2018

Multiple hypothesis testing and RNA-seq differential expression analysis accounting for dependence and relevant covariates

Yet Nguyen
Iowa State University

Multiple hypothesis testing and RNA-seq differential expression analysis accounting for dependence and relevant covariates

by

Yet Nguyen

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dan Nettleton, Major Professor
Alicia Carriquiry
Peng Liu
Jarad Niemi
Dan Nordman

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

For my family.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEGEMENTS

I would like to thank Dr. Dan Nettleton for being my major professor. There are no words that can express how grateful I am to him. His guidance, encouragement, patience, attention to detail, and generosity help me become a better version of myself. I hope that I can apply most lessons he's taught me throughout my future career.

I am so grateful to Dr. Max Morris for his helpful advice and suggestions during two semesters as my mentor in the Preparing Future Faculty program. I've learnt a lot from him about academic life in such a short time.

I truly appreciate Dr. Jack Dekkers, Dr. Christ Tuggle, and Dr. Haibo Liu providing datasets that motivate most of my research.

I would also like to thank my committee members, Dr. Alicia Carriquiry, Dr. Peng Liu, Dr. Jarad Niemi and Dr. Dan Nordman, for their time and consideration.

Lastly, I would like to thank my family for their unconditional support and love.

ABSTRACT

This dissertation is a collection of four papers on the development of statistical methods for the analysis of high-dimensional data, mostly RNA-seq gene expression data. We introduce in the first two papers two covariate-selection strategies for RNA-seq analysis. As in any experiment or observational study, covariates may hold information about heterogeneity of the experimental or observational units used in the investigation. Either ignoring relevant covariates or accounting for irrelevant covariates may be detrimental to RNA-seq analysis. We show through simulation that our methods outperform methods that do not take covariate selection into account. Next, we develop in the third paper a parametric bootstrap algorithm to analyze RNA-seq datasets from repeated measures designs. In such designs, RNA samples are extracted from each experimental unit at multiple time points. The read counts that result from RNA sequencing of the samples extracted from the same experimental unit tend to be temporally correlated. Simulation studies show the advantages of our method over alternatives that do not account for correlation among observations within experimental units. Finally, we develop a new method to estimate and control false discovery rate (FDR) when identifying simultaneous signals in two independent experiments. Our FDR estimation and control procedure is a generalization of the histogram-based FDR estimation and control procedure for one experiment proposed by Nettleton et al. (2006); Liang and Nettleton (2012). We show that our method performs better than other existing methods both in theory and in simulation.

CHAPTER 1. GENERAL INTRODUCTION

This dissertation consists of four separate papers with the same theme: statistical methods for high-dimensional data and multiple testing. Our work is motivated by scientific questions in biological research. The datasets used in this thesis are gene expression datasets produced by RNA-sequencing technology. The methodologies focus on RNA-seq differential expression analysis with complex designs (designs with many covariates and repeated-measures designs) and false discovery rate control and estimation in multiple testing. In this chapter, we present an overview of RNA-seq data, differential expression analysis, false discovery rate control and estimation in multiple testing, and finally an outline of the dissertation.

1.1 RNA-seq Data and Differential Expression Analysis

RNA-seq is a next generation sequencing technology by which one can measure several features, such as measures of gene transcript abundance, often referred to as gene expression levels. The outcome of an RNA-seq experiment is typically represented as a data matrix of counts with rows representing genes and columns representing samples from one or more populations. Each row of counts tends to be positively correlated with transcript abundance levels of the corresponding gene in the samples.

A common challenge in RNA-seq data analysis is to identify differentially expressed genes, i.e., genes whose mean expression levels change across different groups of samples, or, more generally, are associated with one or more variables of interest. Such analysis is called differential expression analysis. Differential expression analysis usually involves carrying out a significance test for each gene. Because RNA-seq data generally contain thousands of genes, differential expression analysis involves testing thousands of hypotheses.

1.2 False Discovery Rate Control and Estimation in Multiple Hypothesis Testing

When conducting thousands or millions of hypothesis tests, familywise error rate (FWER) control procedures such as Bonferroni are often too conservative for practical use. In such situations, false discovery rate (FDR) is a favorable error rate measure. FDR is defined as the expected proportion of true nulls among the rejected hypotheses (where the proportion is defined as zero if no null hypotheses are rejected). Controlling FDR at a level α means that, on average, the proportion of false discoveries among all discoveries is at most α. Hypothesis testing procedures aiming to control FDR tend to be considerably more powerful than procedures aiming to control FWER.

1.3 Dissertation Organization

Following this general introduction in Chapter 1, Chapter 2 and Chapter 3 contribute two methodologies in RNA-seq differential expression analysis that allow for an effective accounting for relevant covariates. In particular, we propose two covariate-selection strategies to select the most relevant covariates whose effects are accounted for in RNA-seq analysis. Chapter 4 contains a method to analyze RNA-seq datasets from repeated-measures designs. We use normalized log-counts and associated precision weights in a general linear model pipeline with continuous autoregressive structure to account for correlation among repeated measures; statistical inference is conducted using a parametric bootstrap procedure. Chapter 5 presents a novel histogram-based FDR estimation and control procedure when identifying simultaneous signals in two independent experiments. Finally, Chapter 6 is devoted to concluding remarks and some directions for future work.

Throughout Chapters 2 to 5, Yet Nguyen developed the proposed methods, conducted evaluations, and wrote the initial manuscripts; Dan Nettleton suggested the proposed methods, their revisions, and contributed significantly to the writing. Jack Dekkers, Christ Tuggle and Haibo Liu provided datasets for Chapter 2, 3 and 4. Megan Orr, Peng Liu and Dan Nettleton proposed

the initial method in Chapter 5. The research was supported by Dan Nettleton through multiple sources: Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) and the joint National Science Foundation (NSF)/NIGMS Mathematical Biology Program under award number R01GM109458, and Plant Science Institute, ISU.

CHAPTER 2. DETECTING DIFFERENTIALLY EXPRESSED GENES WITH RNA-SEQ DATA USING BACKWARD SELECTION TO ACCOUNT FOR THE EFFECTS OF RELEVANT COVARIATES

A paper published in the *Journal of Agricultural, Biological, and Environmental Statistics*

Yet Nguyen, Dan Nettleton, Haibo Liu, and Christopher Tuggle

Abstract

A common challenge in analysis of transcriptomic data is to identify differentially expressed genes, i.e., genes whose mean transcript abundance levels differ across the levels of a factor of scientific interest. Transcript abundance levels can be measured simultaneously for thousands of genes in multiple biological samples using RNA sequencing (RNA-seq) technology. Part of the variation in RNA-seq measures of transcript abundance may be associated with variation in continuous and/or categorical covariates measured for each experimental unit or RNA sample. Ignoring relevant covariates or modeling the effects of irrelevant covariates can be detrimental to identifying differentially expressed genes. We propose a backward selection strategy for selecting a set of covariates whose effects are accounted for when searching for differentially expressed genes. We illustrate our approach through the analysis of an RNA-seq study intended to identify genes differentially expressed between two lines of pigs divergently selected for residual feed intake. We use simulation to show the advantages of our backward selection procedure over alternative strategies that either ignore or adjust for all measured covariates.

2.1 Introduction

A standard challenge in transcriptomic data analysis is to identify genes whose mean transcript abundance levels differ across the levels of a categorical factor of primary scientific interest (e.g.,

treatment, genotype, tissue, or disease state). Such genes are typically referred to as differentially expressed (DE). Currently, the leading technology used to detect DE genes is RNA sequencing (RNA-seq). In raw form, RNA-seq data contain information about the identity of bases in short RNA sequence fragments known as reads. For the purpose of identifying DE genes, the number of reads matching each of thousands of gene sequences is determined for each of several experimental or observational units. These read counts serve as measures of RNA abundance. Typically, a generalized linear model with a log link and a negative binomial response is fit to the count data for each gene, and DE genes are identified by testing, for each gene, whether a model parameter or linear combination of model parameters is zero.

RNA-seq datasets often contain several covariates in addition to the factor of primary scientific interest. As in any experiment or observational study, covariates may hold information about heterogeneity of the experimental or observational units used in the investigation. Other covariates in an RNA-seq dataset may track variation that is created during the complex process of measuring RNA transcript abundance levels using RNA-seq technology. If covariates are ignored when searching for DE genes, the unaccounted for variation in expression levels associated with variation in covariates may obscure the association of expression levels with the primary factor of interest. On the other hand, explicitly accounting for the effects of all covariates in data analysis may be inefficient when some covariates are actually unassociated or only weakly associated with expression levels. Either ignoring relevant covariates or accounting for the effects of irrelevant covariates reduces power for identifying DE genes. Unfortunately, the power problem is exacerbated by the low sample sizes common in expensive RNA-seq experiments.

To address the challenge of identifying DE genes with RNA-seq datasets that include covariates, we propose a backward selection algorithm for selecting a subset of covariates whose effects are estimated and adjusted for when testing for differential expression. Our goal is to find one subset of all available covariates to include in every gene-specific generalized linear model. Although it is possible (and perhaps even likely) that the subset of covariates relevant for one gene is different than the subset of covariates relevant for another, we seek one subset of covariates common to

all genes for two main reasons. First, the number of experimental/observational units is often relatively small in RNA-seq datasets, especially in agricultural applications. Small sample sizes lead to unreliable model selection and considerable uncertainty in models selected separately for tens of thousands of genes. Second, for purposes of interpretability, it is useful to test for differential expression by adjusting for the same set of covariates for all genes. Identifying DE genes involves testing whether one (or more) partial regression coefficients in a generalized linear model is zero. If different covariates are used for different genes, the interpretation of partial regression coefficients – and consequently the definition of differential expression – changes from gene to gene. A shifting definition of what it means for a gene to be DE is undesirable when reporting results. Instead, we choose one subset of covariates for all genes and attempt to answer the following question: If we adjust for the effects of the subset of variables that tends to be most relevant when considering all genes, do we see significant differences in mean transcript abundance levels across the levels of the factor of primary scientific interest?

As a motivating example, we consider RNA-seq measures of transcript abundance in blood samples from 31 pigs of two genetic lines created by selection on the basis of residual feed intake (RFI). RFI is computed as the observed feed intake of an animal minus an estimate of the feed intake that would be expected considering that animal's growth characteristics. Pigs from the high residual feed intake (HRFI) line tend to eat more feed than expected considering their growth, while pigs from the low residual feed intake (LRFI) line tend to eat less than expected considering their growth. Because feed is the largest single cost incurred by US pork producers, pigs of the LRFI line have economically desirable feeding and growth characteristics, and understanding the transcriptional differences between these lines is of scientific interest.

Finding genes differentially expressed between lines is complicated by heterogeneity among pigs, heterogeneity among the blood samples extracted from pigs, and heterogeneity among the processed and measured RNA samples derived from the blood samples. A total of 13 categorical and continuous covariates (described in detail in the Appendix) are available for tracking and accounting for this heterogeneity. The backward selection procedure that we formally define in Section 2.2.3

starts by fitting, for each gene, a full generalized linear model with a negative binomial response and a log link that includes the effects of primary interest due to line as well as effects for all 13 covariates. Using criteria described in Section 2.2.4, the least relevant variable when considering results from all genes is dropped, and the resulting reduced model is fit for all genes. This process continues until the variable identified as least relevant is the factor of scientific interest (line in our example). This backward selection procedure produces a sequence of increasingly smaller subsets of covariates, starting with all covariates and progressing, one removed variable at a time, down to a subset of covariates most strongly associated with transcript abundance levels when considering the results for all genes. From this sequence of subsets of covariates, we determine the subset of covariates that, when accounted for, leads to identification of the greatest number of genes differentially expressed across the levels of the factor of primary scientific interest (i.e., line).

