0.163
		Power	0.997	0.997	0.998	0.998	1.000	1.000	1.000	1.000
	Standard	FDR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Power	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	FANOVA based	FDR	0.050	0.075	0.097	0.147	0.047	0.072	0.089	0.141
		Power	0.998	0.999	0.999	0.999	1.000	1.000	1.000	1.000
	T_g based	FDR	0.071	0.097	0.123	0.177	0.047	0.072	0.112	0.141
		Power	0.999	0.999	0.999	0.999	1.000	1.000	1.000	1.000
	Standard	FDR	0.000	0.000	0.000	0.051	0.000	0.000	0.000	0.000
		Power	0.000	0.000	0.000	0.998	0.000	0.000	0.001	0.001
15	FANOVA based	FDR	0.050	0.071	0.090	0.141	0.047	0.067	0.101	0.150
		Power	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	T_g based	FDR	0.064	0.090	0.120	0.166	0.047	0.067	0.101	0.150
		Power	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Standard	FDR	0.060	0.032	0.046	0.080	0.000	0.000	0.005	0.024
		Power	0.000	0.999	0.999	1.000	0.000	0.000	1.000	1.000

Table A.3: Simulation B: (Valley) Average FDR and power for different methods to estimate $\widehat{\text{FP}}$ along with out test in Section 2.2.3 for simulation setting (B) with $\sigma = 1$. Results for different sample sizes n, observation time points T, and nominal levels α are reported based on 100 replications.

	T		10				15			
n	Methods	α	0.050	0.075	0.100	0.150	0.050	0.075	0.100	0.150
6	FANOVA based	FDR	0.000	0.000	0.000	0.133	0.015	0.031	0.050	0.086
		Power	0.000	0.000	0.000	0.258	0.232	0.276	0.305	0.346
	T_g based	FDR	0.000	0.000	0.000	0.153	0.045	0.071	0.097	0.148
		Power	0.000	0.000	0.000	0.277	0.299	0.331	0.356	0.399
	Standard	FDR	0.000	0.000	0.000	0.000	0.000	0.022	0.039	0.077
		Power	0.000	0.000	0.000	0.000	0.000	0.257	0.291	0.338
9	FANOVA based	FDR	0.041	0.066	0.093	0.137	0.022	0.037	0.062	0.101
		Power	0.286	0.324	0.352	0.390	0.359	0.392	0.428	0.470
	T_g based	FDR	0.050	0.074	0.104	0.152	0.044	0.072	0.086	0.144
		Power	0.302	0.332	0.361	0.402	0.404	0.442	0.456	0.500
	Standard	FDR	0.025	0.041	0.059	0.104	0.015	0.032	0.044	0.086
		Power	0.248	0.286	0.316	0.361	0.340	0.381	0.404	0.456
15	FANOVA based	FDR	0.042	0.063	0.088	0.128	0.022	0.044	0.058	0.104
		Power	0.432	0.459	0.482	0.515	0.496	0.532	0.550	0.590
	T_g based	FDR	0.046	0.068	0.092	0.141	0.044	0.062	0.081	0.146
		Power	0.437	0.462	0.487	0.524	0.532	0.554	0.571	0.616
	Standard	FDR	0.022	0.042	0.063	0.092	0.017	0.032	0.044	0.078
		Power	0.394	0.432	0.457	0.487	0.481	0.514	0.532	0.568

A.2 Real Data Analysis

A.2.1 Detail Filtering Step

First, the values are fitted for the i^{th} observation from the null model, $\hat{y}_{i,\text{null}}$, and alternative models $\hat{y}_{i,\text{alt}}$. The residuals of the model fits are then obtained by subtracting the fitted values from the observed values. Let $SS_i^0 = \sum_i (y_i - \hat{y}_{i,\text{null}})^2$, and $SS_i^1 = \sum_i (y_i - \hat{y}_{i,\text{alt}})^2$ be the sum of squares of the residuals obtained from the null model and the alternative model respectively. The statistic for gene i was constructed as

$$F_i = \frac{SS_i^0 - SS_i^1}{SS_i^1}.$$

The null distribution of these statistics are estimated through a data bootstrap method, where residuals from the alternative model are resampled and added back to the null model. This method simulates the case where the patterns of the expression levels are linear [12]. The empirical p-values from the bootstrap step are recorded for all genes. In our study, under both light and dark condition, we select genes that their p-values in all sections are less than 0.1 for further study.

A.2.2 More Results

Table A.4: Summary of Real Data Analysis at 0.05 FDR level

	After filtering	Number of rejection	Peak	Valley	C_4/PS
Dark	10330	1658	1118	541	22 (13 without isoform)
Light	9063	1020	515	511	15 (7 without isoform)

A.3 Expectation for the Statistic T_n

Statistic T_n in (2.5) plays a critical role in the proposed procedure for detecting differential geometric patterns in gene temporal profiles under different biological conditions. As discussed in Section 2.2, T_n mimics the quantity

Table A.5: Top 10 for Light: top 10 significant DE genes under light with valley different on top and peak location difference at bottom.

gene ID	Method	Annotation
GRMZM2G700188	V	putative cinnamyl-alcohol dehydrogenase family protein
GRMZM2G074672	V	vacuolar iron transporter 1.2-like
AC231745.1_FGT003	V	
GRMZM2G113415	V	uncharacterized LOC100275062
GRMZM2G079348	V	
GRMZM2G122943	V	uncharacterized LOC100276511
GRMZM2G091743	V	
GRMZM2G010460	V	putative ubiquitin-conjugating enzyme E2 25
GRMZM2G099745	V	RPM1-interacting protein 4 (RIN4) family protein
GRMZM2G006468	V	wound responsive protein-like
GRMZM2G404973	P	GATA zinc finger family protein
GRMZM2G376416	P	uncharacterized LOC100501315
GRMZM2G039443	P	uncharacterized LOC100275689
GRMZM2G074604	P	phenylalanine ammonia lyase 3
GRMZM5G839640	P	verprolin
GRMZM5G840909	P	putative cytidine/deoxycytidylate deaminase family protein
GRMZM2G054115	P	uncharacterized LOC100273094
GRMZM2G078143	P	uncharacterized LOC100192461
GRMZM2G134072	P	hypothetical protein ZEAMMB73_Zm00001d051466
GRMZM2G470438	P	DNA topoisomerase 3-alpha

Table A.6: Top 10 for Dark: top 10 significant DE genes under dark with valley different on top and peak location difference at bottom.

gene ID	Method	Annotation
GRMZM2G450233	V	peroxidase 5
GRMZM5G869530	V	
GRMZM2G448001	V	putative WD40-like beta propeller repeat family protein
GRMZM2G174990	V	Putative CRAL/TRIO domain containing, Sec14p-like phosphatidylinositol transfer family protein
GRMZM2G483490	V	
GRMZM5G864815	V	thiamin pyrophosphokinase 1
GRMZM2G078033	V	uncharacterized LOC100191603
GRMZM2G146267	V	prolamin-box binding factor 1
GRMZM2G124365	V	chorismate mutase
GRMZM5G815358	V	phytoene desaturase
GRMZM2G107839	P	Non-specific lipid-transfer protein
GRMZM2G103197	P	uncharacterized LOC100383045
GRMZM2G332976	P	short chain alcohol dehydrogenase 1
GRMZM2G130149	P	Transcription factor MYB48
GRMZM2G101693	P	nudix hydrolase 2
GRMZM2G080168	P	uncharacterized LOC100216597
GRMZM2G131205	P	cinnamoyl CoA reductase 1
GRMZM2G029048	P	phenylalanine ammonia lyase9
GRMZM2G167766	P	Protein COFACTOR ASSEMBLY OF COMPLEX C SUBUNIT B CCB3 chloroplastic
GRMZM2G034855	P	putative receptor-like protein kinase

Table A.7: C4/PS gene under Dark: Top 10 significant DE C_4/PS genes under dark. * means gene shows up in both light and dark condition

gene ID	Method	both DE	Annotation
GRMZM2G178693	P		plasma membrane intrinsic protein
GRMZM2G081843	P		plasma membrane intrinsic protein 1
GRMZM2G129513	P		malate dehydrogenase 6
GRMZM2G047368	P		plasma membrane intrinsic protein 2
GRMZM2G083841	P		phosphoenolpyruvate carboxylase 1
GRMZM2G155253	P		Fructose-bisphosphate aldolase
GRMZM2G083016	P		metacaspase type II
GRMZM2G040933	P	*	plastidic general dicarboxylate transporter
GRMZM2G086258	P	*	plastidic general dicarboxylate transporter
GRMZM2G081192	P		plasma membrane intrinsic protein 2

$$(k_1+k_2)^{-1}\left[\sum_{i=1}^{k_1}\min_{1\le j\le k_2}\{\mu_i^{(1)}-\mu_j^{(2)}\}^2+\sum_{j=1}^{k_2}\min_{1\le i\le k_1}\{\mu_j^{(2)}-\mu_i^{(1)}\}^2\right],$$

which measures the discrepancy between two arrays $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ whose dimensions k_1 and k_2 are not necessarily the same. Given the nonlinearity in T_n, it is necessary to establish its connections with $\sum_{1\le i\le k_1}\min_{1\le j\le k_2}\{\mu_i^{(1)}-\mu_j^{(2)}\}^2+\sum_{1\le j\le k_2}\min_{1\le i\le k_1}\{\mu_j^{(2)}-\mu_i^{(1)}\}^2$. By exploring the lower and upper bounds of $\mathbb{E}(T_n)$ under the Gaussian sequence model, this appendix achieves that goal and provides justifications for employing T_n to detect differential geometric patterns in gene temporal profiles.

As an ideal model that carries most of the insight of nonparametric inference, the Gaussian sequence model ([155]) has received a vast attentions in literature. Hereafter, we consider the finite version of the Gaussian sequence model

$$\theta_{1,i}=\mu_i^{(1)}+\epsilon_{1,i}, \tag{A.1}$$

$$\theta_{2,j}=\mu_j^{(2)}+\epsilon_{2,j}, \tag{A.2}$$

where $\epsilon_{1,i}$'s and $\epsilon_{2,j}$'s are i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables for $i=1,\ldots,k_1$ and $j=1,\ldots,k_2$, k_1,k_2 are finite integers and not necessarily the same. For the convenience of expositions, we let $\sigma^2=1$. Based on (A.1) and (A.2), we consider $Q_1=\sum_{i=1}^{k_1}\min_{1\le j\le k_2}\{\theta_{1,i}-\theta_{2,j}\}^2$ and $Q_2=\sum_{j=1}^{k_2}\min_{1\le i\le k_1}\{\theta_{2,j}-\theta_{1,i}\}^2$, so that $T(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(k_1+k_2)^{-1}(Q_1+Q_2)$, where $\boldsymbol{\theta}_1=(\theta_{1,1},\ldots,\theta_{1,k_1})^T$ and $\boldsymbol{\theta}_2=(\theta_{2,1},\ldots,\theta_{2,k_2})^T$, is analogous to T_n.

Lower bound

First, we study the lower bound of $\mathbb{E}\{T(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\}$. For each fixed $i\in[k_1]$, denote $\xi_j=\theta_{1,i}-\theta_{2,j}$ with $\mathbb{E}(\xi_j)=\mu_i^{(1)}-\mu_j^{(2)}:=\nu_j$ and $\text{Var}(\xi_j)=2$. Let $A_j(t)=\{|\xi_j|>t\}$, then $A(t):=\{\min_j|\xi_j|>t\}=\bigcap_{1\le j\le k_2}A_j(t)$ and $\mathbb{P}\{A(t)\}\ge 1-\sum_{j=1}^{k_2}\mathbb{P}\{A_j(t)^c\}$ by the Bonferroni's inequality. Therefore, by the triangle inequality, for some $A>0$

$$\begin{aligned}\mathbb{E}\left\{\min_{1\le j\le k_2}|\xi_j|^2\right\} &= \int_0^\infty \mathbb{P}\left\{\min_{1\le j\le k_2}|\xi_j|^2 > t\right\} dt \\ &\ge \int_0^A \mathbb{P}\left\{\min_{1\le j\le k_2}\left(|\xi_j-\nu_j|-|\nu_j|\right)^2 > t\right\} dt \\ &\ge \int_0^A 1 - \sum_{1\le j\le k_2}\mathbb{P}\left\{\left(|\xi_j-\nu_j|-|\nu_j|\right)^2 < t\right\} dt.\end{aligned} \tag{A.3}$$

In particular, we first choose $A \le \min_{1\le j\le k_2}\nu_j^2$. By the basic properties of the standard normal distribution

$$\begin{aligned}&\mathbb{P}\left\{\left(|\xi_j-\nu_j|-|\nu_j|\right)^2 < t\right\} \\ =&\mathbb{P}\left\{|\xi_j-\nu_j| < |\nu_j|+\sqrt{t}\right\} - \mathbb{P}\left\{|\xi_j-\nu_j| < |\nu_j|-\sqrt{t}\right\} \\ \le&\sqrt{\frac{1}{\pi}}(|\nu_j|+\sqrt{t}) - \sqrt{\frac{1}{\pi}}(|\nu_j|-\sqrt{t})e^{-(|\nu_j|-\sqrt{t})^2/4}\end{aligned}$$

Hence, the last integral in (A.3) can be bounded from below by

$$\begin{aligned}&\int_0^A 1 - \sqrt{\frac{1}{\pi}}\sum_{j=1}^{k_2}(|\nu_j|+\sqrt{t}) + \sqrt{\frac{1}{\pi}}\sum_{j=1}^{k_2}\left\{(|\nu_j|-\sqrt{t})e^{-(|\nu_j|-\sqrt{t})^2/4}\right\} dt \\ \ge&\int_0^A \left\{1 - \frac{\sum_{j=1}^{k_2}|\nu_j|}{\sqrt{\pi}} - \frac{k_2\sqrt{t}}{\sqrt{\pi}}\right\} dt + \sqrt{\frac{1}{\pi}}\sum_{j=1}^{k_2}\int_0^A \left\{(|\nu_j|-\sqrt{t})e^{-(|\nu_j|-\sqrt{t})^2/4}\right\} dt \\ =&\mathbb{I}_1(A)+\mathbb{I}_2(A).\end{aligned}$$

For any $A \le \min_{1\le j\le k_2}\nu_j^2$,

$$\mathbb{I}_1(A) \ge A - \frac{\sum_{j=1}^{k_2}|\nu_j|}{\sqrt{\pi}}A - \frac{2k_2}{3\sqrt{\pi}}A^{3/2}.$$

Choose $A_0 < \min_{1\le j\le k_2}\nu_j^2$ small such that $\min_{1\le j\le k_2}\int_{(|\nu_j|-\sqrt{A_0})/\sqrt{2}}^{|\nu_j|/\sqrt{2}} ue^{-u^2/2}du \ge c_1$ for some constant $c_1 > 0$, and it yields

$$\mathbb{I}(A_0) \ge \frac{4c_1}{\sqrt{\pi}}\sum_{j=1}^{k_2}|\nu_j| - 4.$$

Let $A = c_2(k_2/\sqrt{\pi})^{-2}\min_{1\le j\le k_2}\nu_j^2 \vee A_0$ for sufficiently small constant $c_2 < 1$,

$$\begin{aligned}\mathbb{E}\left\{\min_{1\le j\le k_2}|\xi_j|^2\right\} &\ge A - \frac{2k_2}{3\sqrt{\pi}}A^{3/2} + \frac{(4c_1 - A)k_2}{\sqrt{\pi}}\min_{1\le j\le k_2}|\nu_j| - 4\\ &\ge C_1\min_{1\le j\le k_2}\nu_j^2 + C_2\end{aligned}$$

where C_1, C_2 are constants depending on c_1, c_2 only.