The mechanics of our backward selection procedure are similar to those of the usual backward selection procedure used in multiple regression in that the variable least significant (by some criterion) is removed at each step. One major difference between our proposed procedure and the usual backward selection procedure for multiple regression is that we are dealing simultaneously with thousands of response variables rather than a single response. A second major difference (related to the first) is that the subset of variables we ultimately select from the sequence of subsets generated by backward selection is determined by maximizing the number of rejected null hypotheses for a test of interest across thousands of response variables. This strategy is motivated by the knowledge that both including irrelevant covariates and excluding relevant covariates can act to reduce power. Thus, selecting the set of covariates that maximizes the number of rejected hypotheses for the test of interest is a natural strategy for identifying the most relevant covariates.

In a simulation study presented in Section 2.4, we show that our backward selection procedure is effective at selecting the truly relevant covariates when the truly relevant covariates are the same for all genes. In this idealized situation, our simulations also show that the false discovery rate (FDR) can be controlled when tests for differential expression are conducted while adjusting for the effects of the covariates selected using our backward selection procedure. We also show that FDR

can still be controlled even when the set of truly relevant covariates differs across genes. However, results must be carefully interpreted if some excluded covariates are associated with the factor of primary scientific interest.

Prior to presenting our differential expression analysis of the RFI RNA-Seq dataset in Section 2.3, we provide more details about generalized linear models and significance testing for RNA-seq read count data in Sections 2.2.1 and 2.2.2. We formally define our proposed backward selection procedure in Section 2.2.3. Section 2.2.4 covers two measures of covariate relevance that can be used to choose covariates for removal in each step of backward selection. We compare the performance of the backward selection algorithm with alternative methods in a simulation study presented in Section 2.4. The paper concludes with a discussion in Section 2.5.

2.2 Methods

2.2.1 Generalized Linear Models for RNA-seq Read Count Data

Consider the analysis of m genes using RNA-seq read count data from n experimental or observational units. For $g = 1, \ldots, m$ and $i = 1, \ldots, n$, let y_{gi} be the read count for gene g from experimental/observational unit i. Let $\boldsymbol{x}_i = (\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{ik})'$ denote a vector of known explanatory variable values for the ith unit. Without loss of generality, we assume that $\boldsymbol{x}_{i1}$ is a vector of zero-one indicator variable values that code for the level of the factor of primary scientific interest associated with unit i. The number of components of $\boldsymbol{x}_{i1}$ is one less than the number of levels of the factor of primary scientific interest. For example, for the RFI dataset discussed in Section 2.1 and in more detail in Section 2.3 and the Appendix, $\boldsymbol{x}_{i1}$ is simply a single indicator variable that takes the value 1 if the ith pig is from the LRFI line and the value 0 if the ith pig is from the HRFI line. Each of the other vectors $\boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik}$ corresponds to either a continuous or categorical covariate that is not of primary scientific interest. Vectors for continuous covariates have only one element while vectors corresponding to categorical covariates consist of indicator variable values with one less indicator than the number of levels of the categorical covariate. Finally, let o_i be the normalization offset computed for unit i. The normalization offsets account for differences in the

thoroughness of sequencing across the units. A variety of normalization offsets have been proposed in the literature (see, e.g., Marioni et al. (2008), Mortazavi et al. (2008), Robinson and Oshlack (2010), Anders and Huber (2010), Bullard et al. (2010), Risso et al. (2014a), Risso et al. (2014b), and references therein). Throughout this paper, we set o_i to be the log of the 0.75 quantile of unit i read counts in accordance with the recommendation of Bullard et al. (2010).

As is popular in RNA-seq data analysis, we use, as a working assumption, that the read counts for gene g $(y_{g1}, \ldots, y_{gn})$ are independent and that $y_{gi} \sim \text{NB}(\mu_{gi}, \omega_g)$, where $\text{NB}(\mu_{gi}, \omega_g)$ is the negative binomial distribution with mean μ_{gi}, dispersion parameter ω_g, and variance $\mu_{gi} + \omega_g \mu_{gi}^2$. Letting $\mathcal{S}$ represent a subset of $\{1, \ldots, k\}$ that contains 1, we consider log-linear models of the form

$$\log(\mu_{gi}) = o_i + \beta_{g0|\mathcal{S}} + \sum_{j \in \mathcal{S}} \boldsymbol{x}'_{ij} \boldsymbol{\beta}_{gj|\mathcal{S}}, \tag{2.1}$$

where $\beta_{g0|\mathcal{S}}$ is an unknown intercept parameter, $\boldsymbol{\beta}_{g1|\mathcal{S}}$ is an unknown parameter vector for the factor of primary scientific interest, and $\boldsymbol{\beta}_{gj|\mathcal{S}}$ is a vector of unknown covariate effects for each $j \in \mathcal{S} \setminus \{1\}$. The set $\mathcal{S}$ is included in the parameter subscripts to emphasize that the meaning of partial regression coefficients depends on all the covariates included in the model. We use $\mathcal{S}^*$ to represent the unknown set containing 1 and the largest subset of $\{2, \ldots, k\}$ such that $j \in \mathcal{S}^* \setminus \{1\}$ implies $\boldsymbol{\beta}_{gj|\mathcal{S}} \neq \mathbf{0}$ for some $g \in \{1, \ldots, m\}$. This makes $\mathcal{S}^* \setminus \{1\}$ the set of all indices corresponding to covariates relevant for at least one gene.

For all $g = 1, \ldots, m$, we wish to test $H_{0g1}^{\mathcal{S}^*} : \boldsymbol{\beta}_{g1|\mathcal{S}^*} = \mathbf{0}$. If $H_{0g1}^{\mathcal{S}^*}$ is false, gene g is said to be differentially expressed (DE). Otherwise, gene g is said to be equivalently expressed (EE). Because the set of all relevant covariates $\mathcal{S}^* \setminus \{1\}$ is unknown, we cannot directly test $H_{0g1}^{\mathcal{S}^*}$ for any gene g. Instead, we use a backward selection procedure to first identify a set of covariates $\hat{\mathcal{S}}^*$ to approximate $\mathcal{S}^*$. Then, for each gene g, we fit the (possibly misspecified) model in which the true equation defining $\log(\mu_{gi})$ in (2.1) is replaced by

$$\log(\mu_{gi}) = o_i + \beta_{g0|\hat{\mathcal{S}}^*} + \sum_{j \in \hat{\mathcal{S}}^*} \boldsymbol{x}'_{ij} \boldsymbol{\beta}_{gj|\hat{\mathcal{S}}^*}. \tag{2.2}$$

Note that the regression coefficients in (2.2) are the same as the partial regression coefficients in (2.1) whenever $\hat{\mathcal{S}}^* = \mathcal{S}^*$. Even if $\hat{\mathcal{S}}^* \neq \mathcal{S}^*$, the partial regression coefficients of interest given by $\boldsymbol{\beta}_{g1|\mathcal{S}^*}$ may be similar to $\boldsymbol{\beta}_{g1|\hat{\mathcal{S}}^*}$ if $\hat{\mathcal{S}}^*$ includes the most relevant covariates. In such situations, reasonable decisions about whether $\boldsymbol{\beta}_{g1|\mathcal{S}^*} = \mathbf{0}$ may be reached by testing $H_{0g1}^{\hat{\mathcal{S}}^*} : \boldsymbol{\beta}_{g1|\hat{\mathcal{S}}^*} = \mathbf{0}$. Thus, for each $g \in \{1, \ldots, m\}$, we test $H_{0g1}^{\hat{\mathcal{S}}^*} : \boldsymbol{\beta}_{g1|\hat{\mathcal{S}}^*} = \mathbf{0}$, and we use the p-values from these m tests to declare a subset of the m genes to be differentially expressed.

2.2.2 Significance Testing for RNA-seq Read Count Data

A variety of methods have been proposed for testing the significance of regression coefficients in generalized linear models for RNA-seq read count data. Some prominent examples include Lu et al. (2005), Robinson and Smyth (2007, 2008), Anders and Huber (2010), Hardcastle and Kelly (2010), Yanming et al. (2011), Van De Wiel et al. (2013), and McCarthy et al. (2012). A recent review of methods was provided by Lorenz et al. (2014). To conduct our tests for differential expression and to assess the significance of covariates, we use the `QuasiSeq` `R` package, which implements the quasi-likelihood testing method developed by Lund et al. (2012). This approach was recently found by Burden et al. (2014) to be the "best performing package in the sense that it achieves a low FDR which is accurately estimated over the full range of p-values."

In brief, the `QuasiSeq` method uses a hierarchical model for gene-specific quasi-dispersion parameters to obtain quasi-dispersion parameter estimates that are stabilized by borrowing information across genes. For each gene, the usual likelihood ratio test statistic for testing the significance of a subvector of regression coefficients is then scaled by the inverse of the estimated quasi-dispersion parameter. This scaled test statistic is then compared to an appropriate central F distribution to obtain an approximate p-value. Approximate control of the false discovery rate (FDR) at any desired level α is obtained by converting the p-values to q-values Storey (2002) and rejecting a null hypothesis if and only if its corresponding q-value is less than α. When computing q-values by the method of Storey (2002), an estimate of m_0, the number of true null hypotheses among all m null hypotheses tested, is required. We use the histogram-based method of Nettleton et al. (2006)

to estimate m_0. Desirable theoretical characteristics of a closely related histogram-based approach were demonstrated by Liang and Nettleton (2012).

The denominator degrees of freedom parameter for the F distribution used to obtain p-values in the quasi-likelihood analysis is bounded below by the sample size minus the number of estimated partial regression coefficients and, all else equal, will decrease as irrelevant covariates are included in the model. Decreased denominator degrees of freedom can result in a loss in power for detecting DE genes. On the other hand, excluding relevant covariates will increase the denominator degrees of freedom at the cost of larger quasi-dispersion parameter estimates due to lack of model fit. Because the estimated quasi-dispersion parameters are the denominators of the F statistics, larger quasi-dispersion parameter estimates lead to smaller F statistics and, again, reduced power for identifying differentially expressed genes. For these reasons, finding the most relevant set of covariates is crucial for differential expression analysis.

2.2.3 The Proposed Backward Selection Algorithm

Let $\mathcal{S}$ be any subset of $\{1,\ldots,k\}$. For any $j \in \mathcal{S}$, let $\boldsymbol{p}_{j|\mathcal{S}}$ denote the vector of m p-values obtained by testing $H_{0gj}^{\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \mathbf{0}$ for each gene $g = 1,\ldots,m$. Let $r(\boldsymbol{p}_{j|\mathcal{S}})$ be a measure of the relevance of $\boldsymbol{x}_j$ in model (2.1); as an example, the simplest of the two relevance measures we consider in this paper (see Section 2.2.4) is the number of elements of $\boldsymbol{p}_{j|\mathcal{S}}$ less than 0.05. Let $\mathcal{S}_1 = \{1,\ldots,k\}$ and consider an iterative procedure whose ℓth iteration is as follows:

1. Compute $\boldsymbol{p}_{j|\mathcal{S}_\ell}$ for all $j \in \mathcal{S}_\ell$.
2. Let $\boldsymbol{q}_\ell$ be the vector of q-values obtained from $\boldsymbol{p}_{1|\mathcal{S}_\ell}$.
3. Let $R_\ell(\alpha)$ be the number of q-values in $\boldsymbol{q}_\ell$ less than or equal to a user-defined FDR threshold α.
4. Find j^* so that $r(\boldsymbol{p}_{j^*|\mathcal{S}}) \leq r(\boldsymbol{p}_{j|\mathcal{S}})$ for all $j \in \mathcal{S}_\ell$.
5. If $j^* = 1$, stop iterating. Otherwise, carry out the $\ell+1$st iteration with $\mathcal{S}_{\ell+1} = \mathcal{S}_\ell \setminus \{j^*\}$.

Suppose the iterative procedure concludes after L iterations, and let ℓ^* be the smallest element of $\{1, \ldots, L\}$ such that $R_{\ell^*}(\alpha) \geq R_\ell(\alpha)$ for all $\ell \in \{1, \ldots, L\}$. We set $\hat{\mathcal{S}}^* = \mathcal{S}_{\ell^*}$ and base our inference about differential expression on the fit of model (2.2). By the definition of ℓ^*, this analysis will maximize the number of genes declared to be differentially expressed (at FDR threshold α) over the L models that correspond to the L explanatory variable index sets $\mathcal{S}_1 \supset \cdots \supset \mathcal{S}_L$. Despite maximizing the number of genes declared differentially expressed over the sequence of models, we show through simulation studies in Section 2.4 that this approach can control the false discovery rate at desired levels.

2.2.4 Measures of Variable Relevance

For a given $\mathcal{S} \subseteq \{1, \ldots, k\}$ and any $j \in \mathcal{S}$, we consider $\boldsymbol{x}_j$ to be an irrelevant variable if

$$H_{0gj}^{\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \mathbf{0} \text{ is true for all } g = 1, \ldots, m. \tag{2.3}$$

When (2.3) holds, each element of $\boldsymbol{p}_{j|\mathcal{S}}$ will be uniformly distributed on $(0, 1)$ whenever the test used to produce the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ has size equal to the significance level for all significance levels in $(0, 1)$. If the test used to produce the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ is unbiased for all significance levels, then an element of $\boldsymbol{p}_{j|\mathcal{S}}$ corresponding to a false null hypothesis will have a distribution stochastically smaller than uniform$(0, 1)$ and a density that is decreasing on the interval $(0, 1)$. Based on this reasoning, the empirical distribution of the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ provides information about the relevance of $\boldsymbol{x}_j$ in the model that includes the explanatory variables whose indices are contained in $\mathcal{S}$. An empirical distribution close to uniform or stochastically larger than uniform implies little relevance while an empirical distribution with a clear excess of small p-values relative to a uniform distribution implies relevance of $\boldsymbol{x}_j$ for at least some appreciable number of genes.

In practice, the tests used to assess significance are only approximate, each observed p-value is only a single draw from its marginal distribution, and dependence among genes leads to dependence among p-values. For all of these reasons, empirical distributions comprised of one p-value from each gene can have shapes that are neither uniform nor stochastically smaller than uniform. Nonetheless, measuring the extent to which an empirical distribution of the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ departs from uniform

towards a distribution with a decreasing density on $(0, 1)$ can provide a useful measure of relevance for variable $\boldsymbol{x}_j$. As examples, the histograms in the first row of Figure 2.1 show the empirical distribution of the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ for each $j \in \mathcal{S} = \{1, \ldots, 14\}$. Based on visual inspection, covariates like *RINb*, *Conca*, *Order*, *Diet*, and *Eosi* appear irrelevant in the full model, while covariates like *Concb*, *Neut*, *Mono*, and *Block* appear relevant.

There are many ways to formally measure the relevance of explanatory variable $\boldsymbol{x}_j$ through definition of a relevance function $r(\cdot)$ that maps $\boldsymbol{p}_{j|\mathcal{S}}$ to the real line. We consider two choices for $r(\cdot)$, one relatively simple and one more complicated. It turns out that both measures of relevance lead to similar performance for our backward selection and testing procedure. As noted in the previous section, the simpler of our two relevance measures sets $r(\boldsymbol{p}_{j|\mathcal{S}})$ to the number of elements of $\boldsymbol{p}_{j|\mathcal{S}}$ less than 0.05. We use p.05 as an abbreviation for this criterion in the remainder of the paper. The more complicated version of $r(\cdot)$ is described as follows.

Given a vector of p-values $\boldsymbol{p} = (p_1, \ldots, p_m)'$, let $\hat{F}_m(\cdot)$ be the empirical distribution function of the elements of $\boldsymbol{p}$. If we were to assume the elements of $\boldsymbol{p}$ were an independent and identically distributed sample from a distribution with cumulative distribution function $F(\cdot)$, then $\hat{F}_m(\cdot)$ is known to be the nonparametric maximum likelihood estimator of $F(\cdot)$. If we were to assume that the distribution defined by $F(\cdot)$ has a non-increasing density, then the nonparametric maximum likelihood estimator of $F(\cdot)$, subject to the constraint of a non-increasing density, is given by $\tilde{F}_m(\cdot)$, the least concave majorant of $\hat{F}_m(\cdot)$ (Grenander, 1956). If we let

$$r(\boldsymbol{p}) = \sqrt{m} \sup_{x \in (0,1)} [\tilde{F}_m(x) - x], \tag{2.4}$$

then $r(\boldsymbol{p})$ is a Kolmogorov-Smirnov-type statistic that measures the extent to which the empirical distribution of the elements of $\boldsymbol{p}$ departs from a uniform$(0, 1)$ distribution towards a distribution with a decreasing density on $(0, 1)$. Henceforth, we refer to this measure of variable relevance as the GKS criterion (short for Grenander-Kolmogorov-Smirnov).

2.3 Analysis of the RFI RNA-Seq Dataset

The proposed backward selection algorithm was used to analyze the RFI RNA-seq dataset introduced in Section 2.1 and described in more detail in the Appendix. Recall that the primary scientific goal is to identify genes whose mean transcript abundance levels, adjusted for relevant covariates, differ between the LRFI and HRFI lines. The dataset consists of read counts for 12280 genes for each of 31 pigs. As is customary in RNA-seq analysis, this dataset excludes genes with predominantly low read counts because genes with low read counts contain little information about differential expression and can lead to computation problems when attempting to fit negative binomial models. Thus, the 12280 genes analyzed in this study each have average read counts of at least 8 and no more than 27 zero counts across the 31 pigs. This same threshold for gene inclusion was used throughout the simulation study described in Section 2.4.

Table 2.1: The first 14 rows show the number of p-values less than 0.05 for each covariate at each iteration of the backward selection algorithm applied to the RFI RNA-seq data. The last row $R_\ell(0.05)$ is the number of q-values less than or equal to 0.05 for the test of the *Line* factor in each iteration.

	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$	$\ell = 7$	$\ell = 8$	$\ell = 9$
RINb	**202**								
Order	235	**324**							
eosi	303	340	**320**						
Conca	450	503	421	**409**					
Diet	392	497	489	507	**335**				
RFI	585	708	1004	1042	879	**917**			
Baso	255	458	742	742	1262	1396	**1275**		
mono	1400	1531	1293	1519	1371	1496	1506	**1432**	
Line	722	1221	1682	1793	1676	1949	2303	2139	**2235**
Concb	1680	2635	3015	3326	4353	4414	4385	4331	4153
RINa	281	681	1939	2020	2100	2118	2543	2594	2997
neut	1138	1704	2123	2155	2290	2350	2352	2987	2919
Lymp	625	818	1119	1251	1350	1393	1354	1606	4225
Block	967	1259	1867	2152	2379	2456	2406	2380	2440
$R_\ell(0.05)$	2	2	1	1	0	46	**448**	337	379

Table 2.1 and Figure 2.1 summarize the results of the backward selection algorithm when p.05 is used as the measure of covariate relevance. The covariate *RINb* was the first to be removed from the full model, followed in subsequent iterations by the covariates *Order*, *Eosi*, *Conca*, *Diet*, *RFI*, *Baso*, and *Mono*. At the 9th iteration, *Line* was judged to be the least relevant factor, and thus,

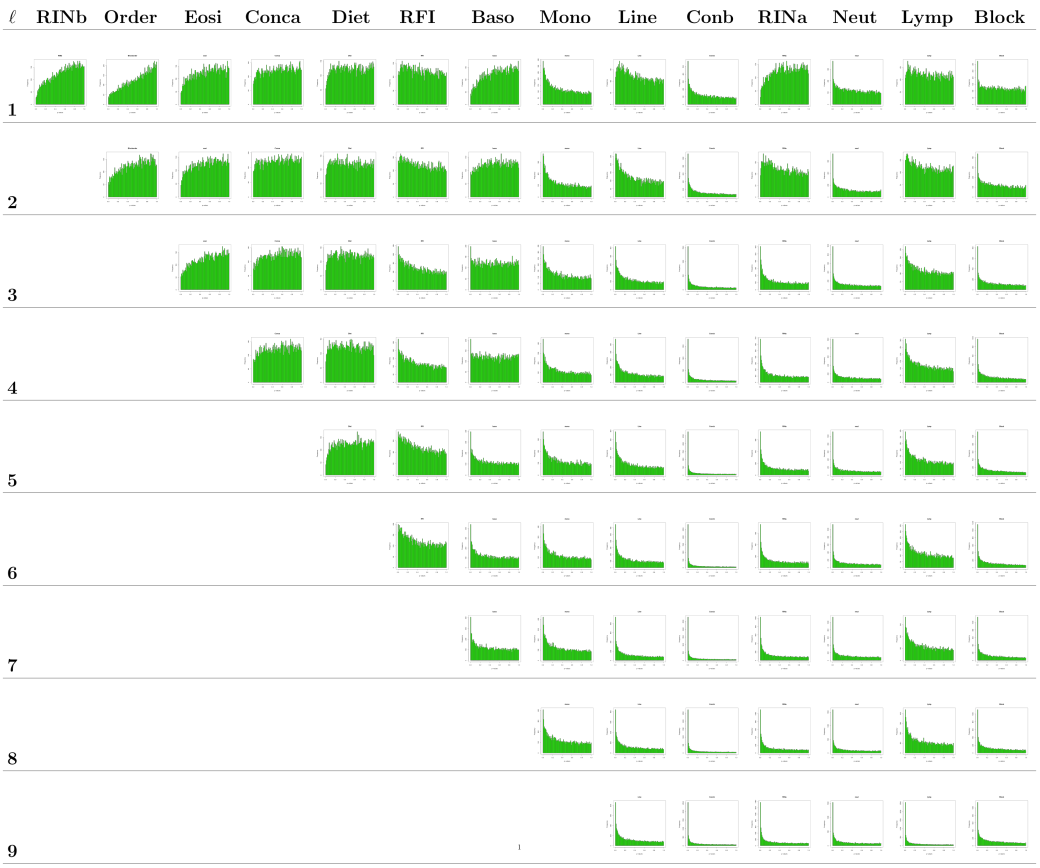

Figure 2.1: Histograms of p-values at each iteration of the backward selection procedure applied to the RFI RNA-seq dataset using the number of p-values less than 0.05 (p.05) as the measure of covariate relevance. Rather than using a common upper limit for each histogram's vertical axis, the upper limit varies across histograms to accommodate the height of the tallest bar in each histogram. Using variable upper limits makes it easier to see differences between the histogram shapes of relevant and irrelevant covariates.

the backward selection procedure terminated. As the bottom row of Table 2.1 shows, the model corresponding to iteration $\ell = 7$ yielded the greatest number ($R_7(0.05) = 448$) of q-values no larger than 0.05 for the *Line* test. Hence, the backward selection procedure resulted in the declaration of 448 genes as differentially expressed between the LRFI and HRFI lines while controlling for the effects of the covariates *Baso*, *Mono*, *Concb*, *RINa*, *Neut*, *Lymp*, and *Block*. For this dataset, the backward selection procedure using the GKS criterion to measure covariate relevance deleted variables in a slightly different order but selected the same final model and, therefore, provided results identical to backward selection using p.05 to measure covariate relevance.

The results of our proposed backward selection procedure can be contrasted with two simple alternative strategies that might be used in practice. The first such strategy is to account for all covariates regardless of whether the data suggest they are relevant. As shown in the first column and last row of Table 2.1, fitting the full model yielded only two genes with q-values less than 0.05 for the *Line* test. The second strategy is to ignore all covariates. This is the only strategy available to researchers who do not measure or record covariates, and it might be the most commonly used strategy, considering that many published RNA-seq studies of differential expression do not mention covariates. When the 13 covariates in the RFI RNA-seq analysis were ignored, 251 genes had q-values less than or equal to 0.05 for the *Line* test. Both of these alternative strategies identified far fewer differentially expressed genes than our proposed backward selection procedure. Via simulation, we evaluate the efficacy of these simple strategies relative to our backward selection procedure in the next section.