Similar results can be derived for $\mathbb{E}\{\min_{1\le i\le k_1}|\theta_{2,j} - \theta_{1,i}|^2\}$ for fixed $j \in [k_2]$. Therefore, $\mathbb{E}(Q_1+Q_2) = \sum_{i=1}^{k_1}\mathbb{E}\{\min_{1\le j\le k_2}\xi_{j,(i)}^2\}+\sum_{j=1}^{k_2}\mathbb{E}\{\min_{1\le i\le k_1}\xi_{i,(j)}^2\} \ge C_1'[\sum_{i=1}^{k_1}\min_{1\le j\le k_2}\{\mu_i^{(1)} - \mu_j^{(2)}\}^2 + \sum_{j=1}^{k_2}\min_{1\le i\le k_1}\{\mu_j^{(2)} - \mu_i^{(1)}\}^2] + C_2'$ for some constants C_1', C_2' depending on k_1, k_2 and c_1, c_2 above. For sufficiently large distinctions between $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$, the lower bound of the expectation for $T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is therefore primarily driven by $\sum_{i=1}^{k_1}\min_{1\le j\le k_2}\{\mu_i^{(1)} - \mu_j^{(2)}\}^2 + \sum_{j=1}^{k_2}\min_{1\le i\le k_1}\{\mu_j^{(2)} - \mu_i^{(1)}\}^2$. Thus, as discussed and motivated in Section 2.2, it implies that T_n is capable to capture the distinctions between arrays $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ whenever the discrepancy is reasonably large.

Upper bound

Employing the same notation used for deriving the lower bound of $\mathbb{E}\{T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}$, the random variable ξ_j has mean ν_j, variance 2, and $\text{Cov}(\xi_j, \xi_{j'}) = 1$ for each $j' \ne j$ given the models (A.1) and (A.2). Denote $B(t) = \{\xi_1 > \sqrt{t}, \xi_2 < -\sqrt{t}, \ldots, \xi_{k_2} > \sqrt{t}\}$. Based on the directions of inequalities and the sign of $\sqrt{t}$, there are $\binom{k_2}{2}$ cases in total. That is, we have $A_\ell(t)$ where $\ell = 1, \ldots, \binom{k_2}{2}$. Consider a particular ℓ,

$$\begin{aligned}\mathbb{P}\{A_\ell(t)\} &= \mathbb{P}\{(\xi_1, -\xi_2, \ldots, \xi_{k_2}) > \sqrt{t}\mathbf{1}\}\\ &= \mathbb{P}\{(\xi_1 - \nu_1, -\xi_2 + \nu_2, \ldots, \xi_{k_2} - \nu_{k_2}) > \sqrt{t}\mathbf{1} + (-\nu_1, \nu_2, \ldots, -\nu_{k_2})\}\\ &= \mathbb{P}\left\{\mathbf{Z} > \widetilde{\boldsymbol{\Sigma}}_\ell^{-1/2}[\sqrt{t}\mathbf{1} + (-\nu_1, \nu_2, \ldots, -\nu_{k_2})]^T\right\}\\ &:= \mathbb{P}(\mathbf{Z} > \mathbf{s})\end{aligned}$$

where $\mathbf{1}^T$ is the k_2-dimensional vectors of ones, $\mathbf{Z}$ is the k_2-dimensional standard multivariate normal random vector, $\widetilde{\mathbf{\Sigma}}_\ell$ is the covariance matrix of $\widetilde{\boldsymbol{\xi}} = (\xi_1, -\xi_2, \ldots, \xi_{k_2})^T$, and $\mathbf{s}$ is the corresponding vector of $\sqrt{t}$ and ν_j's. By the Chernoff bound,

$$\begin{aligned}
\mathbb{P}\{A_\ell(t)\} &= \prod_{j=1}^{k_2} \mathbb{P}(Z_j > s_j) \\
&\leq \exp(-||\mathbf{s}||_2/2) \\
&= \exp\left\{-\frac{1}{2}\left[t\mathbf{1}^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\mathbf{1} + 2\sqrt{t}\mathbf{1}\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell + \widetilde{\boldsymbol{\nu}}_\ell^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell\right]\right\}
\end{aligned}$$

where $\widetilde{\boldsymbol{\nu}}_\ell = (-\nu_1, \nu_2, \ldots, -\nu_{k_2})^T$. Denote $a_\ell = \mathbf{1}^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\mathbf{1}$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\min_{1\leq j\leq k_2} |\xi_j|^2\right\} \\
=&\int_0^\infty \mathbb{P}\left(\min_{1\leq j\leq k_2} |\xi_j|^2 > t\right) dt \\
\leq&\binom{k_2}{2} \max_{1\leq\ell\leq\binom{k_2}{2}} \int_0^\infty \exp\{-ta_\ell/2\}\exp\{-\sqrt{t}\mathbf{1}\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell\}\exp\{-\widetilde{\boldsymbol{\nu}}_\ell^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell\}dt \\
&\leq \binom{k_2}{2} \max_{1\leq\ell\leq\binom{k_2}{2}} \int_0^\infty \exp\{-ta/2\}dt\exp\{-\widetilde{\boldsymbol{\nu}}_\ell^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell\} \\
&\leq 2k_2^2 \max_{1\leq\ell\leq\binom{k_2}{2}} a_\ell^{-1}\exp\{-\widetilde{\boldsymbol{\nu}}_\ell^T\widetilde{\mathbf{\Sigma}}_\ell^{-1}\widetilde{\boldsymbol{\nu}}_\ell\} \\
&\leq 2k_2^2 \max_{1\leq\ell\leq\binom{k_2}{2}} a_\ell^{-1}\exp\left\{-\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_\ell^{-1})k_2 \min_{1\leq j\leq k_2}|\nu_j|^2\right\} \\
&\leq D_1 + D_2 \min_{1\leq j\leq k_2}|\nu_j|^2
\end{aligned}$$

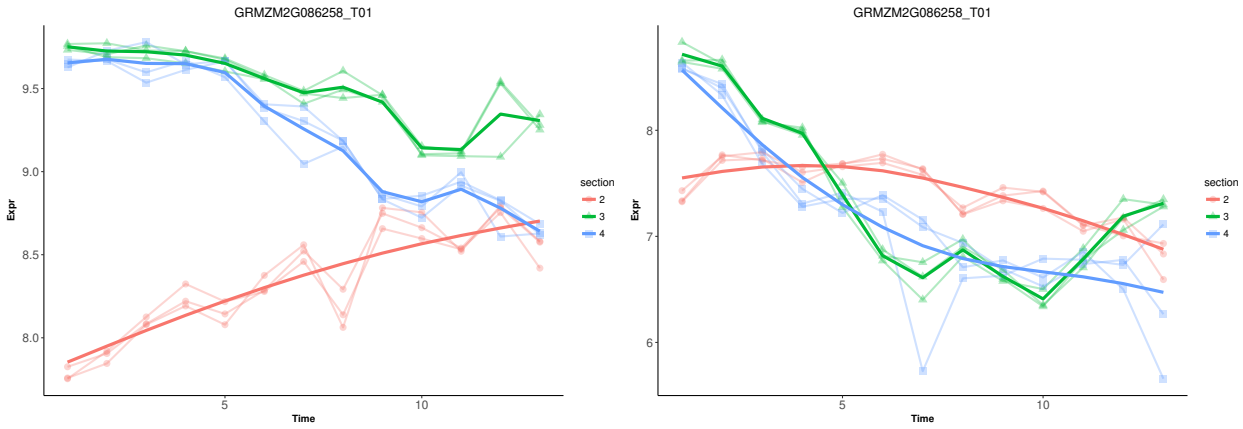
for some constants D_1, D_2 depending on k_1, k_2 and symmetric matrices whose diagonals equal to 2 and off-diagonals equal to ± 1.

Similar results can be derived for $\mathbb{E}\{\min_{1\leq i\leq k_1}|\theta_{2,j} - \theta_{1,i}|^2\}$ for fixed $j \in [k_2]$. Therefore,

$$\mathbb{E}(Q_1 + Q_2) = \sum_{i=1}^{k_1}\mathbb{E}\{\min_{1\leq j\leq k_2}\xi_{j,(i)}^2\} + \sum_{j=1}^{k_2}\mathbb{E}\{\min_{1\leq i\leq k_1}\xi_{i,(j)}^2\}$$

$$\leq D_2' \left[\sum_{i=1}^{k_1} \min_{1 \leq j \leq k_2} \{\mu_i^{(1)} - \mu_j^{(2)}\}^2 + \sum_{j=1}^{k_2} \min_{1 \leq i \leq k_1} \{\mu_j^{(2)} - \mu_i^{(1)}\}^2 \right] + D_1'$$

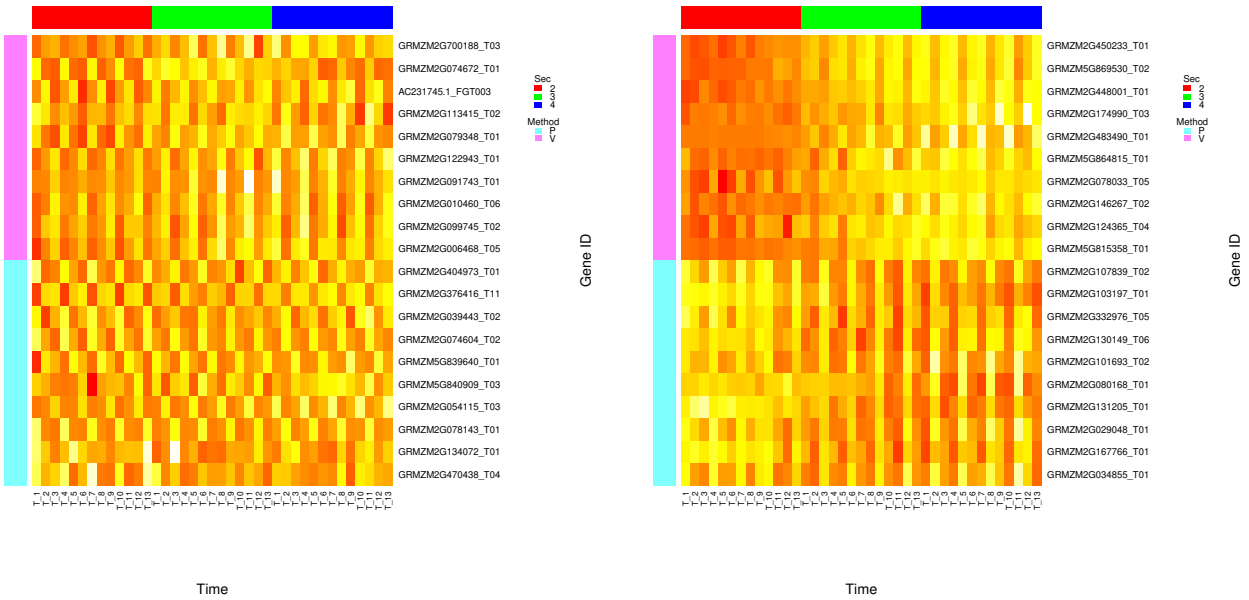
for some constants D_1', D_2'. The upper bound of the expectation for $T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is primarily driven by $\sum_{i=1}^{k_1} \min_{1 \leq j \leq k_2} \{\mu_i^{(1)} - \mu_j^{(2)}\}^2 + \sum_{j=1}^{k_2} \min_{1 \leq i \leq k_1} \{\mu_j^{(2)} - \mu_i^{(1)}\}^2$. In fact, from the above exponential upper bound, the expectation of $T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is controlled from above whenever $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ share a common entry.



(a) Constant light

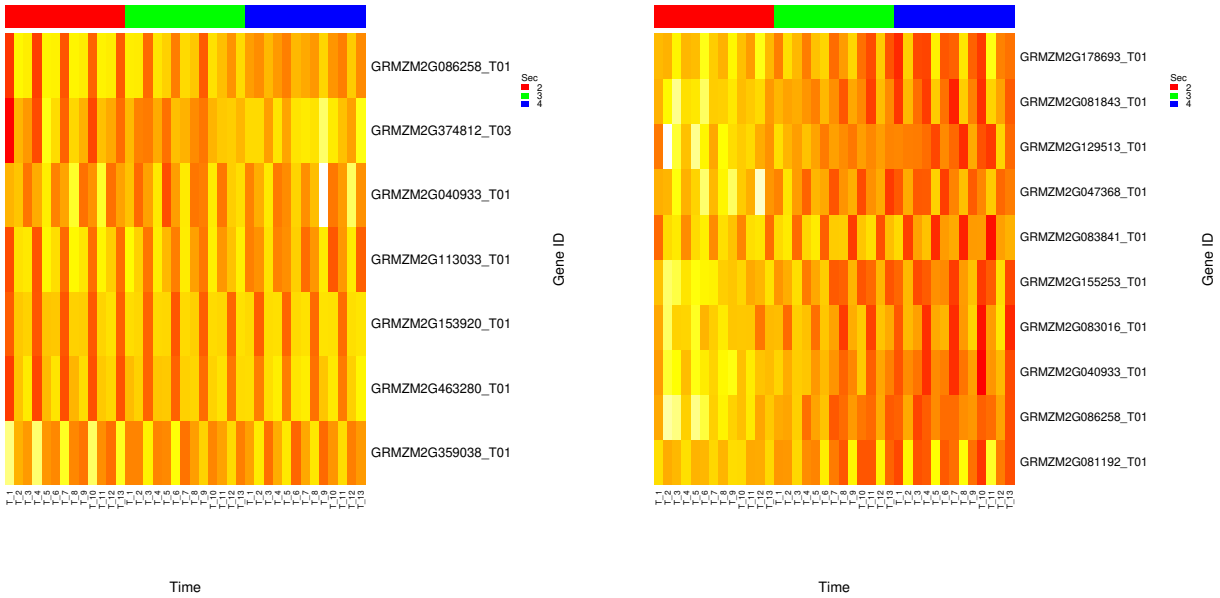
(b) Constant dark

Figure A.1: log RPKM of Gene GRMZM2G086258 with smoothing curve: lighter curves with dots show the actual log scale RPKM data, the darker solid lines show the estimated mean curve using spline model. Color red represents section 2, color green represents section 3 and color blue represents section 4.



(a) Constant light **(b)** Constant dark

Figure A.2: Heatmap of Top 10 genes: Each column is a section and time point combination, and each row is a gene. Heatmap indicates log scale level of gene expression; red, low expression; yellow, high expression. The categorical annotation bars (above heatmap) demonstrate the section label (red, section 2; green, section 3; blue, section 4). The color bar on the left side the method (purple, valley locations are different; light blue, peak locations are different).



(a) Constant light **(b)** Constant dark

Figure A.3: Heatmap of Top C_4 genes: Each column is a section and time point combination, and each row is a gene. Heatmap indicates log scale level of gene expression; red, low expression; yellow, high expression. The categorical annotation bars (above heatmap) demonstrate the section label (red, section 2; green, section 3; blue, section 4).

Appendix B

Supplemental materials for Chapter 3

B.1 Supplementary Materials

The `R` package, `MAPTest`, implements our proposed method and is publicly available at https://github.com/meca7653/MAPTest, where the illustrative numerical example is also included. Users can specify their own estimates of the normalization factors or dispersion parameters.

B.2 Details: A Quasi-Monte Carlo Integration-Assisted Gradient Expectation-Maximization Algorithm for Estimation

In Section 3.2 in the main paper, we remark that the proposed model is estimated by a variant of the expectation-maximization (EM) algorithm. In this section, we detail that estimation procedure, which is a quasi-Monte Carlo-integration assisted gradient EM algorithm.

B.2.1 Estimation

As discussed in Section 3.2 in the main paper, the coefficients of the basis functions, $\boldsymbol{\eta}_{g2}$ and $\boldsymbol{\tau}_g$, as well as η_{g1}, are *de facto* latent variables for the proposed K-component latent Gaussian-Negative Binomial model. As displayed in Section 3.4.1, $\{\mu_1, \sigma_1^2, \boldsymbol{\Psi}, \mathbf{M}\}$ can be viewed as hyper-parameters of the proposed model. For estimation, we introduce a latent vector $\mathbf{Z}_g = (Z_{g1}, \ldots, Z_{gK})$ for gene g, where Z_{gk} is a binary pointer assigning $(\eta_{g1}, \boldsymbol{\eta}_{g2}, \boldsymbol{\tau}_g)$ to the kth component with $k = 1, \ldots, K$. Though our model focuses on $K = 4$, the algorithm is flexible for any finite K. It is common to assume that $\mathbf{Z}_g$'s are independent multinomial random vectors that consist of K categories with probability $\boldsymbol{p} = (p_1, \ldots, p_K)$. Hence, the complete data for the proposed model are $(\mathbf{Y}_g, \eta_{g1}, \boldsymbol{\eta}_{g2}, \boldsymbol{\tau}_g, \mathbf{Z}_g)$, where only $\mathbf{Y}_g$'s are observed and others are latent.

We estimate the non-observables and parameters, $\boldsymbol{\zeta} = \{\mu_1, \sigma_1^2, \text{diag}(\boldsymbol{\Psi}), \text{diag}(\mathbf{M})\}$ and $\boldsymbol{p}$, via an EM algorithm, which is conventionally employed to estimate models with latent variables. Here,

diag$(\mathbf{A})$ denotes the diagonal entries of matrix $\mathbf{A}$. The complete log-likelihood is

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{g=1}^{G} Z_{gk} \log \left\{ F_k(\mathbf{Y_g}; \boldsymbol{\zeta}) \right\} + Z_{gk} \log(p_k) + (1 - Z_{gk}) \log(1 - p_k),$$

by which the conditional expectation of latent variables can be computed. For example,

$$\mathbb{E}\left\{ Z_{gk} | \mathbf{Y}_g, \boldsymbol{\zeta}^{(m)}, \boldsymbol{p}^{(m)} \right\} = \frac{p_k^{(m)} F_k(\mathbf{Y_g}; \boldsymbol{\zeta}^{(m)})}{\sum_\ell p_\ell^{(m)} F_\ell(\mathbf{Y_g}; \boldsymbol{\zeta}^{(m)})}, \tag{B.1}$$

where $\boldsymbol{\zeta}^{(m)} = \{\mu_1^{(m)}, \sigma_1^{2,(m)}, \text{diag}(\boldsymbol{\Psi})^{(m)}, \text{diag}(\mathbf{M})^{(m)}\}$ and $\boldsymbol{p}^{(m)}$ are estimates at step m, and $F_k(\mathbf{Y_g}; \boldsymbol{\zeta}^{(m)})$ denotes the conditional density of $\mathbf{Y}_g$ given parameters from component k.

We then update $\widehat{p}_k$ by

$$\widehat{p}_k = \frac{\sum_g \mathbb{E}\left\{ Z_{gk} | \mathbf{Y}_g, \boldsymbol{\zeta}^{(m)} \right\}}{\sum_k \sum_g \mathbb{E}\left\{ Z_{gk} | \mathbf{Y}_g, \boldsymbol{\zeta}^{(m)} \right\}} \tag{B.2}$$

for each $k = 1, \ldots, K$. This is the so-called **E-step**. Instead of maximizing the weighted log-likelihood

$$\sum_k \sum_g \widehat{Z}_{gk} \log \left\{ \widehat{p}_k F_k(\mathbf{Y_g}; \boldsymbol{\zeta}) \right\} \tag{B.3}$$

in the traditional EM, we employ the gradient EM [156] to improve the computational efficiency. That is, we update $\boldsymbol{\zeta}$ only by moving along the gradient of (B.3) with respect to $\boldsymbol{\zeta}$ for one step whose step size shrinks along update. This is the so-called **GM-step**.

Repeat the **E-step** and **GM-step** until the log-likelihood function changes no more than a small number from the previous iteration, say $G/1000$, which means the improvement is no better than $1/1000$ log-likelihood on average for each gene. The complete algorithm is summarized in Algorithm 1 below.

Initialization

Performance of the EM algorithm is known to depend highly on the initialization. We design an initialization procedure for our proposed model as following.

1. First, for pre-specified basis functions $\mathbf{B}(t)$, we fit a negative binomial regression model for each gene to obtain $\widetilde{\eta}_{g1}$, $\widetilde{\boldsymbol{\eta}}_{g2}$ and $\widetilde{\boldsymbol{\tau}}_g$.

2. Next, we cluster $\widetilde{\eta}_{g1}$'s into 2 groups using method such as K-means. Proportions (p_1, p_2), means (m_1, m_2), and variances (s_1^2, s_2^2) are estimated for each cluster. Without loss of generality, we assume $|m_1| > |m_2|$ and initialize (μ_1, σ_1^2) as $(\widetilde{\mu}_1, \widetilde{\sigma}_1^2) = (m_1, s_1^2)$.

3. Then, we perform a test, such as the likelihood ratio test, on $H_0^g : \boldsymbol{\eta}_{g2} = 0$ to obtain the p-values p_g's. The initialization of $\mathrm{diag}(\mathbf{M})$ is given by diagonal entries of

$$\widetilde{M} = \frac{1}{|\sum_g \mathbb{I}(p_g \le \alpha)|} \sum_{g: p_g \le \alpha} (\widetilde{\boldsymbol{\eta}}_{g2} - \bar{\widetilde{\boldsymbol{\eta}}}_2)'(\widetilde{\boldsymbol{\eta}}_{g2} - \bar{\widetilde{\boldsymbol{\eta}}}_2)'$$

   for cutoff α and $\bar{\widetilde{\boldsymbol{\eta}}}_2 = |\sum_g \mathbb{I}(p_g \le \alpha)|^{-1} \sum_{g: p_g \le \alpha} \widetilde{\boldsymbol{\eta}}_{g2}$. We initialize $\mathrm{diag}(\boldsymbol{\Psi})$ similarly.

4. Last, we initialize $\widetilde{p}_0 = p_2 G^{-1} \sum_g \mathbb{I}\{p_g > \alpha\}$, $\widetilde{p}_1 = p_2 G^{-1} \sum_g \mathbb{I}\{p_g \le \alpha\}$, $\widetilde{p}_3 = p_1 G^{-1} \sum_g \mathbb{I}\{p_g > \alpha\}$ and $\widetilde{p}_4 = p_1 G^{-1} \sum_g \mathbb{I}\{p_g \le \alpha\}$.

Dispersion parameter estimation

We employ the moments estimator for dispersion parameter ϕ_g. For gene g in treatment group i at time point t, $s_{gi}^2(t) = m_{gi}(t) + \phi_g m_{gi}^2(t)$, where $s_{gi}^2(t) = \sum_{j=1}^r \{Y_{gij}(t) - m_{gi}(t)\}^2/(r-1)$ is the sample variance and $m_{gi}(t) = r^{-1} \sum_{j=1}^r Y_{gij}(t)$ is the sample mean. Here, r is the number of replicates. Then, we estimate ϕ_g by

$$\widehat{\phi}_g = \sum_i \sum_t \frac{s_{gi}^2(t) - m_{gi}(t)}{\sum_i \sum_t m_{gi}^2(t)}. \tag{B.4}$$

B.2.2 Quasi-Monte Carlo Approximation

For our model, F_k in (B.1) and (B.3) is an integral without closed form. This brings challenges for both the **E-step** and the **GM-step** as the evaluation of F_k is indispensable. To circumvent this difficulty, we employ the quasi-Monte Carlo method to approximate

Algorithm 1: Estimation and inference of the 4-component latent Gaussian-negative binomial model

Input : Gene matrix: $\{\mathbf{Y}_g\}_{g=1}^G$, nominal FDR level α,
Initial number of QMC nodes: N,
Initial parameters: $\widetilde{\boldsymbol{\zeta}}, \{\widetilde{p}_k\}_{k=1}^{K=4}, \{\widehat{\phi}_g\}_{g=1}^G$ in Section B.2.1,
Hypothesis parameters: $H_0 : \Delta_0$ vs. $H_1 : \Delta_1$.

Output: A list of DE genes corresponding to the hypothesis of interest.

Parameter estimation: while $|\mathcal{L}^{(m)} - \mathcal{L}^{(m+1)}| > \max(1\text{e}{-6}, G/1000)$ **do**

  E-Step: Update p_k's by (B.2).

  GM-Step: $\boldsymbol{\zeta}^{(m+1)} = \boldsymbol{\zeta}^{(m)} + \Delta^{(m)} \partial\mathcal{L}/\partial\boldsymbol{\zeta}|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}^{(m)}}$, where $\mathcal{L}$ is in (B.3), $\Delta^{(m)}$ is the step size chosen to be $1/2^m$ and $N^{(m)} = N/2^m$.

end

Determine λ_α **or** $\Gamma(\alpha)$**: for** *all genes* **do**

  (1) Calculate $\widehat{\delta}_{\text{MAP}}(\mathbf{Y}_g)$ in (4) in the main paper with $\widehat{\pi}(\eta_1, \boldsymbol{\eta}_2, \boldsymbol{\tau}|\widehat{\boldsymbol{\zeta}}, \{\widehat{p}_k\}_{k=1}^{K=4})$ for each $g = 1, \ldots, G$.

  (2) For each $\lambda > 0$, compute

$$\widehat{\text{FDR}}_\Gamma = \frac{\sum_g \widehat{\delta}_{\text{MAP}}(\mathbf{Y}_g)\mathbb{I}\{\mathbf{Y}_g \in \Gamma\}}{\sum_g \mathbb{I}\{\mathbf{Y}_g \in \Gamma\}},$$

  where $\Gamma = \{\mathbf{Y}_g : \widehat{\delta}_{\text{MAP}}(\mathbf{Y}_g) \leq \lambda\}$.

  (3) Choose λ such that $\widehat{\text{FDR}}_\Gamma \leq \alpha$, and denote by $\widehat{\lambda}_\alpha$.

end

Test Construction: for $g = 1, 2, \ldots, G$ **do**

  if $\widehat{\delta}_{MAP}(\mathbf{Y}_g)$ *is smaller than* $\widehat{\lambda}_\alpha$*, i.e.* $\mathbf{Y}_g \in \widehat{\Gamma}(\alpha)$ **then**

    Reject H_0^g;

  else

    Fail to reject H_0^g;

  end

end

$$F(y; \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \int_{\mathbb{R}^m} f(y|\mathbf{x}'\boldsymbol{\beta}, \phi)\varphi(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Lambda})d\boldsymbol{\beta}$$

where $f(y|\mathbf{x}'\boldsymbol{\beta}, \phi)$ is the pdf of a negative binomial distribution and φ is the density of multivariate normal. Among a large number of numerical integration methods in the literature, Monte Carlo method is straightforward but time consuming when the dimension of parameters increases. Compared to the traditional Monte Carlo method, quasi-Monte Carlo (QMC) method generates nodes based on a low discrepancy sequence. As suggested by [157], let

$$D_n * (x_1, \ldots, x_n) = \sup_{a \in [0,1)^d} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{0 \leq x_i < a\} - |[0, a)| \right|,$$

then sequences (x_i) in $[0, 1)^d$ with $D_n * (x_1, \ldots, x_n) = O\left(\log(n)^d/n\right)$ are called low discrepancy sequences. Examples include the van der Corput sequence and the Halton sequence [158].

Following [157], we generate a low discrepancy sequence in $[0, 1)^{2q+1}$, where q is number of basis functions. Then, convert this low discrepancy sequence by inverse cumulative distribution function of standard normal, followed by necessary shift and projection, to generate a quasi-normal sequence $\boldsymbol{\beta}_i = (\eta_{1i}, \boldsymbol{\eta}_{2i}, \boldsymbol{\tau}_i)$ for $i = 1, \ldots, N$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}$. The quasi-Monte Carlo integration is then given by

$$F(y; \boldsymbol{\mu}, \boldsymbol{\Lambda}) \approx \widetilde{F}(y; \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{1}{N} \sum_{i=1}^{N} f(y|\mathbf{x}'\boldsymbol{\beta}_i, \phi).$$

In practice, the integrand might be small at some QMC nodes, which incurs numerical instabilities. Hence, instead of sampling nodes with equal weights $1/N$, we further employ the idea of "sampling with proportion to size" to avoid sampling QMC nodes with overwhelmingly small density values [159] using the history information of the updates. That is, within the EM algorithm, let

$$\widetilde{F}\left(y; \boldsymbol{\mu}^{(m+1)}, \boldsymbol{\Lambda}^{(m+1)}\right) = \sum_{i=1}^{N^{(m)}} w_i f(y|\mathbf{x}'\boldsymbol{\beta}_i^{(m+1)}, \phi),$$

where $\boldsymbol{\beta}_i^{(m+1)}$ are quasi-normal sequence with mean $\boldsymbol{\mu}^{(m+1)}$ and $\boldsymbol{\Lambda}^{(m+1)}$ at step m and

$$w_i = \frac{N^{(m+1)} f(y|\mathbf{x}'\boldsymbol{\beta}_i^{(m)}, \phi)}{\sum_{j=1}^{N^{(m)}} f(y|\mathbf{x}'\boldsymbol{\beta}_j^{(m)}, \phi)}.$$

B.3 More Simulation Results

Figures S.1–S.7 and Tables S.1–S.7 display additional simulation results as discussed in the main paper.

B.3.1 Details on Basis Functions

The basis functions for the traditional Gaussian kernel [54], denoted by GA_2 and GA_3 in Section 4 in the main paper are $\{b_k(t)\}_{k=1}^q$, $q = 2, 3$, where

$$b_k(t) = h_k \exp\{-(c-a)t^2\} H_k(\sqrt{2c}t),$$

for which $H_k(x)$ is the kth order Hermite polynomial, $c = (a^2 + 2ab)^{1/2}$, $h_k^{-2} = \sqrt{a/c}2^k k!$, and we set $a = 1/4$ and $b = 2$ in simulations.

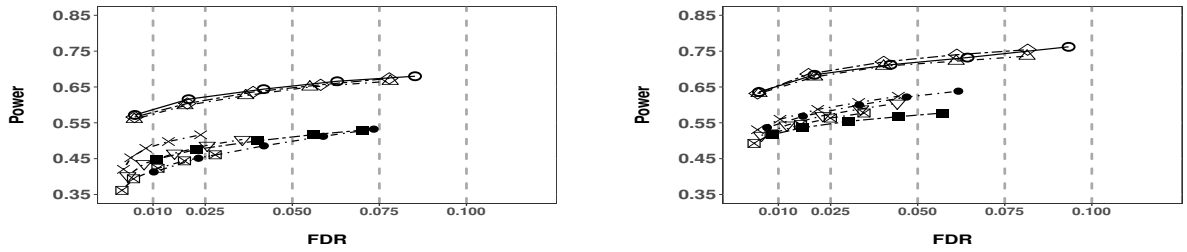
B.3.2 Additional Results on Settings A and B

In Figures S.1–S.6, additional simulation results for testing overall temporal DE genes that is alternative to Δ_0^{DE} are displayed.

- Figures S.1 and S.2 are for Settings A and B, respectively, where $T = 6$.
- Figures S.3 and S.4 are for $\mu_1 = 2$ and $\mu_1 = 4$, respectively; $T = 10$ and both Settings A and B are included for each figure.
- Figures S.5 and S.6 are for $\mu_1 = 2$ and $\mu_1 = 4$, respectively; $T = 6$ and both Settings A and B are included for each figure.

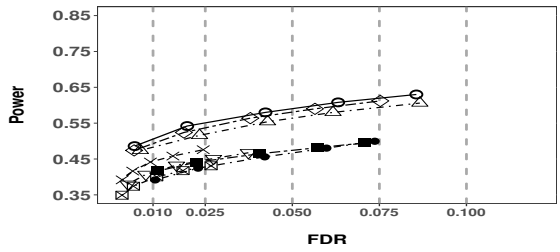
In addition, Figures S.3–S.6 also display results of our method with different basis functions for comparison (blue dot lines in each Figure). We observe that our method is reasonably robust with respect to the mis-specification of the basis function. Similar observations are obtained from Tables 1–2 in the main paper and Tables S.1-S.4 below. This is because that the mean dynamic $\lambda(t)$ in (5) in the main paper is smoothly modeled conditional on a latent zero-mean Gaussian process, whose finite approximation, given the smooth covariance structure, is less sensitive to the choice and the number of eigenfunctions in practice due to Mercer's Theorem.