2.4 Simulation Study

We considered three simulation scenarios described in detail in Sections 2.4.1, 2.4.2, and 2.4.3, respectively. We compared analysis approaches with respect to their ability to identify differentially expressed genes while controlling FDR. Such comparisons require simulated datasets to contain both EE and DE genes. Within each scenario, we varied $\pi_0 =$ the proportion of EE genes over the values 0.6, 0.7, 0.8, and 0.9. Within each scenario and for each value of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$, we

simulated 100 datasets. Each dataset included read counts for 31 pigs and 5000 genes simulated from negative binomial distributions. The log of each negative binomial mean was set to be a linear combination of covariates as in equation (2.1), with $\mathcal{S}$ specifically defined in each scenario. Except where otherwise noted in Section 2.4.3, covariates for the 31 pigs were held fixed at the values observed for the actual RFI data. The true values of partial regression coefficients and negative binomial dispersion parameters were set based on values estimated from the RFI data, and EE genes were established by setting to zero the partial regression coefficient for the *Line* indicator variable as detailed in the following sections.

2.4.1 Simulation Scenario 1: Same Set of Relevant Covariates for Every Gene

The first simulation scenario provides a favorable case for our backward selection procedure in which the same set of covariates is relevant for every gene. As the common set of relevant covariates, we used those identified by our backward selection procedure when applied to the RFI dataset in Section 2.3, i.e., *Line*, *Baso*, *Lymp*, *Mono*, *Neut*, *Concb*, *RINa*, and *Block*. As true parameter values for simulating new data, we used the dispersion parameter estimates and the partial regression coefficient estimates from the fit of the selected model to the RFI data, except that we set partial regression coefficients on the *Line* indicator variable to zero for a subset of genes to permit simulation of EE genes. More specifically, the $\hat{m}_0 = 7795$ least significant partial regression coefficients for *Line* were set to zero, where $\hat{m}_0 = 7795$ is the estimated number of *Line* partial regression coefficients equal to zero when the method of Nettleton et al. (2006) is applied to *Line* p-values from the fit of the selected model to the RFI data. This strategy yielded a parameter vector (consisting of a dispersion parameter and partial regression coefficients) for each of 7795 EE genes and each of $12280 - 7795 = 4485$ DE genes. To simulate any particular dataset for a given value of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$, we randomly sampled $5000 \cdot \pi_0$ EE gene parameter vectors and $5000 \cdot (1 - \pi_0)$ DE gene parameter vectors. The selected parameter vectors and observed covariates for the 31 pigs were used to simulate a 5000×31 dataset of negative binomial read counts. Random

selection of parameters and generation of data was independently repeated 100 times to obtain the 100 datasets for each value of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$ as described in the introduction to Section 2.4.

2.4.2 Simulation Scenario 2: Different Sets of Relevant Covariates for Different Genes

The second simulation scenario is designed to evaluate our backward selection procedure when, contrary to our working assumption, different sets of covariates are relevant for different genes within each dataset. In simulation scenario 2, each dataset was simulated using exactly the same procedure described in Section 2.4.1, except that instead of generating data for all 5000 genes using one set of relevant covariates, data for 1250 genes were simulated from each of four covariate sets. The covariate sets we considered are sets $\mathcal{S}_6$, $\mathcal{S}_7$, $\mathcal{S}_8$, and $\mathcal{S}_9$, which correspond to iterations $\ell = 6, 7, 8$ and 9 form the RFI data analysis is Section 2.3. The largest of these covariate sets ($\mathcal{S}_6$) contains the covariate *RFI* in addition to the covariates considered in Section 2.4.1 (those in $\mathcal{S}_7$). Covariate sets $\mathcal{S}_8$ and $\mathcal{S}_9$ differ from $\mathcal{S}_7$ by the exclusion of covariates *Baso* and both *Baso* and *Mono*, respectively.

2.4.3 Simulation Scenario 3: Orthogonal Covariates

As described in the Appendix, the covariate *RFI* provides a continuous measure of residual feed intake for each of the 31 pigs in the study. Because the LRFI and HRFI lines were created by selecting on residual feed intake for several generations, it is not surprising that the LRFI pigs in our study tend to have lower *RFI* values than the HRFI pigs in our study. Thus, the *RFI* covariate is strongly associated with the factor *Line* in our dataset. This association makes it difficult to distinguish the direct effects of *Line* from the direct effects of *RFI* on transcript abundance levels. To remove this partial confounding in the third simulation scenario, the average *RFI* value for pigs from the LRFI line was subtracted from each LRFI pig's *RFI* value. Likewise, the average *RFI* value for pigs from the HRFI line was subtracted from each HRFI pig's *RFI* value. After these subtractions, the altered *RFI* values sum to zero within each line so that the altered *RFI* variable is

orthogonal to the *Line* factor. The simulation strategy described in Section 2.4.2 was then repeated with the altered *RFI* values in place of the original *RFI* values.

2.4.4 Simulation Results

We analyzed the simulated datasets using model (2.2) with five different strategies for choosing $\hat{\mathcal{S}}^*$: all available covariates (Full), only the factor of primary interest (Line Only), the backward selection procedure with the p.05 measure of covariate relevance (Backward), the backward selection procedure with the GKS measure of covariate relevance, and $\hat{\mathcal{S}}^* = \mathcal{S}^*$, i.e., using the true set of covariates that was actually used to simulate the data for each gene (Oracle). Of course, the Oracle procedure cannot be used in practice, but its inclusion provides a useful reference measure of the performance achieved if covariate selection were perfect.

For all five analysis strategies, the `QuasiSeq` `R` package was used to compute a p-value for testing the significance of the partial regression coefficient on the *Line* indicator variable for each gene. These p-values were converted to q-values (as described in Section 2.2.2), and genes with q-values no larger than 0.05 were declared DE. We evaluated each procedure's performance according to three criteria: the incurred FDR when FDR is nominally controlled at 5%, the number of true positive (NTP) declarations of differential expression, and the partial area under the receiver operating characteristic curve (PAUC) corresponding to false positive rates less than or equal to 0.05. These performance criteria assess error control, power, and the ability to distinguish EE and DE genes from one another, respectively. In all scenarios and for all performance measures, the results for our backward selection procedure with the p.05 variable relevance criterion were very similar to the results when using the GKS variable relevance criterion. To simplify figures, we have shown results only for the simpler p.05 version of backward selection.

A summary of the results for simulation scenario 1 is displayed in the left column of Figure 2.2. All methods provided approximate control of the FDR at or below 5%. The Full approach was slightly conservative while the Line Only approach was very conservative, with actual FDR around 1%. In terms of power for detecting DE genes and the ability to distinguish EE genes from DE

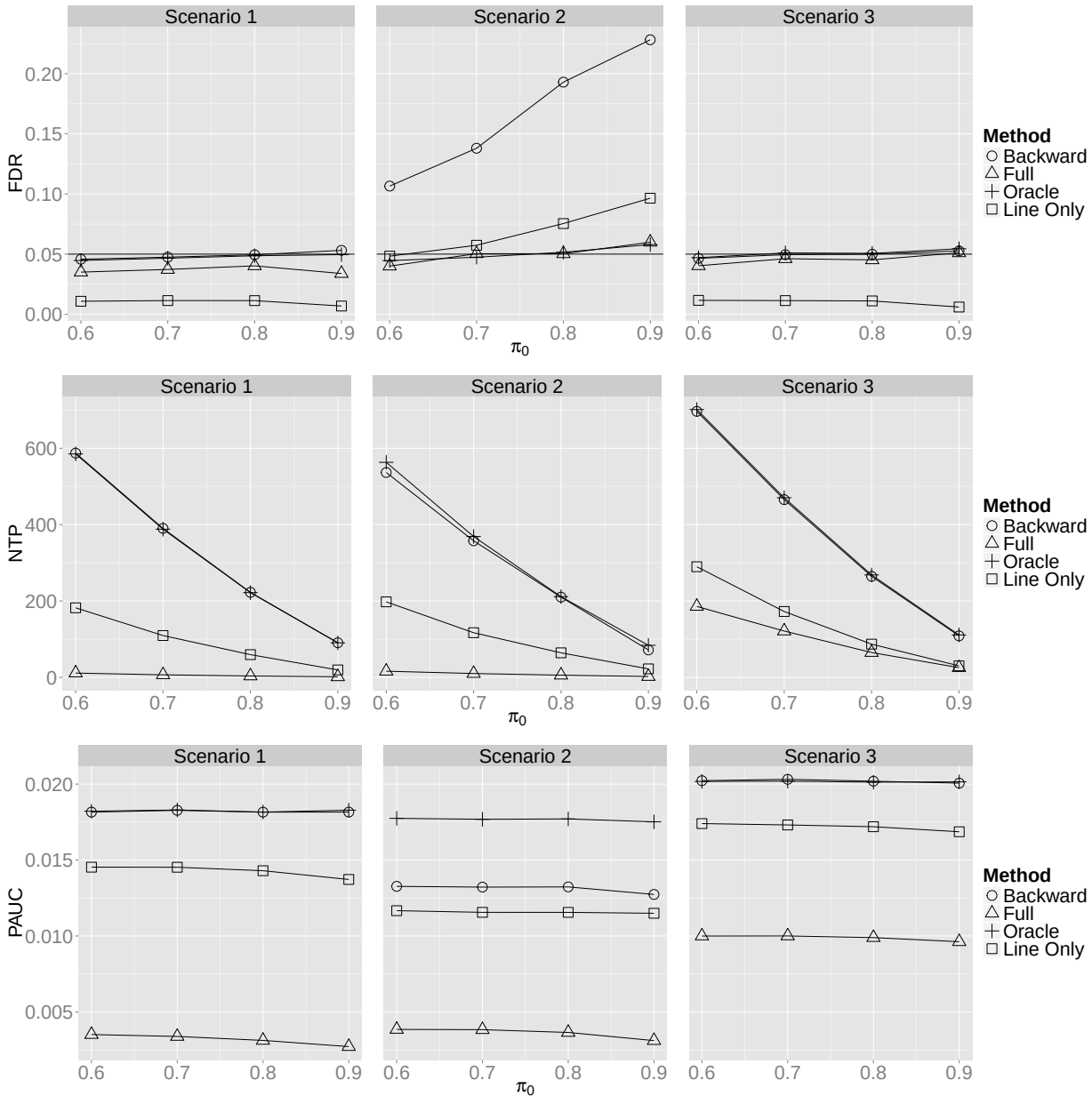

Figure 2.2: Empirical estimates of incurred false discovery rate (FDR), the average number of true positive (NTP) detections of differential expression, and the average partial area under the receiver operating characteristic curve (PAUC) from 100 replicates as a function of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$ for Backward, Full, Line Only, and Oracle methods and all three simulation scenarios. Standard errors of means (not shown to improve clarity of plots) were no larger than 0.0125, 4.6, and 0.00017 for FDR, NTP, and PAUC, respectively.

genes, Backward performed as well as Oracle, while the Full and Line Only procedures exhibited far lower NTP and PAUC on average. The backward selection procedure was able to match the Oracle procedure in this scenario because the correct set of relevant covariates was chosen by backward selection ($\hat{\mathcal{S}}^* = \mathcal{S}^*$) for around 80% of the datasets. When backward selection failed to identify the exact set of relevant covariates ($\hat{\mathcal{S}}^* \neq \mathcal{S}^*$), the selected set was typically a small superset of the true set ($\hat{\mathcal{S}}^* \supset \mathcal{S}^*$) so that the fitted model was correct, though slightly more complicated than necessary due to the inclusion of one or (rarely) more irrelevant covariates.

The results for simulation scenario 2 are summarized in the second column of Figure 2.2. Backward selection matched the Oracle procedure with respect to power (as measured by average NTP) and outperformed all methods except Oracle with respect to PAUC. Backward selection, however, failed to control FDR at 5% for all values of $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$. The incurred FDR rate was more than four times the nominal level when $\pi_0 = 0.9$. The Line Only method also failed to control FDR for $\pi_0 = 0.8$ and 0.9, but the departures from the target 5% rate were not as severe for Line Only as for Backward.

The failure of the backward method to control FDR can be explained as follows. In simulation scenario 2, the true set of covariates is $\mathcal{S}_6$ for 1250 genes, $\mathcal{S}_7$ for 1250 genes, $\mathcal{S}_8$ for 1250 genes, and $\mathcal{S}_9$ for 1250 genes. Despite different sets of relevant covariates for different genes, the backward procedure, by design, selects one common set of covariates for all genes for reasons explained in Section 2.1. Backward selections chose $\mathcal{S}_7$ for more than 90 of the 100 datasets on average across $\pi_0 \in \{0.6, 0.7, 0.8, 0.9\}$. Because $\mathcal{S}_6 \supset \mathcal{S}_7 \supset \mathcal{S}_8 \supset \mathcal{S}_9$, selecting $\mathcal{S}_6$ would guarantee that all relevant covariates were included in the model for each gene. However, $\mathcal{S}_6$ includes *RFI*, which is strongly associated with *Line* as discussed in Section 2.4.3. The lack of orthogonality between the *Line* indicator variable and *RFI* reduces the significance of the *Line* partial regression coefficient in models that include both *Line* and *RFI*. Decreased significance of the *Line* partial regression coefficient reduces the number of *Line* q-values less than or equal to 0.05 and discourages selection of $\mathcal{S}_6$ by our backward selection procedure. For EE genes whose true covariate set is $\mathcal{S}_6$, the partial regression coefficients for *RFI* and *Line* are nonzero and zero, respectively. However, when models

excluding *RFI* are selected (e.g., $\mathcal{S}_7$) and fit to the data, the association between gene expression and *RFI* and between *RFI* and *Line* leads to a nonzero partial regression coefficient for *Line* in the fitted model. In the notation of Section 2.2, we have $\boldsymbol{\beta}_{g1|\mathcal{S}_6} = \mathbf{0}$ and $\boldsymbol{\beta}_{g1|\mathcal{S}_7} \neq \mathbf{0}$ for EE genes whose true covariate set is $\mathcal{S}_6$. When the selected covariate set is $\mathcal{S}_7$ for such genes, the null hypothesis $H_{0g1}^{\mathcal{S}_7} : \boldsymbol{\beta}_{g1|\mathcal{S}_7} = \mathbf{0}$ is correctly rejected, but this leads to a false discovery of differential expression in our simulation set up because $\boldsymbol{\beta}_{g1|\mathcal{S}_6} = \mathbf{0}$. Table 2.2 confirms that the vast majority of false discoveries by the Backward procedure occurred for genes whose relevant covariate set is $\mathcal{S}_6$.

Table 2.2: The average number of false discoveries over 100 replicates as a function of π_0 and the true covariate set used to generate data in simulation scenario 2.

π_0	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$	$\mathcal{S}_9$	Total
0.6	46.57	5.47	5.62	6.44	64.10
0.7	44.40	4.14	3.99	4.85	57.3
0.8	41.44	2.77	2.64	3.45	50.30
0.9	23.58	1.17	1.43	1.44	27.62

Results for simulation scenario 3 are presented in the third column of Figure 2.2. Recall that scenario 3 is identical to scenario 2 except that the strong association between the *RFI* covariate and the *Line* factor was eliminated by centering *RFI* values on zero within each line by subtracting within-line *RFI* averages. The resulting orthogonality between *RFI* and *Line* improved the performance of all methods with respect to all performance criteria when compared to both simulation scenarios 1 and 2. Backward performed as well as Oracle even though the relevant set of covariates differed from gene to gene. For approximately 75% of the datasets, covariate set $\mathcal{S}_6$ was selected so that fitted models included the relevant covariates, along with 0, 1, 2, or 3 extra covariates depending on the gene. The loss of denominator degrees of freedom for including up to three irrelevant covariates was negligible in this case. However, the loss in power was substantial when all covariates were used, as shown by the relatively poor performance of the Full method.

2.5 Discussion

The proposed backward selection algorithm provides a practical method for identifying and controlling for the effects of covariates relevant for all genes in the analysis of RNA-seq data. In the past, we have used visual inspection of p-value histograms (like those in Figure 2.1) to identify and remove irrelevant covariates from models for RNA-seq data. The proposed backward selection algorithm provides a well-defined formalization of this process. This section discusses limitations, variations, and extensions of the backward selection algorithm.

2.5.1 Combining Model Selection and Inference

Caution is in order any time the same dataset is used both to select a model and to perform statistical inference with the selected model (see Miller (2002), for example). We may avoid some problems associated with double use of data because of an important difference between the work we have presented here and traditional work on model selection and inference. While most past work focuses on a single response variable, we combine information from thousands of response variables when choosing the common set of variables to include in the model for each response. Although our backward selection algorithm uses data from all genes, excluding the data from any one gene would be very unlikely to change the set of selected covariates. Thus, we can view the model used to make inferences about any single gene as being selected using data from other genes. This separation between the data used for model selection and data used for inference could be partly responsible for the good inferential performance following backward selection exhibited in the simulation results of Section 2.4.4.

2.5.2 Backward Selection with Other RNA-seq Analysis Methods

The reasoning behind our backward selection algorithm rests on the claim that including irrelevant covariates and excluding relevant covariates in models for gene expression analysis results in power loss for scientific discovery. Support for this claim is given in Section 2.2.2 and in the simulation results of Section 2.4.4. Our argument depends to some extent on the quasi-likelihood

approach implemented in QuasiSeq and does not directly apply to other inference methods that do not account for lack of model fit with quasi-dispersion parameter estimates and do not account for model complexity with denominator degrees of freedom. Thus, further study is required before our backward selection algorithm could be recommended for use with other RNA-seq analysis packages. However, of the many methods available for RNA-seq analysis other than QuasiSeq, one approach does stand out as a good candidate for use with backward selection. The voom approach (Law et al., 2014) in conjunction with the R package `limma` (Ritchie et al., 2015) involves weighted linear model analysis of log-transformed RNA-seq read counts. The limma estimates of linear model error variances are analogous to the quasi-dispersion estimates of QuasiSeq, and both methods of inference involve F statistics whose denominator degrees of freedom are derived from the same basic argument. For these reasons, we expect the proposed backward selection to work well with voom/limma analysis.

2.5.3 Measures of Covariate Relevance

We have suggested two related measures of covariate relevance to use in backward selection. We have found that both measures perform very similarly across the simulation scenarios we considered. The p.05 criterion has an advantage of simplicity but could be criticized because of the somewhat arbitrary 0.05 p-value threshold. Alternative thresholds could be considered, but we do not expect much variation in performance across thresholds near 0.05 because of the similar performance of p.05 and GKS, which is threshold free. Both the p.05 and GKS criteria provide reasonable ways to detect departures from uniformity toward distributions stochastically smaller than uniform, and both criteria produce similar sequences of models that permit effective model selection.

2.5.4 Direct Versus Indirect Associations and Automatic Covariate Inclusion

For model selection, we have proposed choosing the model (from those in the backward selection sequence) that maximizes the number of declarations of differential expression subject to control of FDR at a desired nominal level. Despite the greedy nature of this selection criterion, we found

that the approach worked well except for challenging genes where there is no direct association between gene expression and the primary factor of interest, but rather only indirect association that results from strong association between the primary factor of interest and a covariate that is directly associated with gene expression. In this situation, illustrated with simulation scenario 2 in Section 2.4.2, the proposed backward selection procedure failed to control FDR. However, most false discoveries in this case were not incorrect conclusions if the declarations of differential expression are stated as associations between gene expression and the primary factor of interest while controlling for the effects of the selected covariates. In our analysis of the actual RFI RNA-seq dataset in Section 2.3, we must be careful to acknowledge that some of the gene expression levels declared to be significantly associated with *Line* may be indirectly associated with *Line* and only directly associated with *RFI* or other covariates that were not included in the selected set.

In the RFI application, we are fortunate that either direct or indirect associations between expression and *Line* are of interest. In other applications where researchers are specifically interested in distinguishing direct effects of a primary factor of interest from indirect effects due to a covariate, such covariates should be automatically included in the model. More generally, if scientific questions of interest dictate that one or more covariates be included in the model, the fate of such covariates should not be decided by backward selection; rather, such covariates should be part of every model considered, just as the intercept term was, by default, part of every model we fit to the RFI dataset. The backward selection algorithm's primary purpose is to identify and account for covariates that are not of *a priori* interest but are relevant in the sense that they explain non-negligible residual variation in transcript abundance levels beyond that explained by the primary factor of interest. Accounting for such covariates can boost power for discovery of differential expression that is of primary scientific interest.

2.5.5 Backward Selection to Account for Unobserved Covariates

In contrast to our paper, which has focused on adjusting for the effects of observed covariates, Leek and Storey (2007) and Leek (2014) have proposed surrogate variable analysis (SVA) as a

strategy for dealing with unobserved covariates in differential expression analysis of microarray data and RNA-Seq data, respectively. It is possible to combine SVA with our backward selection strategy to simultaneously account for both observed and unobserved covariates in RNA-seq analysis. Using our RFI RNA-seq dataset as an example, we applied the approach of Leek (2014) as implemented in the `sva` R package available on Bioconductor (Gentleman et al. (2004)). After accounting for the effects of all 14 observed variables in our dataset, SVA detected one unobserved covariate and estimated values for a surrogate variable to be used in place of the unobserved covariate. We then applied our backward selection algorithm as described in Section 2.2.3, except that we included the surrogate variable among our other covariates. The surrogate variable was removed from the model on the fourth iteration, and the final selected model was identical to the model chosen in Section 2.2.3. Although considering unobserved covariates turned out to be irrelevant for the analysis of our example dataset, accounting for such variables may be crucial in other cases.

2.5.6 Backward Selection When Multiple Factors are of Interest

We have described our method for the important special case where a single categorical factor is of primary scientific interest. The backward selection algorithm can be trivially extended to handle cases where a single quantitative variable is of primary interest. If multiple factors (quantitative, categorical, or a combinations of the two) are of interest, there are multiple variations of the algorithm that could be considered. We will highlight two options by focusing on the case where two factors (say A and B) are of interest.

First, suppose the part of the model involving the factors of interest is specified a priori so that backward selection will focus only on eliminating irrelevant covariates from the model. For example, suppose we will include A, B, and $A \times B$ interaction effects in our model regardless of what the data imply about the significance of these effects. To choose what covariates to include in a model with A, B, and $A \times B$ interaction effects, we could apply our backward selection algorithm as before by treating the A, B, and $A \times B$ interaction effects as the effects associated with a single factor of primary interest. In the notation of Section 2.2, we would define $\boldsymbol{x}_{i1}$ and $\boldsymbol{\beta}_{g1|\mathcal{S}}$ so that $\boldsymbol{x}_{i1}'\boldsymbol{\beta}_{g1|\mathcal{S}}$

represents the sum of the appropriate A, B, and $A \times B$ partial regression coefficients for unit i in the model with variables indicated by $\mathcal{S}$. Backward selection could then proceed exactly as defined in Section 2.2.3. The joint significance of the A, B, and $A \times B$ partial regression coefficients would determine when to stop backward selection and which model in the backward selection sequence to choose.

Now suppose the part of the model involving the factors of interest is not fully specified a priori but instead will be chosen based on an examination of the data. For example, suppose the researchers are interested in the main effects of factors A and B but do not want to study $A \times B$ interaction effects unless the data indicate that these effects are important. One strategy is to treat the $A \times B$ interaction effects as we would the effects of any other categorical covariate. Without loss of generality, interaction effects for unit i could be coded in $\boldsymbol{x}_{i2}$ and other covariates specified by $\boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ik}$. Additive effects for A and B and unit i would be coded in $\boldsymbol{x}_{i1}$, and joint significance of the partial regression coefficients on $\boldsymbol{x}_{i2}$ would be used to stop backward selection and choose the model. If the selected model includes $A \times B$ interaction effects, subsequent inferences would be made for each gene using a model that includes A, B, and $A \times B$ interaction effects, along with any other selected covariates. If interaction effects are removed by the backward selection algorithm, then subsequent inferences for each gene could focus on A and B main effects while accounting for the effects of other relevant covariates without further consideration of $A \times B$ interactions.