For competing methods in simulations, to perform the DE analysis on time-course RNA-seq data, the F-statistics are used for `edgeR`, the likelihood ratio statistics are used for `DESeq2`,

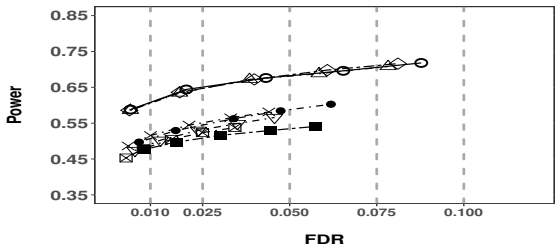


(a) $\sigma_1^2 = 1, r = 3, \mu = 2$

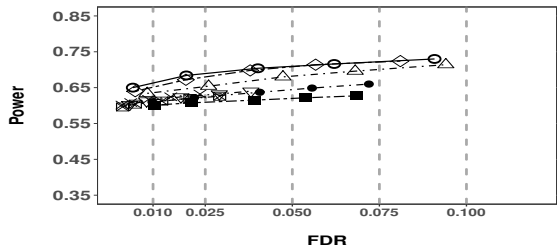
(b) $\sigma_1^2 = 1, r = 6, \mu = 2$



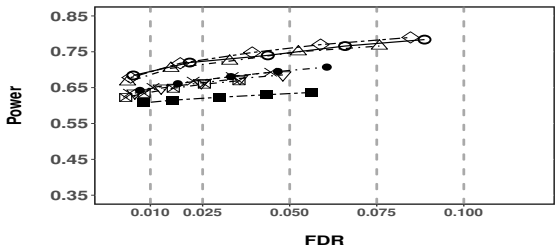
(c) $\sigma_1^2 = 2, r = 3, \mu = 2$



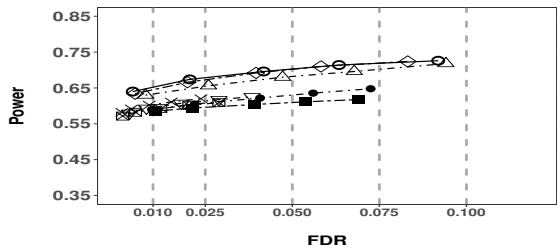
(d) $\sigma_1^2 = 2, r = 6, \mu = 2$



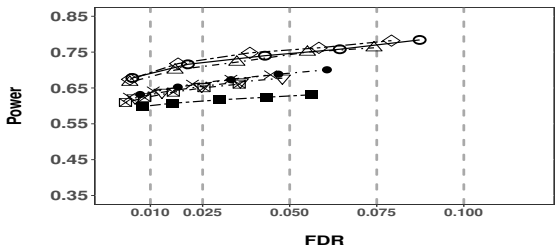
(e) $\sigma_1^2 = 1, r = 3, \mu = 4$



(f) $\sigma_1^2 = 1, r = 6, \mu = 4$



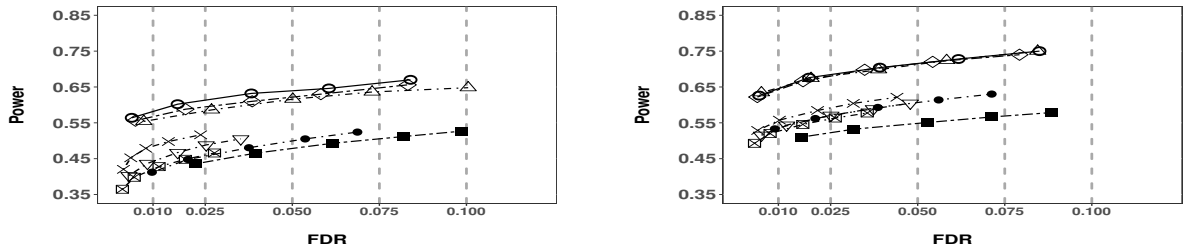
(g) $\sigma_1^2 = 2, r = 3, \mu = 4$



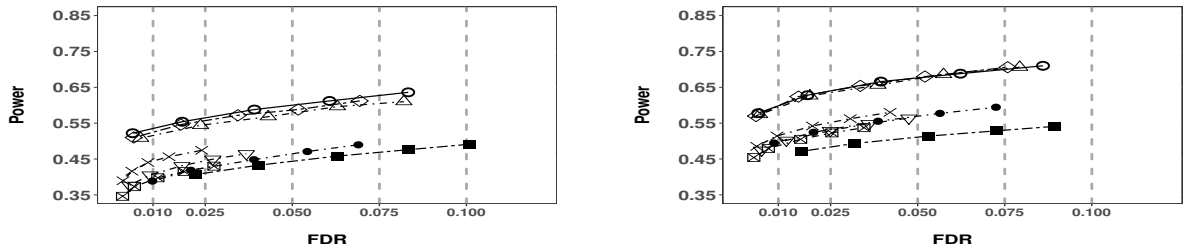
(h) $\sigma_1^2 = 2, r = 6, \mu = 4$

Figure S.1: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$)and `DESeq2` ($\times$) for Setting A. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 6$ and based on 100 replications.

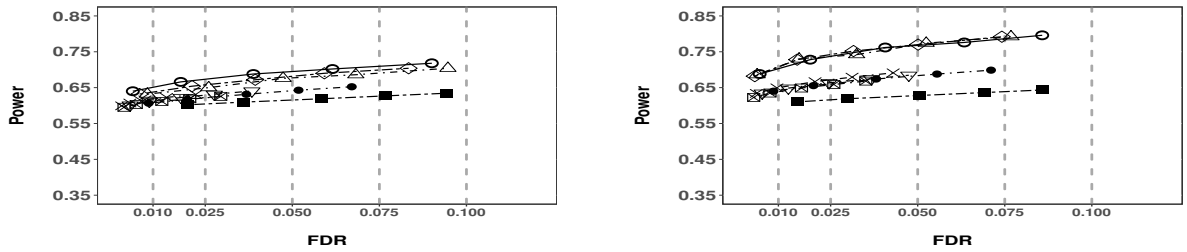
`magSigPro-GLM`, and `ImpulseDE2`, and the empirical Bayes moderated F-statistics are used for `splineTC`.



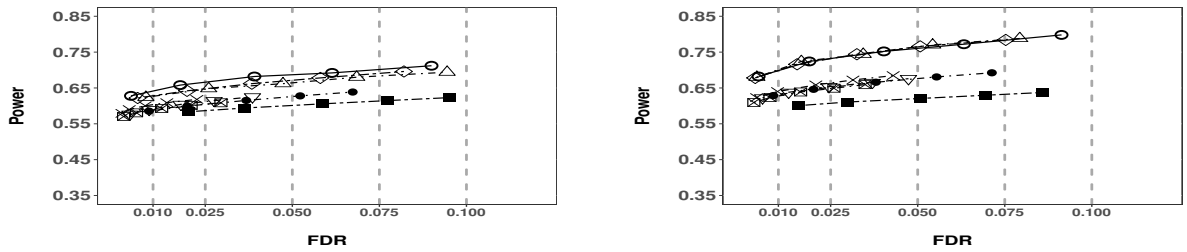
(a) $\sigma_1^2 = 1, r = 3, \mu = 2$ **(b)** $\sigma_1^2 = 1, r = 6, \mu = 2$



(c) $\sigma_1^2 = 2, r = 3, \mu = 2$ **(d)** $\sigma_1^2 = 2, r = 6, \mu = 2$



(e) $\sigma_1^2 = 1, r = 3, \mu = 4$ **(f)** $\sigma_1^2 = 1, r = 6, \mu = 4$



(g) $\sigma_1^2 = 2, r = 3, \mu = 4$ **(h)** $\sigma_1^2 = 2, r = 6, \mu = 4$

Figure S.2: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$)and `DESeq2` ($\times$) for Setting B. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 6$ and based on 100 replications.

From the plots, the control of empirical FDRs by `magSigPro-GLM` is better when the underlying mean dynamic is close to a quadratic function (such as PL_2 in Setting A) rather than a more sophisticated form (such as PL_3 in Setting B). This is because that `magSigPro-GLM` models the

mean temporal dynamic using a less flexible quadratic regression and does not utilize the protocol of borrowing information across genes. On the other hand, for small T, the difference between PL_2 and PL_3 is not significant based on a small number of realizations of $\lambda(t)$ at T time points. This explains the reasonable control on empirical FDRs by `magSigPro-GLM` for relative small T in simulations.

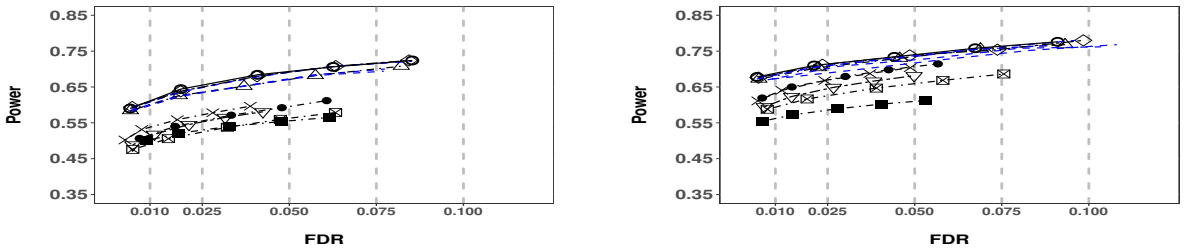
In Tables S.1 and S.2, we present additional results on comparing the proposed method using different basis functions with `edgeR` and `DESeq2` for testing DE genes with relative mean shift when $\mu_1 = 2$ and 4, respectively, under Setting B. Tables S.3 and S.4 report similar results for testing NPDE genes when $\mu_1 = 2$ and 4, respectively, under Setting B as well.

Table S.1: Comparison of empirical FDRs and powers for testing DE genes with relative mean shift by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting B. In simulations, $\mu_1 = 2$, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

		(T, r, σ_1^2)							
		$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
GA_2	FDR	0.073	0.060	0.053	0.051	0.054	0.055	0.050	0.049
	Power	0.910	0.847	0.927	0.860	0.933	0.927	0.973	0.947
GA_3	FDR	0.077	0.065	0.059	0.054	0.057	0.059	0.050	0.050
	Power	0.903	0.843	0.927	0.853	0.933	0.913	0.973	0.947
FO_2	FDR	0.083	0.073	0.070	0.070	0.054	0.078	0.082	0.085
	Power	0.917	0.853	0.923	0.863	0.933	0.917	0.977	0.950
FO_3	FDR	0.088	0.071	0.066	0.059	0.057	0.070	0.067	0.070
	Power	0.910	0.853	0.930	0.853	0.933	0.917	0.973	0.947
PL_2	FDR	0.054	0.049	0.047	0.044	0.054	0.048	0.046	0.044
	Power	0.910	0.833	0.927	0.860	0.933	0.923	0.973	0.950
Oracle	FDR	0.059	0.052	0.049	0.045	0.058	0.049	0.046	0.042
	Power	0.903	0.840	0.927	0.847	0.933	0.913	0.973	0.947
True	FDR	0.040	0.042	0.041	0.044	0.043	0.044	0.044	0.045
	Power	0.906	0.847	0.927	0.850	0.937	0.917	0.973	0.947
`edgeR`	FDR	0.101	0.102	0.060	0.060	0.065	0.067	0.047	0.047
	Power	0.798	0.741	0.880	0.820	0.867	0.806	0.917	0.864
`DESeq2`	FDR	0.018	0.018	0.033	0.032	0.029	0.029	0.036	0.036
	Power	0.796	0.726	0.879	0.812	0.865	0.795	0.916	0.860

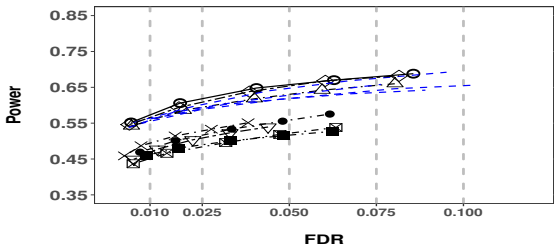
Table S.2: Comparison of empirical FDRs and powers for testing DE genes with relative mean shift by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting B. In simulations, $\mu_1 = 4$, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

		(T, r, σ_1^2)							
		$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
GA_2	FDR	0.068	0.069	0.018	0.050	0.047	0.053	0.006	0.050
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
GA_3	FDR	0.083	0.080	0.053	0.056	0.064	0.062	0.010	0.048
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
FO_2	FDR	0.055	0.059	0.027	0.054	0.047	0.061	0.034	0.060
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
FO_3	FDR	0.078	0.078	0.052	0.056	0.054	0.061	0.013	0.057
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
PL_2	FDR	0.063	0.060	0.018	0.048	0.038	0.053	0.013	0.047
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
Oracle	FDR	0.071	0.066	0.052	0.050	0.059	0.059	0.015	0.050
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
True	FDR	0.019	0.043	0.007	0.049	0.010	0.049	0.005	0.043
	Power	0.997	0.987	1.000	0.993	1.000	0.993	1.000	0.993
`edgeR`	FDR	0.094	0.095	0.058	0.058	0.065	0.065	0.048	0.048
	Power	0.996	0.976	0.999	0.988	0.999	0.986	0.999	0.992
`DESeq2`	FDR	0.017	0.017	0.032	0.032	0.030	0.030	0.036	0.036
	Power	0.996	0.974	0.999	0.987	0.999	0.986	0.999	0.992

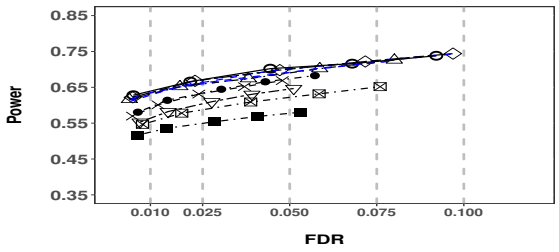


(a) $\sigma_1^2 = 1, r = 3, q = 2$ (Setting A)

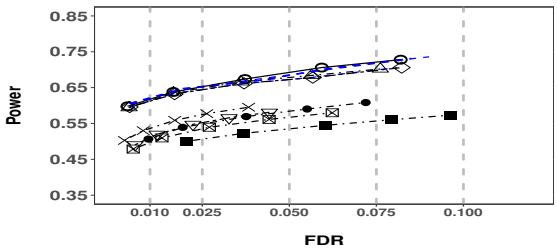
(b) $\sigma_1^2 = 1, r = 6, q = 2$ (Setting A)



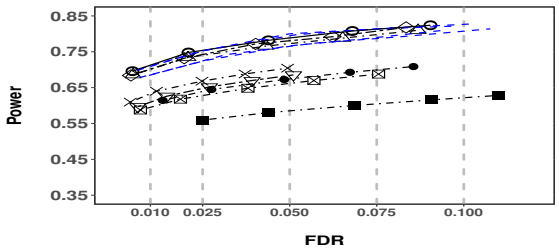
(c) $\sigma_1^2 = 2, r = 3, q = 2$ (Setting A)



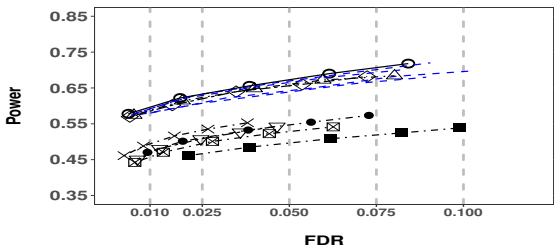
(d) $\sigma_1^2 = 2, r = 6, q = 2$ (Setting A)



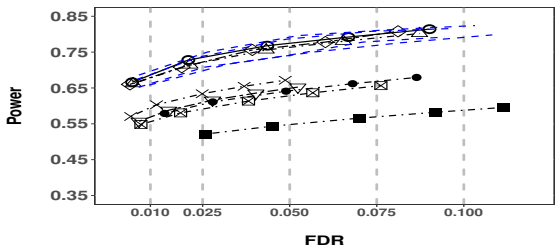
(e) $\sigma_1^2 = 1, r = 3, q = 3$ (Setting B)



(f) $\sigma_1^2 = 1, r = 6, q = 3$ (Setting B)

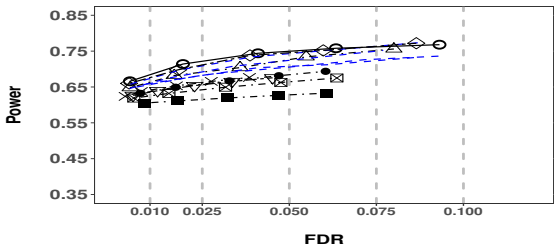


(g) $\sigma_1^2 = 2, r = 3, q = 3$ (Setting B)

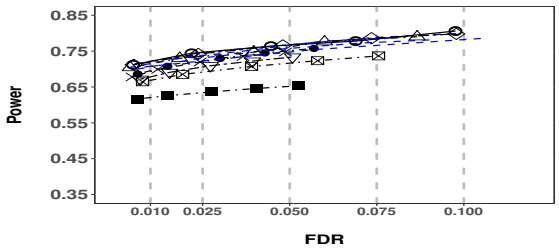


(h) $\sigma_1^2 = 2, r = 6, q = 3$ (Setting B)

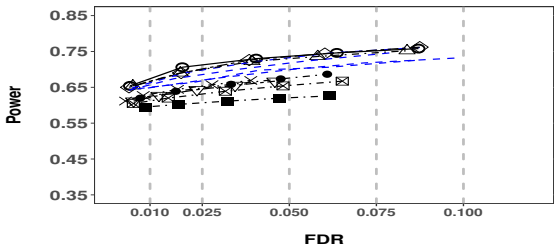
Figure S.3: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$), and `DESeq2` ($\times$). The blue dot lines ($\cdots\cdots$) denote results for other bases described in Section 5 in the main paper. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 10$, $\mu_1 = 2$ and based on 100 replications.