Even in the relatively simple two-factor scenario described above, there are other strategies worth considering that we have not described here. Determining the relative merits of various strategies requires a more formal problem statement, including a clear description of the tests to be conducted after model selection, the desired error control properties for each set of tests, and priorities for discovery of the multiple types of differential expression that arise when multiple factors are of scientific interest. Such details are beyond the scope of the current article but worth considering in future research.

2.6 Appendix: Description of Variables in the RFI Dataset

$\boldsymbol{x}_{\cdot 1}$ = ***Line*** is the categorical factor of primary scientific interest. Line has two levels, which correspond to the HRFI and LRFI selection lines. Among the 31 pigs in this study, 15 were from the LRFI line and 16 were from the HRFI line.

$\boldsymbol{x}_{\cdot 2}$ = ***RFI*** is a continuous covariate that provides a measure of the residual feed intake for each of the 31 pigs from which blood samples were drawn for RNA-seq analysis. Pigs in the HRFI line tend to have high *RFI* values, while pigs in the LRFI line tend to have low *RFI* values.

$\boldsymbol{x}_{\cdot 3}$ = ***Diet*** is a categorical factor with two levels corresponding to the two diets (high fiber, low energy vs. low fiber, high energy) that were fed to the pigs in this study. Approximately half the pigs within each line were fed each diet. Because RNA-seq analysis was performed on blood samples collected prior to the initiation of the two diets, this factor is not expected to be associated with the transcript abundance levels measured by RNA-seq.

$\boldsymbol{x}_{\cdot 4}$ = ***Baso*** is a continuous covariate that provides a measure of the concentration of basophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 5}$ = ***Eosi*** is a continuous covariate that provides a measure of the concentration of eosinophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 6}$ = ***Lymp*** is a continuous covariate that provides a measure of the concentration of lymphocyte cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 7}$ = ***Mono*** is a continuous covariate that provides a measure of the concentration of monocyte cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 8}$ = ***Neut*** is a continuous covariate that provides a measure of the concentration of neutrophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 9}$ = ***Concb*** is a continuous measure of the RNA concentration in each sample before globin depletion (a step that is necessary to focus sequencing efforts on messenger RNA molecules other than highly abundant globin messenger RNA in each blood sample).

$\boldsymbol{x}_{\cdot 10}$ = ***Conca*** is a continuous measure of the RNA concentration in each sample after globin depletion.

$\boldsymbol{x}_{\cdot 11}$ = ***RINb*** is a continuous measure of RNA integrity within each sample before globin depletion.

$\boldsymbol{x}_{\cdot 12}$ = ***RINa*** is a continuous measure of RNA integrity within each sample after globin depletion.

$\boldsymbol{x}_{\cdot 13}$ = ***Block*** is a categorical factor with four levels corresponding to the four blocks used to organize sample collection and processing. Initially, each block involved eight samples, two for each combination of *Line* and *Diet*. One LRFI sample from the first block was removed from the study due to low-quality RNA.

$\boldsymbol{x}_{\cdot 14}$ = ***Order*** is a categorical factor with eight levels indicating the random order samples were processed within each block.

Acknowledgments

This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), and by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) and the joint National Science Foundation (NSF)/NIGMS Mathematical Biology Program under award number R01GM109458. Yet Nguyen was funded in part by a grant from the Vietnam Education Foundation (VEF). The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of USDA, NSF, NIH, or VEF.

Bibliography

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11(1):94.

Burden, C. J., Qureshi, S. E., and Wilson, S. R. (2014). Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ*, 2:e576.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2):125–153.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):e161.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Lorenz, D. J., Gill, R. S., Mitra, R., and Datta, S. (2014). Using RNA-seq data to detect differentially expressed genes. In *Statistical Analysis of Next Generation Sequencing Data*, pages 25–49. Springer.

Lu, J., Tomfohr, J. K., and Kepler, T. B. (2005). Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6(1):165.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5):1544–6115.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 2 edition.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628.

Nettleton, D., Hwang, J. T. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014a). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014b). The role of spike-in standards in the normalization of RNA-seq. In *Statistical Analysis of Next Generation Sequencing Data*, pages 169–190. Springer.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):479–498.

Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128.

Yanming, D., W, S. D., S, C. J., and H, C. J. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1):1–28.

CHAPTER 3. IDENTIFYING RELEVANT COVARIATES IN RNA-SEQ ANALYSIS BY PSEUDO-VARIABLE AUGMENTATION

A paper in preparation

Yet Nguyen and Dan Nettleton

Abstract

RNA-sequencing (RNA-seq) technology enables the detection of differentially expressed genes, i.e., genes whose mean transcript abundance levels vary across conditions. In practice, an RNA-seq dataset often contains some explanatory variables that will be included in analysis with certainty in addition to a set of covariates that are subject to selection. Some of the covariates may be relevant to gene expression levels, while others may be irrelevant. Either ignoring relevant covariates or attempting to adjust for the effect of irrelevant covariates can be detrimental to identifying differentially expressed genes. We address this issue by proposing a covariate selection method using pseudo-covariates to control the expected proportion of selected covariates that are irrelevant. We show that the proposed method can accurately choose the most relevant covariates while holding the false selection rate below a specified level. We also show that our method performs better than methods for detecting differentially expressed genes that do not take covariate selection into account, or methods that use surrogate variables instead of the available covariates.

3.1 Introduction

A common challenge in analysis of RNA-seq data is detection of differentially expressed (DE) genes, i.e., genes whose mean transcript abundance levels vary across conditions of interest. Typically, some explanatory variables will be included in models used for differential expression analysis

with certainty because of scientific interest or because the design of the study or experiment dictates their inclusion. In many cases, there are other available covariates that are subject to variable selection. These covariates may hold important information about the experimental/observational units, or factors associated with the complex process of measuring RNA transcript abundance levels using RNA-seq technology. Because not all covariates are necessarily important, selecting the subset of covariates that are truly relevant may have a critical role in differential expression analysis. In what follows, we use *included variables* to refer to explanatory variables that are included a priori in models for differential expression analysis. Such included variables are not subject to selection. We use *covariates* to describe variables that are subject to variable selection.

Nguyen et al. (2015) addressed the covariate selection problem in the context of RNA-seq differential expression analysis by proposing a backward selection strategy to choose the most relevant covariates. They showed that their method works well when the included variables are uncorrelated or weakly correlated with one or more covariates. However, if one or more covariates are strongly associated with the included variables, their method fails to detect the truly relevant covariates and may result in the misidentification of spurious signals, resulting in the failure to control false discovery rate (FDR).

In this paper, we introduce a new covariate selection strategy in RNA-seq analysis that can overcome the limitation of the aforementioned method. In particular, we demonstrate that our method can accurately detect the most relevant covariates even when relevant covariates are strongly correlated with the included variables.

Our method is based on a variable selection approach originally designed to control the false selection rate (FSR) in linear regression for one response variable (Wu et al., 2007). Wu et al. (2007)'s method involves augmenting the set of available covariates with pseudo-covariates that, by construction, are known to be irrelevant variables that should not be included in the model for the single response variable. By studying the propensity of a selection method to include pseudo-covariates among the selected set of covariates, it is often possible to tune selection to control FSR below a specified threshold. The method of Wu et al. (2007) uses backward (or forward) selection

with the p-value associated with each covariate as the measure of covariate relevance. In this paper, we extend the method of Wu et al. (2007) to multiple response variables so that the method can be used in differential expression analysis, where there is one response variable for each of thousands of genes. Instead of having only one p-value, a covariate in the case of multiple response variables has a vector of p-values, with one p-value for each response variable. To take the whole vector of p-values associated with a covariate into account, we propose a simple covariate relevance measure that can be described, informally, as the ratio of the number of small p-values to the number of large p-values. For a given covariate, the greater the ratio, the more relevant the covariate is judged to be.

We illustrate the performance of our method using a backward selection procedure and the differential expression analysis method voom (Law et al., 2014). If any of the covariates are uncorrelated or weakly correlated with the included variables, the performance of our method is similar to that of Nguyen et al. (2015). On the other hand, if there are covariates strongly correlated with the included variables, we show that our method outperforms the method of Nguyen et al. (2015). We also compare the differential expression analysis produced by our method with several alternatives, such as the method of Nguyen et al. (2015), methods including or excluding all covariates, and methods using surrogate variables (Leek and Storey, 2007; Lee et al., 2017). As a result of accounting for truly relevant covariates, our approach outperforms others in terms of power, FDR control, and ability to distinguish the true and false signals with respect to the included variables.

The paper is organized as follows. In Section 3.2, we first give general preliminaries about RNA-seq dataset, available covariates, and the voom method; then, we explain our covariate selection method using pseudo-covariates aiming to control FSR. In Section 3.3, we apply our method to analyze an RNA-seq dataset with many covariates from a residual feed intake (RFI) experiment conducted to find genes differentially expressed between two RFI lines. In Section 3.4, we conduct an extensive data-driven simulation to investigate the performance of our method relative to competing methods in the ability to select truly relevant covariates and to identify DE genes. Finally, Section 3.5 is devoted to a conclusion and discussion.

3.2 Methods

3.2.1 Notations and Preliminaries

Consider the analysis of m genes using RNA-seq read count data from n subjects. Let $\boldsymbol{x}_{i\cdot} = (\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{ij}, \ldots, \boldsymbol{x}'_{ik})'$ be the vector of known k explanatory variable values for subject $i = 1, \ldots, n$, and let $\boldsymbol{x}_{\cdot j}$ be the set of column vectors corresponding to the j-th variable for $j = 1, \ldots, k$. Vector $\boldsymbol{x}_{ij}$ for a continuous variable j has only one element while vector $\boldsymbol{x}_{ij}$ corresponding to a categorical variable j consists of indicator variable values with one less indicator than the number of levels of the categorical variable. Without loss of generality, we assume $\{\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot \ell}\}$ to be the included variables, i.e., the subset of explanatory variables that are not subject to selection. These variables will always be included in the model because of scientific interest, experimental design or study design considerations. The covariates $\{\boldsymbol{x}_{\cdot \ell+1}, \ldots, \boldsymbol{x}_{\cdot k}\}$ are subject to variable selection. These covariates are measured during the experiment and may or may not be relevant to the expression values.

Let c_{gi} be the read count from gene g and sample i, and let $R_i = \sum_{g=1}^{m} c_{gi}$, which is known as the library size of RNA-seq sample for subject i. The log-counts per million (Law et al., 2014) is defined as

$$y_{gi} = \log_2 \left(\frac{c_{gi} + 0.5}{R_i + 1} \times 10^6 \right). \tag{3.1}$$

The counts are offset away from zero by 0.5 to avoid taking the log of zero, and to reduce the variability of y_{gi} for lowly expressed genes, while the library size is offset by 1 to ensure that $(c_{gi} + 0.5)/(R_i + 1)$ is strictly greater than zero and less than 1. In general, R_i can be any normalization offset computed for subject i. The normalization offsets account for random differences in the thoroughness of sequencing across the RNA-seq samples. Many normalization offsets have been proposed in the literature (see, e.g., Marioni et al. (2008), Mortazavi et al. (2008), Robinson and Oshlack (2010), Anders and Huber (2010), Bullard et al. (2010), Risso et al. (2014a), Risso et al. (2014b), and references therein). Throughout this paper, we set R_i to be the 0.75 quantile of RNA-seq sample read counts from subject i according to the recommendation of Bullard et al. (2010).

With this choice of normalization offsets, the y_{gi} values are no longer "counts per million mapped read" on the log scale, but this interpretation is irrelevant for the differential expression analysis that is the focus of our work. Henceforth, we use $\boldsymbol{c}_g = (c_{g1}, \ldots, c_{gn})'$ and $\boldsymbol{y}_g = (y_{g1}, \ldots, y_{gn})'$ to denote the vector of count values and the vector of log-counts values for gene g, respectively.

3.2.2 The voom Procedure

There are many differential expression analysis methods developed for RNA-seq data such as QuasiSeq (Lund et al., 2012; Lun et al., 2016), edgeR (McCarthy et al., 2012), DESeq2 (Love et al., 2014), and voom (Law et al., 2014) among many others. In this paper, we use the voom method because of its FDR control, power, and computational speed. voom is based on linear model analysis that incorporates the mean-variance relationship of the log-counts by introducing a precision weight for each observation according to the following algorithm.

1. First, let
$$y_{gi} = \log_2\left(\frac{c_{gi} + 0.5}{R_i + 1} \times 10^6\right)$$
where c_{gi} is the read count from gene g and sample i, R_i is the 0.75 quantile of RNA-seq sample read counts from subject i.

2. Let $\mathcal{S}$ represent a subset of $\{1, \ldots, k\}$ that contains $\{1, \ldots, l\}$, the indices of the ℓ explanatory variables always included in the model. For each gene g, assume a linear model
$$y_{gi} = \beta_{g0|\mathcal{S}} + \sum_{j \in \mathcal{S}} \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{gj|\mathcal{S}} + \varepsilon_{gi|\mathcal{S}}, \quad \varepsilon_{gi|\mathcal{S}} \sim \mathcal{N}(0, \sigma^2_{g|\mathcal{S}}), \quad g = 1, \ldots, m;\ i = 1, \ldots, n \tag{3.2}$$
or equivalently, in vector form,
$$\boldsymbol{y}_g = \boldsymbol{X}_{\mathcal{S}}\boldsymbol{\beta}_{g|\mathcal{S}} + \boldsymbol{\varepsilon}_{g|\mathcal{S}}.$$
where $\boldsymbol{X}_{\mathcal{S}}$ is the design matrix consisting of an intercept column $\mathbf{1}$ and all columns in $\boldsymbol{x}_{\cdot j}$ for all $j \in \mathcal{S}$; $\boldsymbol{\beta}_{g|\mathcal{S}}$ is the vector of regression coefficients consisting of $\beta_{g0|\mathcal{S}}$ and all $\boldsymbol{\beta}_{gj|\mathcal{S}}$ for $j \in \mathcal{S}$.

3. Let $\widehat{\boldsymbol{\beta}}_{g|\mathcal{S}} = (\boldsymbol{X}'_{\mathcal{S}}\boldsymbol{X}_{\mathcal{S}})^{-1}\boldsymbol{X}'_{\mathcal{S}}\boldsymbol{y}_g$ and $s_{g|\mathcal{S}} = \sqrt{\frac{\left(\boldsymbol{y}_g - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{g|\mathcal{S}}\right)'\left(\boldsymbol{y}_g - \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{g|\mathcal{S}}\right)}{n - \text{rank}(\boldsymbol{X}_{\mathcal{S}})}}$ be the ML and REML estimates of $\boldsymbol{\beta}_{g|\mathcal{S}}$ and $\sigma_{g|\mathcal{S}}$, respectively. Let $\widehat{\boldsymbol{y}}_g = \boldsymbol{X}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{g|\mathcal{S}}$.

4. Let $\tilde{c}_g = \frac{1}{n}\sum_{i=1}^{n} y_{gi} + \frac{1}{n}\log_2\left(\prod_{i=1}^{n}(R_i+1)\right) - \log_2(10^6)$ be the mean log-count value for each gene g.

5. Let $\text{lo}(\cdot)$ be the predictor obtained by fitting a LOWESS regression (Cleveland, 1979) of $s_{g|\mathcal{S}}^{1/2}$ on $\tilde{c}_g$. The precision weight for y_{gi} is calculated by

$$w_{gi} = \left[\text{lo}\left(\widehat{y}_{gi} + \log_2(R_i+1) - \log_2(10^6)\right)\right]^{-4}.$$

The normalized log-counts and their associated precision weights then enter the limma pipeline (Smyth, 2004; Ritchie et al., 2015) for downstream analysis, including shrinkage estimation of error variances, calculation of moderated t-statistics or moderated F-statistics for partial regression coefficients. These statistics are then compared to a central t or F distribution to obtain p-values. These p-values are converted to q-values by the method of Storey (2002). When computing q-values, we need an estimate of m_0, the number of true null hypotheses among all null hypotheses tested. We use the histogram-based method by Nettleton et al. (2006) to estimate m_0. Desirable theoretical characteristics of a closely related histogram-based approach were demonstrated by Liang and Nettleton (2012).

3.2.3 Measure of Covariate Relevance

For a given $j \in \mathcal{S} \setminus \{1,\ldots,\ell\}$ where $\{1,\ldots,\ell\} \subseteq \mathcal{S} \subseteq \{1,\ldots,k\}$, let $\boldsymbol{p}_{j|\mathcal{S}}$ be the vector of m p-values obtained by testing $H_{0gj|\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \boldsymbol{0}$ for each gene $g = 1,\ldots,m$. Assuming (3.2), we consider covariate j irrelevant if

$$H_{0gj} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \boldsymbol{0} \text{ is true for all } g = 1,\ldots,m. \tag{3.3}$$

If (3.3) holds, each element of $\boldsymbol{p}_{j|\mathcal{S}}$ is uniformly distributed on $(0,1)$ whenever the test used to produce the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ has size equal to the significance level for all significance levels in

$(0, 1)$. If the test used to produce the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ is unbiased for all significance levels in $(0, 1)$, then an element of $\boldsymbol{p}_{j|\mathcal{S}}$ corresponding to a false null hypothesis will have a distribution stochastically smaller than uniform(0, 1) and a density decreasing on the interval $(0, 1)$. Therefore, the empirical distribution of the elements of $\boldsymbol{p}_{j|\mathcal{S}}$ provides useful information about the relevance of covariate j. An empirical distribution of $\boldsymbol{p}_{j|\mathcal{S}}$ close to the uniform(0, 1) or stochastically larger than the uniform(0, 1) implies little relevance while an empirical distribution with clear excess of small p-values relative to the uniform(0, 1) implies more relevance of covariate j for at least some appreciable number of genes.

There are many different ways to formally measure the relevance of covariate j using its $\boldsymbol{p}_{j|\mathcal{S}}$. Nguyen et al. (2015) considered two measurements: 1) the number of elements of $\boldsymbol{p}_{j|\mathcal{S}}$ less than 0.05 and 2) the Kolmogorov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1948) measuring the discrepancy between the uniform(0, 1) distribution and the Grenander estimate (Grenander, 1956) of a non-increasing distribution computed from the elements of $\boldsymbol{p}_{j|\mathcal{S}}$. These two measurements are natural choices given the aforementioned behavior of the empirical distribution of p-values. In this paper, we propose another intuitive covariate relevance measure which is formally stated in the following definition.

Definition 3.2.1 *Let $\{1, \ldots, \ell\} \subseteq \mathcal{S} \subseteq \{1, \ldots, k\}$. For any $j \in \mathcal{S}$, let $\boldsymbol{p}_{j|\mathcal{S}}$ be the p-values vector obtained by testing $H_{0gj|\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \mathbf{0}$ for each gene $g = 1, \ldots, m$. A relevance measure for covariate j is defined as*

$$r(\boldsymbol{p}_{j|\mathcal{S}}) = \frac{\text{Card}\{g : p_{gj|\mathcal{S}} \leq 0.05, 1 \leq g \leq m\}}{\max\{\text{Card}\{g : p_{gj|\mathcal{S}} > 0.75, 1 \leq g \leq m\}/5, 1\}}. \tag{3.4}$$

The measurement r defined in Definition 3.2.1 is also motivated from the relationship between the relevance level of covariate j and the behavior of the empirical distribution of p-values obtained from testing $H_{0gj|\mathcal{S}} : \boldsymbol{\beta}_{gj|\mathcal{S}} = \mathbf{0}$. We use the max operator in the denominator of r to avoid division by zero. It is clear that r is non-negative. If $r(\boldsymbol{p}_{j|\mathcal{S}}) \leq 1$, the number of large p-values exceeds the number of small p-values, which suggests covariate j is irrelevant or less important. On the other hand, if $r(\boldsymbol{p}_{j|\mathcal{S}}) >> 1$, the number of small p-values is much greater than the number of large p-values, which suggests covariate j is relevant or highly important (see, Fig. 3.1). In other

words, the greater $r(\boldsymbol{p}_{j|\mathcal{S}})$ is, the more relevant covariate j is; the smaller $r(\boldsymbol{p}_{j|\mathcal{S}})$ is, the less relevant covariate j is. The comparison of r value to 1 intuitively depicts the relevance level of the covariate and this measurement can be used easily in combination with the FSR variable selection method (Wu et al., 2007) which we are going to review in the next subsection. In what follows, we drop the indices $\mathcal{S}$ and $|\mathcal{S}$ to simplify notations, and we bear in mind that the inference about any particular covariate is conducted by conditioning on the other explanatory variables included in the model.

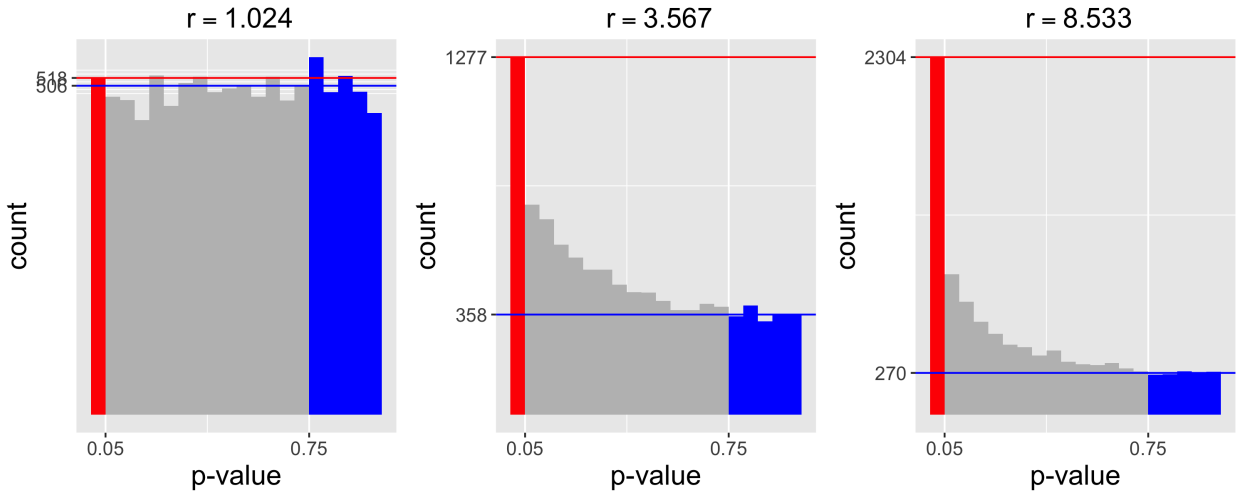

Figure 3.1: An example showing covariate relevance level measured by r function. Each subplot represents (from left to right, respectively) an instance of an irrelevant ($r = 1.024$), a relevant ($r = 3.567$), and a highly relevant ($r = 8.533$) covariate.

3.2.4 False Selection Rate Variable Selection Method

Commonly used approaches to select an important subset of variables in general multiple regression problems include all subset, backward selection, forward selection, and step-wise selection using some measurement of variable importance. These methods produce a number of candidate subsets, then an appropriate selection criterion is utilized to determine the optimal one. Selection criteria are usually based on minimizing prediction error or minimizing some information criteria such as AIC, BIC, etc. In addition, there are many other variable selection methods using regu-

larization, see, e.g., LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Least Angle Regression (Efron et al., 2004), etc.

A good variable selection procedure will on average include a high percentage of the relevant covariates and a low percentage of the irrelevant covariates. The false selection rate (FSR) method proposed by Wu et al. (2007) uses a novel criterion aiming to control the average proportion of selected variables that are irrelevant. Their method is based on a simple idea that the tendency of a variable selection method to overfit or underfit can be revealed by the use of pseudo-covariates. For completeness, we review in the next subsection the detail of their method, adapted for the gene expression data. While the original work in Wu et al. (2007) was demonstrated for a forward selection strategy, in this paper, we describe the method for a backward selection strategy, for the purpose of illustrating the FSR method in RNA-seq analysis.