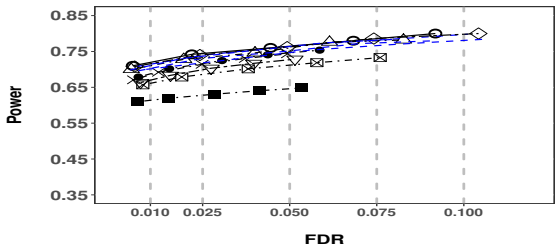
(a) $\sigma_1^2 = 1, r = 3, q = 2$ (Setting A)



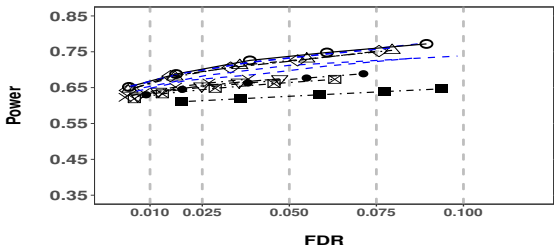
(b) $\sigma_1^2 = 1, r = 6, q = 2$ (Setting A)



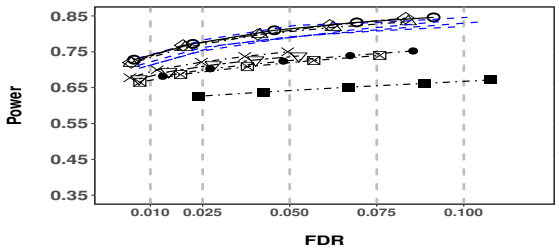
(c) $\sigma_1^2 = 2, r = 3, q = 2$ (Setting A)



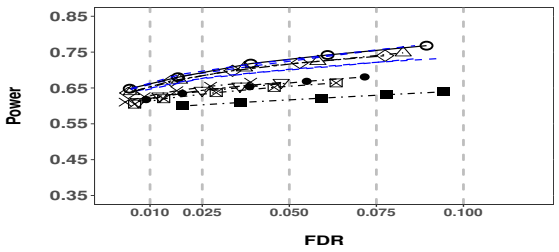
(d) $\sigma_1^2 = 2, r = 6, q = 2$ (Setting A)



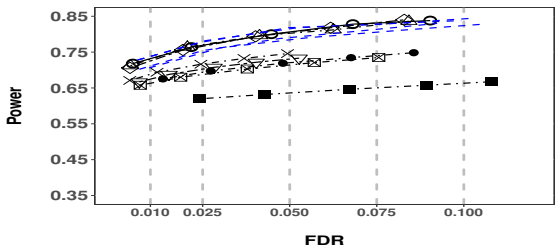
(e) $\sigma_1^2 = 1, r = 3, q = 3$ (Setting B)



(f) $\sigma_1^2 = 1, r = 6, q = 3$ (Setting B)

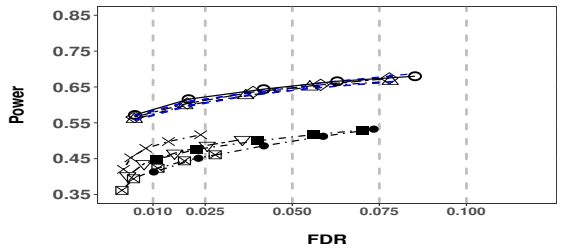


(g) $\sigma_1^2 = 2, r = 3, q = 3$ (Setting B)

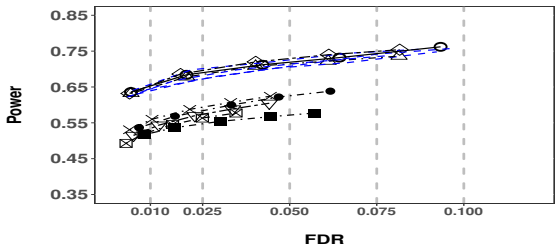


(h) $\sigma_1^2 = 2, r = 6, q = 3$ (Setting B)

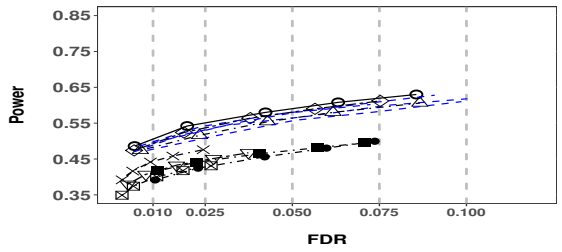
Figure S.4: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$), and `DESeq2` ($\times$). The blue dot lines ($\cdots\cdots$) denote results for other bases described in Section 5 in the main paper. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 10$, $\mu_1 = 4$ and based on 100 replications.



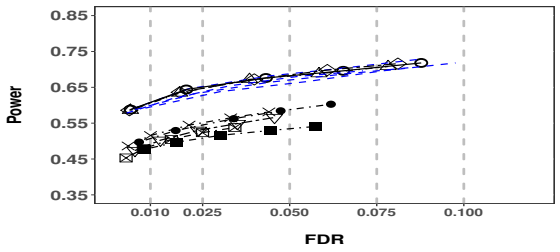
(a) $\sigma_1^2 = 1, r = 3, q = 2$ (Setting A)



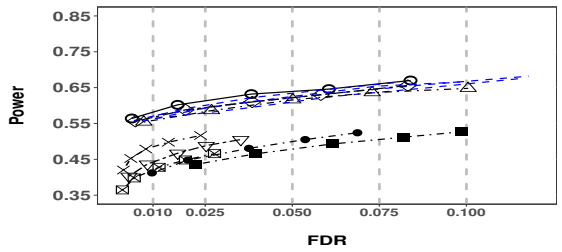
(b) $\sigma_1^2 = 1, r = 6, q = 2$ (Setting A)



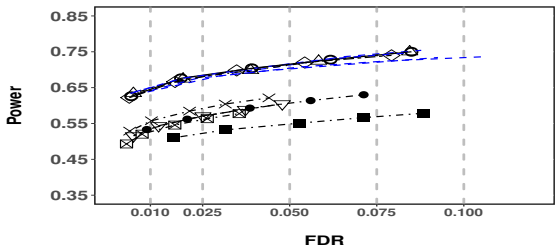
(c) $\sigma_1^2 = 2, r = 3, q = 2$ (Setting A)



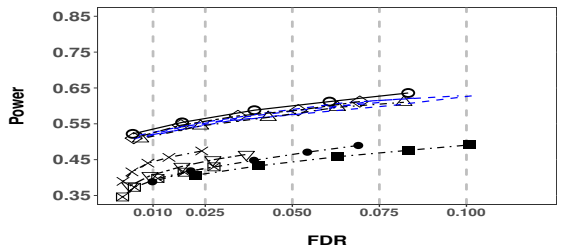
(d) $\sigma_1^2 = 2, r = 6, q = 2$ (Setting A)



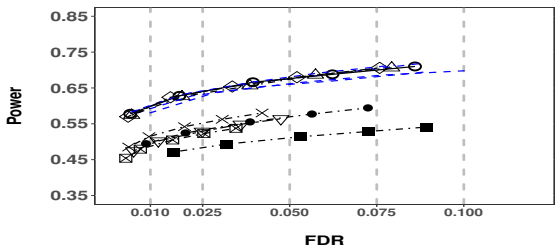
(e) $\sigma_1^2 = 1, r = 3, q = 3$ (Setting B)



(f) $\sigma_1^2 = 1, r = 6, q = 3$ (Setting B)

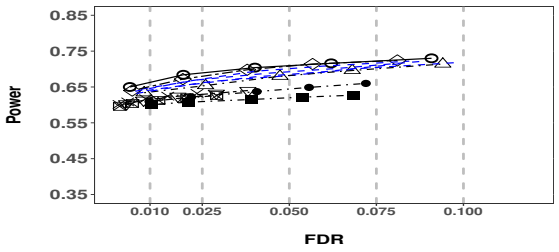


(g) $\sigma_1^2 = 2, r = 3, q = 3$ (Setting B)

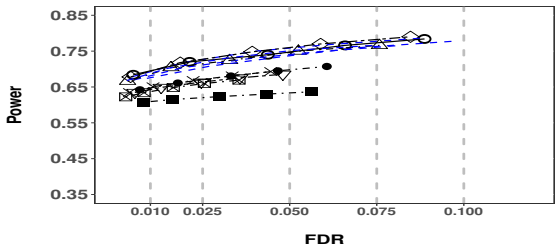


(h) $\sigma_1^2 = 2, r = 6, q = 3$ (Setting B)

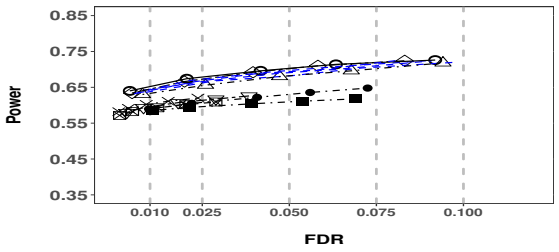
Figure S.5: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$), and `DESeq2` ($\times$). The blue dot lines ($\cdots\cdots$) denote results for other bases described in Section 5 in the main paper. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 6$, $\mu_1 = 2$ and based on 100 replications.



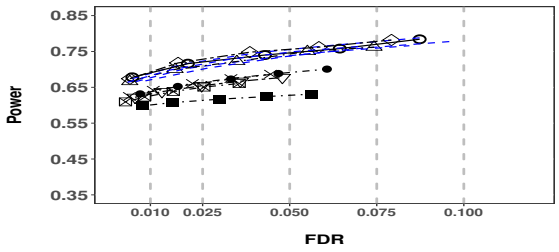
(a) $\sigma_1^2 = 1, r = 3, q = 2$ (Setting A)



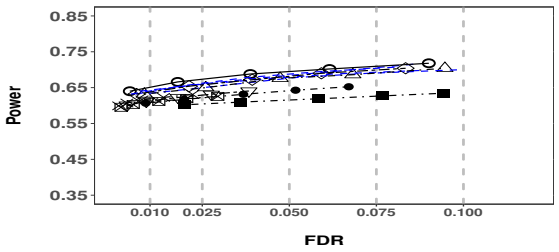
(b) $\sigma_1^2 = 1, r = 6, q = 2$ (Setting A)



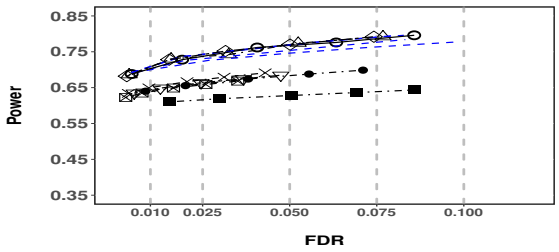
(c) $\sigma_1^2 = 2, r = 3, q = 2$ (Setting A)



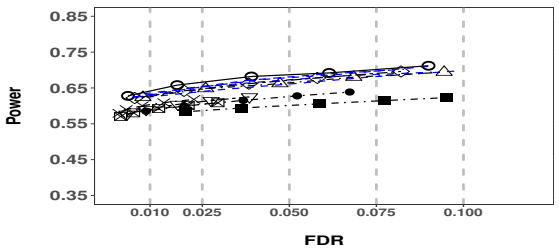
(d) $\sigma_1^2 = 2, r = 6, q = 2$ (Setting A)



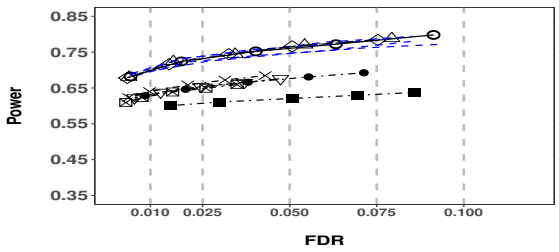
(e) $\sigma_1^2 = 1, r = 3, q = 3$ (Setting B)



(f) $\sigma_1^2 = 1, r = 6, q = 3$ (Setting B)



(g) $\sigma_1^2 = 2, r = 3, q = 3$ (Setting B)



(h) $\sigma_1^2 = 2, r = 6, q = 3$ (Setting B)

Figure S.6: Empirical FDRs and powers for testing the overall temporal DE genes by our method using GA_3 basis (Δ), compared to those of the oracle and true test ($\circ$ and $\diamond$), `edgeR` ($\bullet$), `maSigPro-GLM`($\blacksquare$), `splineTC` (∇), `ImpulseDE2` ($\boxtimes$), and `DESeq2` ($\times$). The blue dot lines ($\cdots\cdots$) denote results for other bases described in Section 5 in the main paper. Each point displays the empirical FDR and power of the corresponding method at a given nominal FDR level (the vertical gray dashed lines). Results are for $T = 6$, $\mu_1 = 4$ and based on 100 replications.

Table S.3: Comparison of empirical FDRs and powers for testing NPDE genes by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting B. In simulations, $\mu_1 = 2$, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

		(T, r, σ_1^2)							
		$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
GA_2	FDR	0.031	0.035	0.038	0.038	0.036	0.037	0.049	0.054
	Power	0.067	0.083	0.217	0.260	0.110	0.107	0.407	0.403
GA_3	FDR	0.046	0.031	0.030	0.031	0.026	0.027	0.038	0.038
	Power	0.083	0.077	0.200	0.223	0.090	0.090	0.350	0.323
FO_2	FDR	0.045	0.059	0.056	0.062	0.037	0.060	0.058	0.063
	Power	0.070	0.087	0.207	0.233	0.100	0.063	0.327	0.297
FO_3	FDR	0.038	0.049	0.035	0.043	0.026	0.042	0.049	0.049
	Power	0.060	0.083	0.183	0.213	0.083	0.063	0.313	0.297
PL_2	FDR	0.029	0.041	0.035	0.035	0.037	0.034	0.047	0.050
	Power	0.067	0.080	0.207	0.257	0.107	0.103	0.403	0.373
Oracle	FDR	0.023	0.034	0.028	0.028	0.026	0.027	0.035	0.035
	Power	0.063	0.087	0.183	0.220	0.083	0.083	0.357	0.330
True	FDR	0.037	0.034	0.037	0.037	0.032	0.033	0.041	0.042
	Power	0.100	0.100	0.203	0.247	0.113	0.113	0.373	0.360
`edgeR`	FDR	0.108	0.133	0.061	0.061	0.053	0.053	0.043	0.046
	Power	0.007	0.007	0.077	0.077	0.049	0.046	0.215	0.214
`DESeq2`	FDR	0.005	0.053	0.041	0.037	0.030	0.022	0.031	0.035
	Power	0.005	0.005	0.074	0.071	0.047	0.044	0.206	0.203

Table S.4: Comparison of empirical FDRs and powers for testing NPDE genes by the proposed method with different bases, `edgeR`, and `DESeq2` for Setting B. In simulations, $\mu_1 = 4$, T, r and σ_1^2 are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.

		(T, r, σ_1^2)							
		$(6,3,1)$	$(6,3,2)$	$(6,6,1)$	$(6,6,2)$	$(10,3,1)$	$(10,3,2)$	$(10,6,1)$	$(10,6,2)$
GA_2	FDR	0.036	0.034	0.038	0.037	0.038	0.032	0.048	0.050
	Power	0.077	0.073	0.287	0.260	0.177	0.090	0.377	0.400
GA_3	FDR	0.024	0.022	0.028	0.029	0.026	0.022	0.037	0.037
	Power	0.073	0.077	0.253	0.227	0.160	0.077	0.320	0.337
FO_2	FDR	0.059	0.058	0.065	0.063	0.061	0.063	0.061	0.062
	Power	0.097	0.093	0.243	0.243	0.143	0.070	0.303	0.307
FO_3	FDR	0.037	0.038	0.047	0.043	0.047	0.045	0.051	0.048
	Power	0.080	0.073	0.233	0.220	0.137	0.077	0.307	0.303
PL_2	FDR	0.031	0.031	0.036	0.034	0.038	0.031	0.048	0.048
	Power	0.087	0.073	0.277	0.247	0.167	0.107	0.357	0.397
Oracle	FDR	0.024	0.023	0.027	0.027	0.022	0.022	0.036	0.035
	Power	0.080	0.063	0.250	0.220	0.147	0.077	0.333	0.333
True	FDR	0.030	0.038	0.038	0.036	0.032	0.032	0.043	0.040
	Power	0.087	0.083	0.277	0.230	0.180	0.113	0.373	0.363
`edgeR`	FDR	0.107	0.111	0.058	0.058	0.060	0.054	0.043	0.045
	Power	0.009	0.008	0.087	0.085	0.053	0.050	0.219	0.220
`DESeq2`	FDR	0.018	0.019	0.036	0.037	0.027	0.033	0.032	0.037
	Power	0.007	0.007	0.081	0.080	0.053	0.050	0.211	0.211

B.3.3 Additional Results on Different DE Proportions and Dispersion Estimators in Our Method

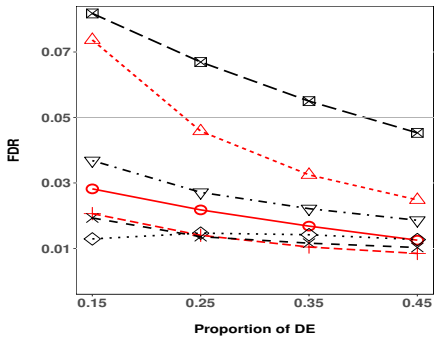
In this section, we report additional simulations to investigate the effects of the proportion of temporally DE genes and the dispersion estimation on the performance of the proposed method. We focus on testing the overall temporal DE genes, which is the alternative to Δ_0^{DE} in Section 3.2 in the main paper. We consider two settings for the proportions and compositions of different types of DE genes based on our proposed model in (5) and (6) in the main paper. Specifically, the proportions for the four components in our model, *null genes, NPDE genes with only time-by-treatment interaction, PDE genes, and NPDE genes with both treatment and time-by-treatment effects* are set as:

- setting (I): the proportion of null genes are $0.85, 0.75, 0.65$, and 0.55; and the proportions for DE genes in three components are $(0.075, 0.04, 0.035)$, $(0.20, 0.04, 0.01)$, $(0.125, 0.04, 0.185)$, and $(0.150, 0.04, 0.26)$;

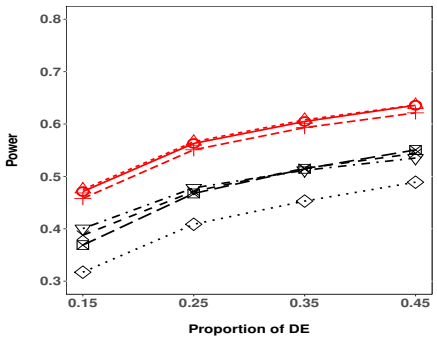
- setting (II): the proportion of null genes are $0.85, 0.75, 0.65$, and 0.55; and the proportions for DE genes in three components are $(0.10, 0.04, 0.01)$, $(0.20, 0.04, 0.01)$, $(0.30, 0.04, 0.01)$, and $(0.40, 0.04, 0.01)$.

For both settings, the proportions of all DE genes are $0.15, 0.25, 0.35$ and 0.45 but with different compositions. For setting (II), the only type of DE genes with changing proportions are from the 2nd component in our model, which are the NPDE genes with only time-by-treatment interaction. For setting (I), in addition to the 2nd component, DE genes from the 4th component in our model, which are the NPDE genes with both treatment and time-by-treatment effects, also have changing proportions. We consider `edgeR`, `splineTC`, `ImpulseDE2`, and `DESeq2` for comparison. Also, we employ three different dispersion estimates for our methods (with GA_3 for model fitting as in Section 4 in the main paper), empirical dispersion $\widehat{\phi}_g$ in (B.4), common dispersion estimate $G^{-1}\sum_{g=1}^{G}\widehat{\phi}_g$ used for all genes, and dispersion estimated by `DESeq2`. In simulations, we generate

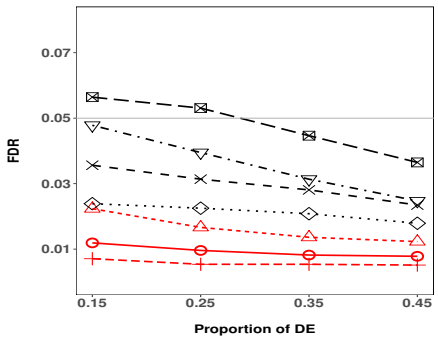
$G = 1,000$ genes using basis function PL_2 with $\sigma_1^2 = 1, T = 6$, and $\mu_1 = 2$. All results are based on 100 repetitions.



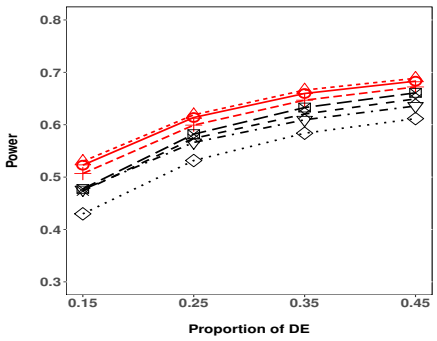
(a) $r = 3$, proportion setting (I).



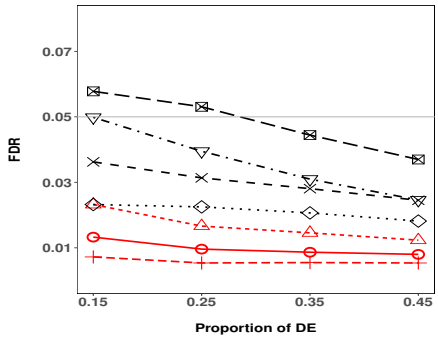
(b) $r = 3$, proportion setting (I).



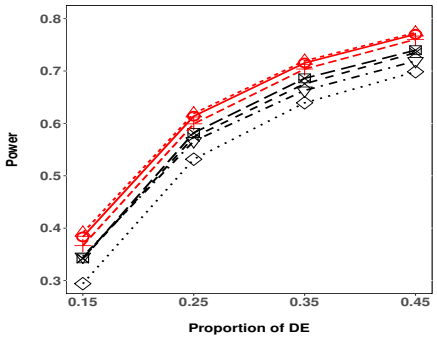
(c) $r = 6$, proportion setting (I).



(d) $r = 6$, proportion setting (I).



(e) $r = 6$, proportion setting (II).



(f) $r = 6$, proportion setting (II).

Figure S.7: Empirical FDRs and powers for testing the overall temporal DE genes with different proportion settings of DE genes. Displayed methods include `edgeR` (⊠), `splineTC` (▽), `ImpulseDE2` (◇), `DESeq2` (×), and our methods using empirical dispersion (△), common dispersion (○), and dispersion estimated by `DESeq2` (+).

Overall, the empirical FDRs for most methods are satisfactorily controlled with respect to the nominal 0.05 level under both settings. When both the proportion of DE genes and the number of replicates r are small, `edgeR` and our method with empirical dispersion estimator in (B.4) have empirical FDRs slightly inflated. However, when the number of replicates r increases, the empirical FDRs improve for `edgeR` and our method with empirical dispersion estimator. In addition, even for small number of replicates, the empirical FDRs for these two methods substantially improve as the proportion of DE genes increases. For both settings, the empirical FDRs of the proposed method with either common dispersion estimator or the dispersion estimated by `DESeq2` are relatively robust against the proportion of DE genes under both settings for larger number of replicates, say $r = 6$.

In terms of the empirical power, the proposed method with different dispersion estimators perform similarly under both settings, and all outperform other methods for different DE proportions. The proposed method with dispersion estimated from `DESeq2` is slightly conservative than the other two, which reflects the shrinkage effect of `DESeq2` on estimating dispersion. When the number of replicates r is small, the advantage of our method is more substantial in comparison to competing methods. For low DE proportions, all methods may encounter challenges while ours still provides substantially better powers than others. For example, in panel (f) in Figure S.7, the proposed method with different dispersion estimators still have empirical powers about 15% better than those of `DESeq2`, `splineTC`, and `edgeR` when the DE proportion is only 0.15. For other methods, `ImpulseDE2` is usually less powerful than others, which can be explained by its specific yet relatively stringent parametric model on the mean dynamics; `splineTC` performs slightly better than `edgeR` and `DESeq2` when the DE proportion is low and r is small.

Furthermore, Figure S.7 also suggests that the dispersion estimate of `DESeq2` is the key to its own performance. In fact, many studies show that estimation of the dispersion parameter in the small RNA-seq experiments may affect the empirical FDR, which is one of the motivations to borrow information across genes. In terms of borrowing information to estimate dispersion parameters, it is known [50, 66] that the weighted likelihood approach in `edgeR` performs better

than the local regression approach used in DESeq2. By inspecting simulation results of edgeR and DESeq2 in Figures 1–2 and Tables 1–2 in the main paper as well as Figures S.1–S.6 and Tables S.1–S.4, we observe that the empirical FDR of edgeR is closer to the nominal level than that of DESeq2. In summary, the dispersion estimate affects DESeq2's performance on DE analysis of time-course RNA-seq data.

Finally, comparing to existing methods in the simulation study, DESeq2 usually has conservative empirical FDRs while its empirical power is somewhat comparable to other methods from literature. A possible explanation is that the time-course RNA-seq experiment is relatively large compared to the traditional RNA-seq experiment. That is, the sample size (or the number of libraries) is usually rT rather than r for estimating the dispersion parameters. Therefore, in practice, DESeq2 is still a reasonable method to be considered for DE analysis on time-course RNA-seq data.

B.3.4 More Results on Testing Composite Hypotheses

The null space Δ_0 in (1) in the main paper can be flexibly defined for different composite hypotheses, which are used to model biological questions of particular interest. Besides those specified in (7) in the main paper, we can also set, for example, $\Delta_0 = \{\eta_{g1} \in (-\infty, 0]\}$ to test whether the mean gene expressions over time in the second group is higher than that in the first. We can also set $\Delta_0 = \{|\eta_{g1}| : |\eta_{g1}| \leq d\}$ with $d = \log 2$ to detect genes with mean log fold-changes of expressions, over time, greater than 2. The proposed framework can accommodate to arbitrary Δ_0 as a subset of $\mathbb{R}^{2q+1}$ while most existing methods for analyzing time-course RNA-seq data only allow the simple hypothesis. This is an important advantage of our method.

To demonstrate this, we consider a simulation study to detect NPDE genes with significant mean shift, which simply corresponds to identifying genes from the fourth component of the proposed model. That is, we set the null hypothesis as $\Delta_0 = \{\eta_{g1} = 0\} \cup \{\boldsymbol{\eta}_{g2} = \mathbf{0}\}$, which is a composite null. In order to detect these genes using edgeR or DESeq2, we must employ a two-step approach. Using edgeR or DESeq2, we first test the treatment effect and then the

Table S.5: Comparisons of the empirical FDRs and powers for testing NPDE with significant mean shift (model is fitted using basis function GA_3) along those of `DESeq2` and `edgeR`. The nominal FDR level is 0.05. The simulation is based on 100 replications.

	Our method fitting with GA_3	`DESeq2`	`edgeR`
FDR	0.083	0.455	0.000
Power	0.280	0.080	0.040

time-by-treatment effect. Genes that are significant for both hypotheses will be considered as the desired DE genes. In this simulation, we generate $G = 1,000$ genes from model (5) in the main paper with basis function PL_2, proportions of four components are $0.75, 0.15, 0.05,$ and 0.05, $T = 6, r = 3, \sigma_1^2 = 0.5$, and $\mu_1 = 8$. That is, 5% of the total genes are of interest. For our method, we use 300 Monte-Carlo nodes to evaluate likelihood functions. Results are displayed in Table S.5. While our method controls the empirical FDR and provides reasonable power for this challenging problem, both `DESeq2` and `edgeR` have compromised powers and `DESeq2` has inflated empirical FDR. In fact, `DESeq2` falsely treats genes with large treatment effect without time-by-treatment effect as the target genes, which leads to many false detection.

B.3.5 Computational Complexity

In this section, we conduct small simulations to demonstrate the computational cost of the proposed method with others along their performance in terms of empirical FDR and power. First, we consider Setting A (PL_2) from the simulation study in the main paper with $G = 1,000$ genes generated by model in (5) and (6), where $\mu_1 = 2, \sigma_1^2 = 1$, $r = 3$, $T = 6$, and the proportion of four components in (6) are $0.65, 0.125, 0.04$ and 0.01 (that is, the proportion of temporal DE genes is 0.35). For implementation, the number of the quasi-Monte Carlo nodes used in numerical integration is 300. The nominal level of FDR is 0.05 and results are based on 100 replications. Results for testing overall temporal DE genes are displayed in Table S.6. We can see that, as expected, our method is more demanding in computation than others but it is also the most powerful.

Table S.6: A simulation study on computational costs of different methods for analyzing time-course RNA-seq data along their performances. (Time in seconds.)

	Ours	DESeq2	ImpulseDE2	splineTC	edgeR	maSigPro-GLM
FDR	0.032	0.011	0.014	0.022	0.055	0.054
Power	0.608	0.515	0.453	0.512	0.514	0.527
Time	790	< 10	590	< 1	1.1	7.40

The quasi-Monte Carlo integration with shrinking number of nodes from Algorithm 1 in Section B.2 is used for the evaluation of likelihood. To further demonstrate the computational intensity, we conduct another small simulation to study the influence of the numbers of Monte-Carlo nodes and observed time points on the computational cost (measured in seconds). For results displayed in Table S.7, we still consider Setting A (PL_2) from the simulation study in the main paper with $G = 1,000$ genes generated by model in (5) and (6), where $\mu_1 = 4, \sigma_1^2 = 1$, $r = 3$, and the proportion of four components in (6) are $0.75, 0.2, 0.04$ and 0.01 (that is, the proportion of temporal DE genes is 0.25). The number of Monte-Carlo nodes varies from $300, 500, 700$ to 1000, and $T = 6, 8, 10, 12$. We use a linux machine with an Intel Xeon E5-2680 v3 @2.50GHz CPU and 8GB RAM. As we observed, though the computational demands gradually increase as the number of nodes or T increasing, the trend is linear rather than exponential. In practice, the evaluation of the likelihood function at Monte-Carlo nodes can be carried out using parallel computing, which substantially reduce the computation cost.

Table S.7: Computational intensity of the proposed method with respect to the number of Monte-Carlo nodes N and the number of observed time points T.

		T			
	N	6	8	10	12
Time (seconds)	300	559.10	629.46	656.20	729.91
	500	882.29	917.88	954.24	1033.96
	700	1169.09	1160.34	1261.33	1434.98
	1000	1535.34	1593.35	1750.32	1915.68

B.4 Additional Results for Real Data Analysis

B.4.1 Preprocessing the *P.t.* Data

As suggested by [3], we first filter out genes with overall low expression or low variation across libraries. If a gene does not display much variation across time for both groups, it possesses very limited temporal patterns and might be analyzed using traditional RNA-seq analysis methods. For this analysis, we focus on those genes with temporal dynamics. After filtering, $10,597$ genes remain for further analysis. For computational stability, we map the time domain from $(0, 24)$ hours to $(0, 1)$. Normalization factors S_{ij}'s are pre-computed using the standard TMM method [50]. In addition, to facilitate the proposed method, we assume that there is an overall mean trend across all genes for both groups to be adjusted. For example, for the high light group it assumes that $\lambda_{g1j}(t) = S_{1j} \exp\{\mathbf{B}(t)\boldsymbol{\gamma}_2\} \exp\{\mathbf{B}(t)\boldsymbol{\eta}_{g2}\}$, where $\boldsymbol{\gamma}_2$ is adjusted by fitting all data from the high light group to the traditional negative-binomial generalized linear model and setting $\widehat{\boldsymbol{\gamma}}_2 = \sum_{g=1}^{G} \widehat{\boldsymbol{\eta}}_{2g}^{\text{NBglm}}$ and $S_{1j} \exp\{\mathbf{B}(t)\widehat{\boldsymbol{\gamma}}_2\}$ will be treated as the normalization factors. Similar pre-processing is applied to the low light group.

B.4.2 Test Results for the *P.t.* Data

In Figure S.8, top 10 genes identified by the proposed method with only significant relative mean shift are displayed.

B.4.3 Gene Set Analysis for the *P.t.* Data

Gene set enrichment analysis is a powerful and revealing follow-up step for RNA-seq analysis [1, 160] and it has been successfully employed to identify gene sets with longitudinally changing patterns for time-course RNA-seq data [37]. By scrupulously inspecting predefined gene sets, we can not only verify statistical discoveries but more importantly also identify critical pathways responsive to the treatment variations.

For *P.t.*, it is known that the 32-gene set for porphyrin and chlorophyll biosynthesis (PCB), the 52-gene set for oxidative phosphorylation (OP), and the 15-gene set for galactoglycerolipid biosyn-

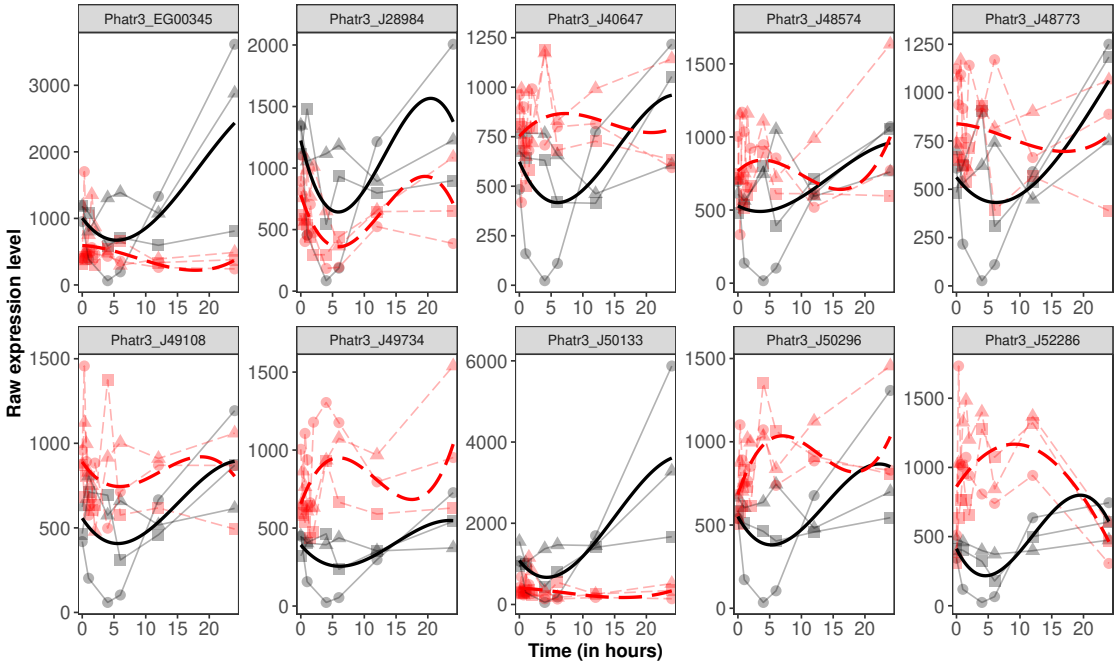


Figure S.8: Top 10 genes identified by the proposed method with only significant relative mean shift. The red solid curves represent data from high light group and the blue dot curves represent data from the low light group. Dots represent the real data points while the bold smooth curves display smooth estimations using orthogonal polynomials. Captions are the gene tags from [2].

thesis (GB) are predicted photoacclimation relevant pathways. Based on our DE analysis results above, the enrichment analyses are conducted using `R` package `gsEasy` with results displayed in Table S.8. From Table S.8, PCB and OP are enriched for all alternatives under consideration while GB is not enriched for the relative mean shift but only NPDE. The traditional heatmap visualization for genes within GB are displayed in Figure S.9, from which we observe that the overall dynamics indeed alter more than the simple level shift when the light environment changes. Visualizations for PCB and OP are included in the Web Appendix.

Anatomically, the plastid of a diatom is the subcellular organelle where photosynthesis takes place. Galactolipids are categorized as those containing galactose as the polar head groups. They are integral components of the membranes of plastids and are not found in significant amounts in other membranes [161]. Photoacclimation to low light leads to a large increase in plastid volume, and is hypothesized to house increases in the light harvesting apparatus to capture more light [68]. Our finding on the enrichment of GB, particularly the NPDE, provides statistical evidence on

Table S.8: Results of the gene set enrichment analysis for the three photoacclimation relevant gene sets. The p-values are derived based on the results of DE analysis using our proposed method and the functional enrichment analysis proposed by [1]. Enrichment of a gene set for certain hypothesis is significant when the p-value is less than 0.05.

Pathway	Overall temporal DE	Relative mean shift Δ_0^{Mean}	NPDE Δ_0^{NPDE}
Porphyrin and Chlorophyll Biosynthesis	0.002	$< 1.0e\text{-}5$	0.002
Oxidative Phosphorylation	$< 1.0e\text{-}5$	$< 1.0e\text{-}5$	$< 1.0e\text{-}5$
Galactoglycerolipid Biosynthesis	0.040	0.095	0.024

the importance of this predicted biosynthesis pathway of galactolipids [162] and reveals how the dynamics of the whole pathway vary between different light conditions. These results show that our proposed method does provide important insights into transcriptional changes that result in major alterations to an organism's biochemistry.

Visualizations of the results on the gene set enrichment analysis for gene sets PCB and OP are displayed in Figures S.10 and S.11.

B.4.4 Results on Fission Yeast Data Analysis in Introduction

Figure S.12 provides a Venn Diagram to summarize the number of overall temporal DE genes from the fission yeast data [3] detected by our method, `DESeq2`, and the LRT procedure described in Section 1 in the main paper. As discussed in the main paper, the proposed method detects the most temporal DE genes, or equivalently, the estimated FDR for the proposed test is the smallest if we declared the same number of significant genes for the three methods in this small illustrative example. Detecting the most DE genes suggests the outstanding power of a test if the FDR control is guaranteed, however follow-up experiments and downstream analysis are needed to confirm the detected genes and will help the evaluation of the method. For example, the gene set enrichment analysis we conduct for the *P.t.* data study reported in Section 5 in the main paper and Section B.4.3 in this Web Appendix.

Figure S.9: Visualization of genes within the set Galactoglycerolipid Biosynthesis (GB). No color (blank) encodes for time points that no data are collected for the high light group. GB is enriched for NPDE but not the relative mean shift.

All genes detected by LRT are detected by either our method or `DESeq2`. Though our proposed method and `DESeq2` detect many common DE genes, there are some discrepancies. This can be explained by the implicit smoothness assumption on the mean dynamics pattern in our method: our method is more powerful when the mean pattern across time of a gene can be modeled by a smooth function. Alternatively, we can replace the smooth basis functions in (5) in the main paper by step functions $\mathbb{I}(t = t_{ij})$ to handle this situation at the cost of more parameters to be further modeled in (6).

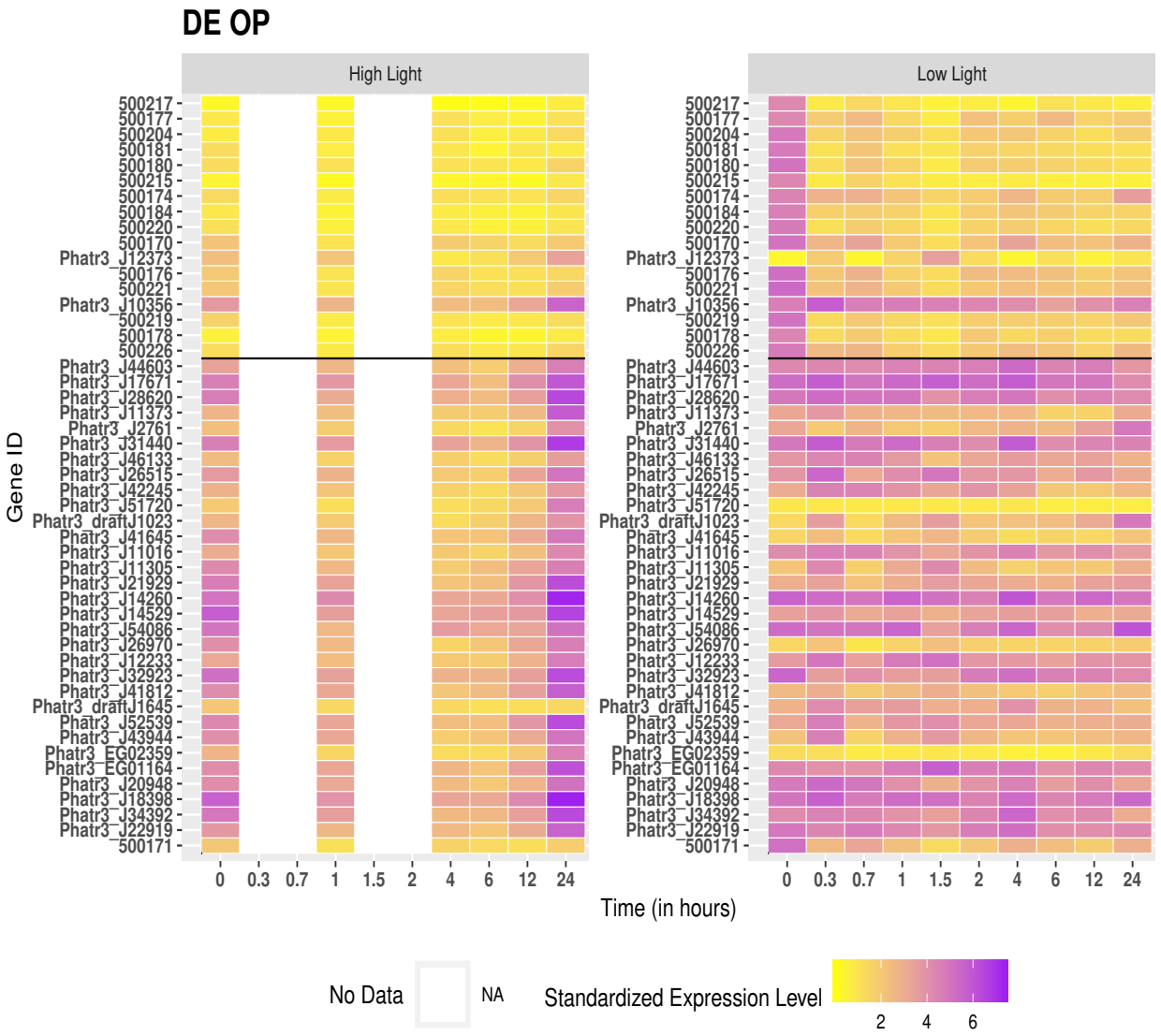


Figure S.10: Visualization of genes within the set Oxidative Phosphorylation (OP). No color (blank) encodes for time points that no data are collected for the high light group. OP is enriched for both the relative mean shift and NPDE.

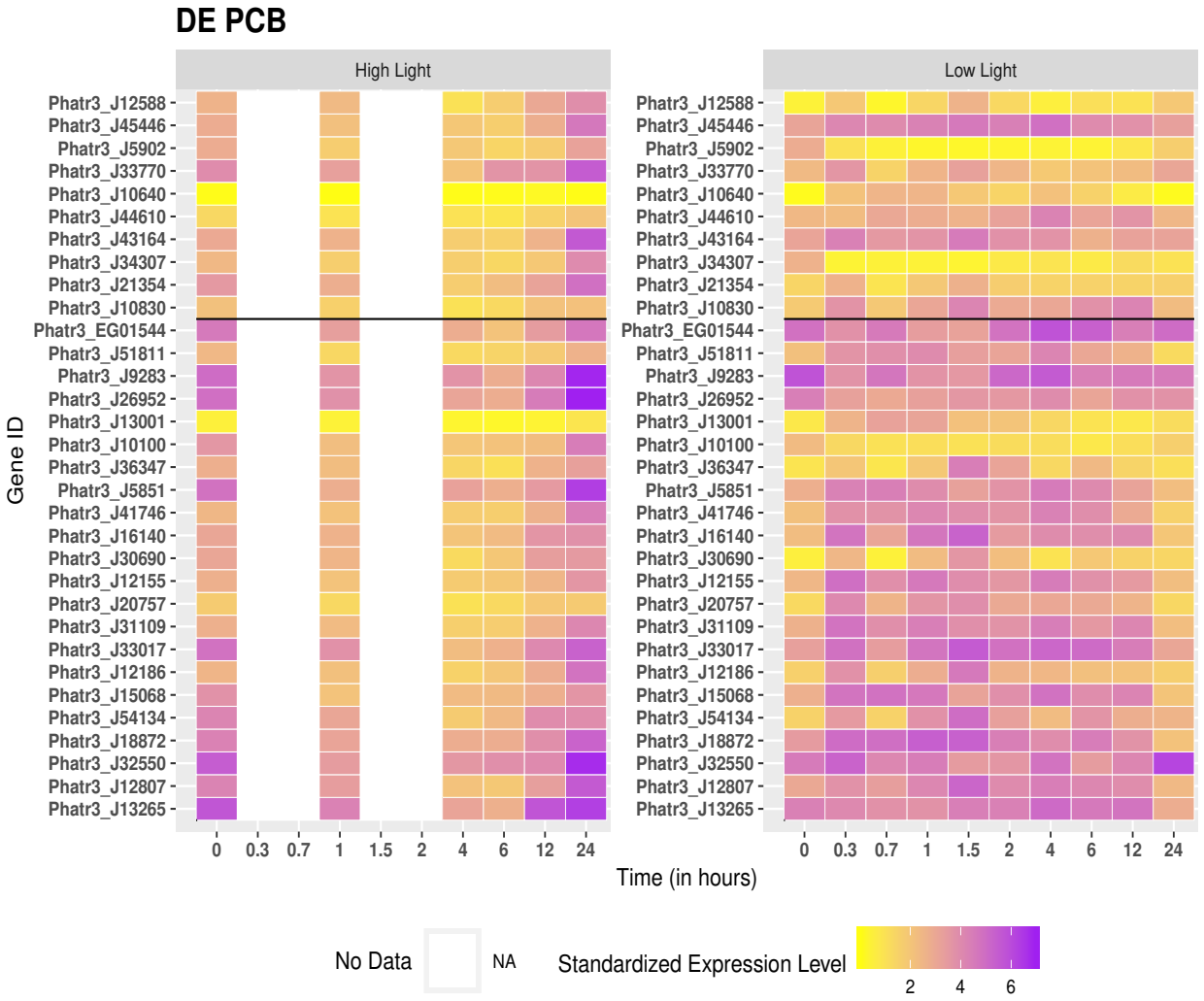


Figure S.11: Visualization of genes within the set Porphyrin and Chlorophyll Biosynthesis (PCB). No color (blank) encodes for time points that no data are collected for the high light group. PCB is enriched for both the relative mean shift and NPDE.

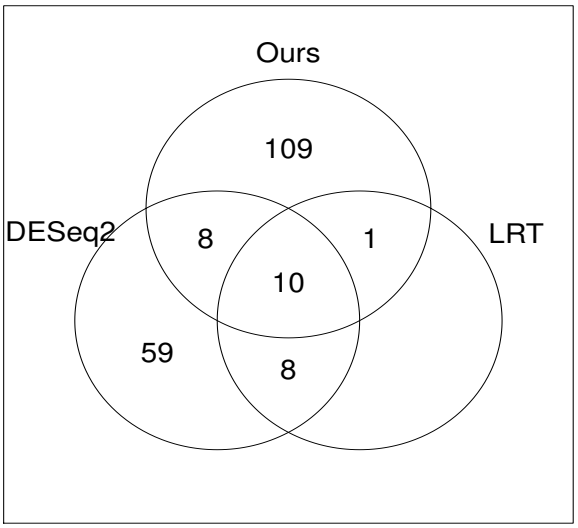


Figure S.12: Number of overall temporal DE genes identified by the proposed method, `DESeq2`, and LRT method for the fission yeast data from [3].

Appendix C

Supplemental materials for Chapter 4

C.0.1 Weighted Least Square

We rewrite each component of $\mathbf{Y}_{j_g}$ to a distribution in exponential family ([163]), taking the form

$$\ell = \{y_{j_g,ijk}\boldsymbol{\theta}_{j_g,ijk} - b(\boldsymbol{\theta}_{j_g,ijk})\}/a(r_{j_g}) + c(y_{j_g,ijk}, r_{j_g}).$$

In our case, given $a(r_{j_g}) = 1$, $\boldsymbol{\theta}_{j_g,ijk} = \log\left(\frac{m_{j_g,ijk}}{r_{j_g}+m_{j_g,ijk}}\right)$, $b(\boldsymbol{\theta}_{j_g,ijk}) = r_{j_g}\log(r_{j_g}) - r_{j_g}\log(1 - \exp(\boldsymbol{\theta}_{j_g,ijk}))$, with r_{j_g} as known constant. We will have, by definition,

$$\frac{\partial b(\boldsymbol{\theta}_{j_g,ijk})}{\partial \boldsymbol{\theta}_{j_g,ijk}} = m_{j_g,ijk}$$

and

$$\frac{\partial^2 b(\boldsymbol{\theta}_{j_g,ijk})}{\partial \boldsymbol{\theta}^2_{j_g,ijk}} = \frac{r_{j_g}\exp(\boldsymbol{\theta}_{j_g,ijk})}{(1-\exp(\boldsymbol{\theta}_{j_g,ijk}))^2} = V_{j_g,ijk}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_{j_g,ijk}} = \frac{(y_{j_g,ijk} - m_{j_g,ijk})}{a(r_{j_g})}\frac{1}{V_{j_g,ijk}}\frac{d\boldsymbol{\eta}_{j_g,ijk}}{dm_{j_g,ijk}}x_{j_g,ijk}.$$

The maximum likelihood equation for $\boldsymbol{\beta}_{j_g,ijk}$ are given by

$$\sum_{j_g}\sum_{k}\frac{(y_{j_g,ijk} - m_{j_g,ijk})}{a(r_{j_g})}\frac{1}{V_{j_g,ijk}}\frac{dm_{j_g,ijk}}{d\boldsymbol{\eta}_{j_g,ijk}}x_{j_g,ijk} = 0,$$

which is the same as

$$\sum_{j_g}\sum_{k}\frac{W_{j_g,ijk}(y_{j_g,ijk} - m_{j_g,ijk})}{a(r_{j_g})}\frac{d\boldsymbol{\eta}_{j_g,ijk}}{dm_{j_g,ijk}}x_{j_g,ijk} = 0,$$

where

$$W_{j_g,ijk} = V_{j_g,ijk}^{-1} \left(\frac{dm_{j_g,ijk}}{d\boldsymbol{\eta}_{j_g,ijk}} \right)^2 = \frac{r_{j_g} m_{j_g,ijk}}{r_{j_g} + m_{j_g,ijk}},$$

$\boldsymbol{\eta}_{j_g,ijk} = x_{j_g,ijk}^T \boldsymbol{\beta}_{j_g,ijk}$.

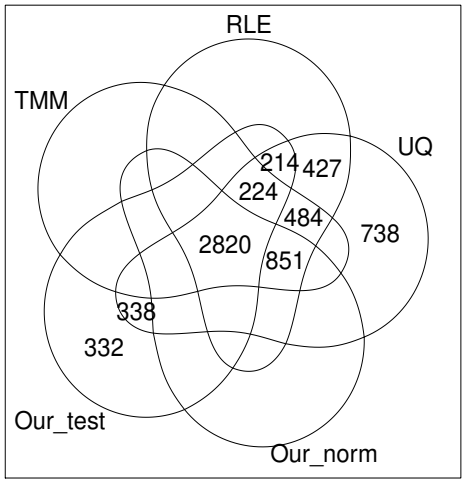
C.0.2 Real Data Analysis

In Table S.1, DE sRNAs identified by the proposed method that may be misannotated are displayed.

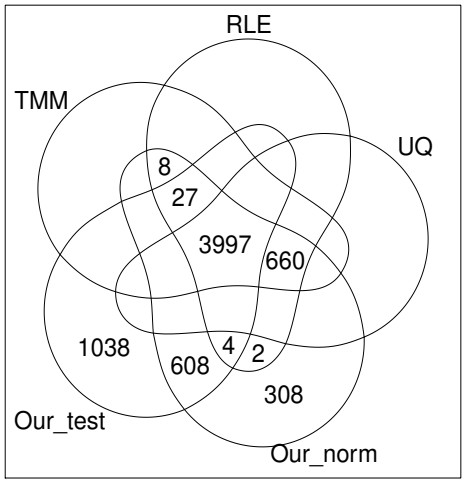
Table S.1: sRNAs that may be misannotated.

Experiment	Feature	Class
wt-vs-prg-1	21ur-8412	piRNA
	21ur-15576	piRNA
	21ur-15116	piRNA
	21ur-10492	piRNA
	miR-78	miRNA
	miR-42-5p	miRNA
	miR-4936	miRNA
	miR-54-5p	miRNA
	miR-797-3p	miRNA
	miR-785	miRNA
	miR-73-5p	miRNA
	miR-74-5p	miRNA
	miR-5549-3p	miRNA
	miR-85-3p	miRNA
	miR-2214-5p	miRNA
	miR-41-5p	miRNA
wt-vs-mut16	miR-260	miRNA
	miR-2210-3p	miRNA
	miR-4816-5p	miRNA
	miR-785	miRNA

Figure S.1 displays the comparisons of the proposed method with `edgeR`.



(a) wt-vs-mut16



(b) wt-vs-prg-1

Figure S.1: differential expression analysis results comparing the proposed method with `edgeR`.

Appendix D

Supplemental materials for Chapter 5

D.1 Log-likelihood, score vector and Fisher information matrix

The log-likelihood of the observations $\{r_i\}_{i=1}^n$ is given by

$$\begin{aligned}\ell\left(\boldsymbol{\beta}; \{r_i\}_{i=1}^n\right) &= \sum_{i=1}^n \ln \mathrm{P}\left[R_i = r_i\right] \\ &= \sum_{i=1}^n \ln\left\{\mathrm{P}\left[R_i = r_i \mid B_i = r_i\right]\mathrm{P}\left[B_i = r_i\right] + \mathrm{P}\left[R_i = r_i \mid B_i = m+1\right] b(m+1)\right\} \\ &= \sum_{i=1}^n \ln\left\{b(r_i) + \pi_i(r_i \mid \boldsymbol{\beta}) b(m+1)\right\} \qquad \text{(D.1)}\end{aligned}$$

where $\pi_i(r_i \mid \boldsymbol{\beta})$ is the Poisson probability mass function with mean $\mu_i = z_i \exp(\boldsymbol{x_i}'\boldsymbol{\beta})$. Let $g(r_i) = b(r_i) + \pi_i(r_i \mid \boldsymbol{\beta}) b(m+1)$. Then the jth component of the score vector is

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial}{\partial \beta_j} \ln\{g(r_i)\} = \sum_{i=1}^n \frac{1}{g(r_i)} x_{ij} b(m+1)\left(\frac{-e^{-\mu_i}\mu_i^{r_i+1}}{r_i!} + \frac{r_i \mu_i^{r_i} e^{-\mu_i}}{r_i!}\right) \\ &= \sum_{i=1}^n f_j(r_i)\end{aligned}$$

for $j = 1, 2, \ldots, p$.

The Fisher information matrix is defined as $\mathcal{I}(\boldsymbol{\beta}) = -\mathrm{E}\left(\partial^2 \ell / \partial\beta_j \partial\beta_k\right)$ for $j, k = 1, 2, \ldots, p$. Now

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^n \frac{\partial}{\partial \beta_k} f_j(r_i) \\ &= \sum_{i=1}^n \Bigg[\{-f_j(r_i) f_k(r_i)\} + \frac{x_{ij} x_{ik} b(m+1)}{g(r_i) r_i!} \times \\ &\quad \left\{e^{-\mu_i}\mu_i^{r_i+2} - (r_i+1)e^{-\mu_i}\mu_i^{r_i+1} + r_i^2 \mu_i^{r_i} e^{-\mu_i} - e^{-\mu_i} r_i \mu_i^{r_i+1}\right\}\Bigg],\end{aligned}$$

so,

$$\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) \quad = \quad \sum_{i=1}^{n}\left[x_{ik}x_{ij}\left\{-b(m+1)\mu_i + b(m+1)\sum_{r=0}^{m}\frac{\pi(r)b(r)}{g(r)}(r-\mu_i)^2\right\}\right].$$

Two special cases of the information matrix are of interest. If $b(m+1) = 0$, then no responses are true, and the data contain no information about the model parameters: $\mathrm{E}\left(\partial^2\ell/\partial\beta_j\partial\beta_k\right) = 0$ for all j and k. If $b(0) = b(1) = \cdots = b(m) = 0$, then all responses are true, and the information matrix is that of ordinary Poisson regression.

D.2 Code

The simulation and empirical results of this paper were obtained using our `R` package `QRRT`. The `R` language and environment for statistical computing [164] is freely available and runs on many computing platforms (UNIX, Windows, MacOS). From within `R`, the `QRRT` package is downloadable via GitHub (install the package `devtools` first, if necessary) using the following commands:

```
library(devtools)
install_github("meca7653/QRRT")
library(QRRT)
```

The following reproducible example is included with the code and accessed with `help(QRRT)`. The example uses one simulated realization from the true, additive model

$$\ln \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_4 \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_5 x_{i1} x_{i2}, \tag{D.2}$$

as described in the simulation section. It fits those simulated data using the larger-than-necessary interaction model

$$\begin{aligned}&\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_4 \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_5 x_{i1} x_{i2} \\ &+ \beta_6 x_{1i} \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_7 x_{1i} \mathbf{1}_{\{x_{i3}=\mathrm{C}\}} + \beta_8 x_{2i} \mathbf{1}_{\{x_{i3}=\mathrm{B}\}} + \beta_9 x_{2i} \mathbf{1}_{\{x_{i3}=\mathrm{C}\}},\end{aligned} \tag{D.3}$$

via the following specification:

```
fit_2way <-
  QRRT(
    Formula = Ri ~ (x1 + x2 + as.factor(x3)) ^ 2,
    Data = Sim_Data,
    Disperse = 1,
    beta = NULL,
    n_times = 10,
    offset = NULL,
    b_distribution = c(6, 7, 4, 2, 2, 1, 1, 1, 1, 25) / 50
  )
```

Here, the `Formula` uses standard `R` syntax to specify the model with all two-way interactions; similarly,

```
fit_truemodel <-
  QRRT(
    Formula = Ri ~ (x1 + x2) ^ 2 + as.factor(x3),
    Data = Sim_Data,
    Disperse = 1,
    beta = NULL,
    n_times = 10,
    offset = NULL,
    b_distribution = c(6, 7, 4, 2, 2, 1, 1, 1, 1, 25) / 50
  )
```

would specify the true model (D.2). Other examples, including the use of an offset, accompany the code.

Next, `Data` specifies a data frame `Sim_Data` consisting of the three covariates $\{x_{1i}\}$, $\{x_{2i}\}$, $\{x_{3i}\}$ and the observed responses $\{r_i\}$. Because the starting value is specified as `beta = NULL`, the code selects `n_times = 10` different random starts for the $\boldsymbol{\beta}$ coefficients, using independent normal random variables with mean zero and standard deviation `Disperse = 1`. The `offset` is not used in this example, but takes its default null value (a vector of zeroes on the logarithmic scale) The `b_distribution` argument specifies the $b(r)$ distribution from

$$(b(0), b(1), \ldots, b(8), b(9)) = \frac{1}{50}(6, 7, 4, 2, 2, 1, 1, 1, 1, 25). \tag{D.4}$$

The code then runs the EM algorithm to convergence from each random start, finally returning the fitted model with highest likelihood:

	Estimate	Std.Error	t-statistic	Pr(>|t|)
(Intercept)	1.397	0.152	9.205	3.41e-20
x1	1.047	0.145	7.223	5.08e-13
x2	-0.557	0.073	-7.663	1.82e-14
as.factor(x3)B	0.644	0.183	3.528	4.18e-04
as.factor(x3)C	0.504	0.183	2.751	0.006
x1:x2	0.203	0.062	3.279	0.001
x1:as.factor(x3)B	-0.164	0.173	-0.948	0.343
x1:as.factor(x3)C	-0.122	0.175	-0.698	0.485
x2:as.factor(x3)B	0.058	0.051	1.126	0.260
x2:as.factor(x3)C	0.039	0.052	0.757	0.449

The `Estimate` column of the above output shows point estimates of the true regression coefficients

$$(\beta_0, \ldots, \beta_5, \beta_6, \ldots, \beta_9) = (1.5, 1.0, -0.5, 0.4, 0.3, 0.2, 0, 0, 0, 0),$$

with excellent agreement relative to the asymptotic standard errors (`Std.Error`). That is, the fitted model correctly identifies the non-zero coefficients $(\beta_0, \ldots, \beta_5)$, with large `t-statistic`

(estimate over standard error) and small p-values (`Pr(>|t|)`) and gives point estimates consistent with the true values. It also correctly identifies the zero coefficients, $(\beta_6, \ldots, \beta_9)$, with small t-statistics and large p-values.

To test the hypothesis that model (D.2) fits as well as model (D.3), we compute the log-likelihood ratio via

```
-2 * fit_true$Maximized_Log_Likelihood
    + 2 * fit_2way$Maximized_Log_Likelihood.
```

The resulting test statistic is 1.82364, with corresponding p-value of 0.7681545, computed via

```
1 - pchisq(q = 1.82364, df = 4)
```

from the χ^2 distribution with 4 degrees of freedom.

The Monte Carlo experiment of this paper repeats the above simulation, estimation and hypothesis test 1000 times. There is no evidence to reject the null hypothesis that model (D.2) fits as well as model (D.3).