3.2.4.1 FSR for Backward Selection

For simplicity, let $\boldsymbol{C}$ be the matrix of counts RNA-seq data with one row for each gene and one column for each sample, i.e., $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m]'$. Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, where $\boldsymbol{X}_1 = [\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot \ell}]$, and $\boldsymbol{X}_2 = [\boldsymbol{x}_{\cdot \ell+1}, \ldots, \boldsymbol{x}_{\cdot k}]$. The backward selection procedure starts by fitting the voom procedure as in Section 3.2.2 to $\boldsymbol{C}$ and $\boldsymbol{X}$. Using the measure r described in Section 3.2.3, the least relevant variable (its r value is recorded) in $\boldsymbol{X}_2$ is dropped, and the resulting reduced model is fit again using voom. The process continues until the last variable in $\boldsymbol{X}_2$ is dropped. This backward selection procedure produces a sequence of increasingly smaller subsets of variables, starting with all variables in $\boldsymbol{X}$ and progressing, one removed covariate (in $\boldsymbol{X}_2$) at a time, down to $\boldsymbol{X}_1$. For a given level-to-leave α, let $BS((\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2), \alpha)$ denote the subset of $\boldsymbol{X}_2$ selected by this backward selection. Define $S(\alpha) = \text{Card}\{BS((\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2), \alpha)\}$. Then $S(\alpha) = I(\alpha) + U(\alpha)$, where $U(\alpha)$, $I(\alpha)$ denote the number of selected irrelevant and relevant covariates, respectively.

The method of Wu et al. (2007) aims at on average at most a small proportion of covariates included in a model to be irrelevant. To do this, Wu et al. (2007) defined two FSR functions as

$$\gamma_{ER}(\alpha) = E\left(\frac{U(\alpha)}{1 + S(\alpha)}\right) = E\left(\frac{U(\alpha)}{1 + I(\alpha) + U(\alpha)}\right) \tag{3.5}$$

and

$$\gamma_{RE}(\alpha) = \frac{E(U(\alpha))}{E(1+S(\alpha))} = \frac{E(U(\alpha))}{E(1+I(\alpha)+U(\alpha)))}. \tag{3.6}$$

The constant 1 is added to $S(\alpha)$ in (3.5) and (3.6) primarily because most models have an intercept, and also because it avoids division by zero. The FSR variable selection method aims to determine α_* such that $\gamma_\cdot(\alpha_*) \leq \gamma_0$ for some pre-specified FSR threshold γ_0, say $\gamma_0 = 0.05$; where $\gamma_\cdot$ denotes either γ_{RE} or γ_{ER}. Formally, α_* is defined as

$$\alpha_* = \inf_{\alpha}\{\alpha : \gamma_\cdot(\alpha) \leq \gamma_0\}.$$

Because $\gamma_\cdot(\cdot)$ is unknown, α_* cannot be determined directly. Wu et al. (2007) showed that it can be estimated approximately using Monte-Carlo generated pseudo-covariates as follows.

For some integer number B and each $b = 1, \ldots, B$, suppose that $\boldsymbol{Z}_b$ is a set of k_P pseudo-covariates that are randomly generated to be independent of the response variables $\boldsymbol{C}$. The backward selection procedure described previously is applied to $\boldsymbol{C}$ and $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Z}_p\}$ where now $\{\boldsymbol{X}_2, \boldsymbol{Z}_p\}$ is the set of covariates that are subject to variable selection. Let $S_{P,b}(\alpha)$ be the total number of selected covariates. Let $I_{P,b}(\alpha)$ and $U_{P,b}(\alpha)$ be the number of selected relevant and irrelevant covariates, respectively. Let $U^*_{P,b}(\alpha)$ be the number of selected pseudo-covariates. Then $S_{P,b}(\alpha) = I_{P,b}(\alpha) + U_{P,b}(\alpha) + U^*_{P,b}(\alpha)$. To estimate FSR $\gamma_\cdot(\alpha)$ using these Monte-Carlo samples of pseudo-covariates, Wu et al. (2007) assumed further that

(A1) $E(U(\alpha)) = E(U_{P,b}(\alpha)) = E(U^*_{P,b}(\alpha))k_U/k_P$, where k_U is the unknown number of irrelevant covariates.

(A2) $E(I_{P,b}(\alpha)) = E(I(\alpha))$.

Assumption (A1) states that on average real irrelevant covariates and pseudo-covariates have the same probability of being selected. Assumption (A2) states that on average the truly relevant covariates have the same probability of being selected whether or not pseudo-covariates are present. These assumptions will also serve as guiding principles for generating pseudo-covariates, estimating FSR and estimating α_*. In next subsections, we describe how FSR method works using each of the FSR functions.

3.2.4.2 FSR Method Based on Estimating $\gamma_{RE}(\alpha)$

First, define

$$\gamma_{RE,P}(\alpha) = \frac{E(U^*_{P,b}(\alpha))}{E(1 + S_{P,b}(\alpha))}.$$

Then using assumptions (A1) and (A2), we have

$$\begin{aligned}
\gamma_{RE,P}(\alpha) &= \frac{E(U^*_{P,b}(\alpha))}{E(1 + I_{P,b}(\alpha) + U_{P,b}(\alpha) + U^*_{P,b}(\alpha))} \\
&= \frac{E(U^*_{P,b}(\alpha))}{1 + E(I_{P,b}(\alpha)) + E(U_{P,b}(\alpha)) + E(U^*_{P,b}(\alpha))} \\
&= \frac{k_P/k_U E(U(\alpha))}{1 + E(I(\alpha)) + E(U(\alpha)) + k_P/k_U E(U(\alpha))} \\
&= \frac{k_P/k_U E(U(\alpha))}{E(1 + I(\alpha) + U(\alpha)) + k_P/k_U E(U(\alpha))} \\
&= \frac{k_P/k_U \frac{E(U(\alpha))}{E(1+I(\alpha)+U(\alpha))}}{\frac{E(1+I(\alpha)+U(\alpha))+k_P/k_U E(U(\alpha))}{E(1+I(\alpha)+U(\alpha))}} \\
&= \frac{k_P/k_U \gamma_{RE}(\alpha)}{1 + k_P/k_U \gamma_{RE}(\alpha)} \\
&= \frac{k_P \gamma_{RE}(\alpha)}{k_P \gamma_{RE}(\alpha) + k_U}.
\end{aligned} \tag{3.7}$$

Moreover, $\gamma_{RE,P}(\alpha)$ is estimated by

$$\widehat{\gamma}_{RE,P}(\alpha) = \frac{\bar{U}^*_P(\alpha)}{1 + \bar{S}_P(\alpha)}, \tag{3.8}$$

where

$$\bar{U}^*_P(\alpha) = B^{-1} \sum_{b=1}^{B} U^*_{P,b}(\alpha), \bar{S}_P(\alpha) = B^{-1} \sum_{b=1}^{B} S_{P,b}(\alpha)$$

Using (3.7) and (3.8), $\gamma_{RE}(\alpha)$ is estimated by the solution to the following equation

$$\widehat{\gamma}_{RE,P}(\alpha) = \frac{k_P \gamma_{RE}(\alpha)}{k_P \gamma_{RE}(\alpha) + k_U},$$

whose the right-hand side is a monotone increasing function of $\gamma_{RE}(\alpha)$. Therefore, if k_U is known, for a given FSR level γ_0, an estimate $\widehat{\alpha}_*$ of α_* can be obtained as

$$\widehat{\alpha}_* = \inf_{\alpha} \left\{ \alpha : \widehat{\gamma}_{RE,P}(\alpha) \leq c := \frac{k_p \gamma_0}{k_p \gamma_0 + k_U} \right\}. \tag{3.9}$$

In general, we usually do not know k_U. This fact is similar to the situation in multiple hypothesis testing in that we usually do not know the number of true null hypotheses. Therefore, a reliable estimate of k_U is important for the estimation of FSR. Wu (2004) proposed an iterative algorithm to obtain an estimate of k_U which in turn is used to estimate α_*. The algorithm is described as follows.

FSR Procedure Using $\gamma_{RE}(\alpha)$

1. Pick a target false selection rate γ_0, e.g., $\gamma_0 = 0.05$.

2. Generate sets of k_P pseudo-covariates B times. For α in $\mathcal{A}$, e.g., $\mathcal{A} = \{1, 1.01, 1.02, \cdots, 10\}$, counts the average total number of selected pseudo-covariates $\bar{U}^*(\alpha)$ and the average number of selected variables $\bar{S}_P(\alpha)$, then calculate

$$\widehat{\gamma}_{RE,P}(\alpha) = \frac{\bar{U}_P^*(\alpha)}{1 + \bar{S}_P(\alpha)}.$$

3. Obtain an initial value cut-off $c^{(0)}$ from the formula

$$c^{(0)} = \frac{k_P\gamma_0}{k_P\gamma_0 + k_T},$$

where k_T is the number of real covariates considered for selection. Define $\widehat{\alpha}_*^{(0)}$ as follows

$$\widehat{\alpha}_*^{(0)} = \min\{\alpha : \widehat{\gamma}_{RE,P}(\alpha) \le c^{(0)},\ \alpha \in \mathcal{A}\}.$$

4. Run backward selection on the original set of covariates X_2 without pseudo-covariates using the significance-level-to-leave $\widehat{\alpha}_*^{(0)}$. Denote the size of the selected model by $\widehat{k}_I^{(0)}$ and set $\widehat{k}_U^{(0)} = k_T - \widehat{k}_I^{(0)}$.

5. Update the cut-off by

$$c^{(1)} = \frac{k_P\gamma_0}{k_P\gamma_0 + \widehat{k}_U^{(0)}},$$

and then find

$$\widehat{\alpha}_*^{(1)} = \min\{\alpha : \widehat{\gamma}_{RE,P}(\alpha) \le c^{(1)},\ \alpha \in \mathcal{A}\}.$$

6. Go back to Step 4 and iterate until there is no change in $\widehat{k}_U^{(i)}$. The final $\widehat{\alpha}_*^{(i)}$ is used in a final backward selection on the original set of data.

3.2.4.3 FSR Method Based on Estimating $\gamma_{ER}(\alpha)$

Similar to the FSR method based on estimating $\gamma_{RE}(\alpha)$ using pseudo-covariates, Wu et al. (2007) defined

$$\begin{aligned}\gamma_{ER,P}(\alpha) &= \frac{E(U^*_{P,b}(\alpha))}{1+S(\alpha)} \\ &= \frac{k_P/k_U E(U(\alpha))}{1+S(\alpha)} \\ &\approx k_P/k_U E\left(\frac{U(\alpha)}{1+S(\alpha)}\right) \\ &\approx k_P/k_U \gamma_{ER}(\alpha). \end{aligned} \tag{3.10}$$

On the other hand, $\gamma_{ER,P}(\alpha)$ is estimated by

$$\widehat{\gamma}_{ER,P}(\alpha) = \frac{\bar{U}^*_P(\alpha)}{1+S(\alpha)}. \tag{3.11}$$

Combining (3.10) and (3.11), $\gamma_{ER}(\alpha)$ is estimated by the solution to the following equation

$$\widehat{\gamma}_{RE,P}(\alpha) = \frac{k_P}{k_U}\gamma_{ER}(\alpha).$$

Therefore, if k_U is known, for a given FSR level γ_0, an estimate of α_* is obtained by

$$\widehat{\alpha}_* = \inf_{\alpha}\left\{\alpha : \widehat{\gamma}_{ER,P}(\alpha) \le c := \frac{k_P}{k_U}\gamma_0\right\}.$$

Then, an FSR algorithm for estimation of k_U and α_* using γ_{ER} is described as follows.

FSR Procedure Using $\gamma_{ER}(\alpha)$

1. Pick a target false selection rate γ_0, e.g., $\gamma_0 = 0.05$.

2. Generate sets of k_P pseudo-covariates B times. For α in $\mathcal{A}$, e.g., $\mathcal{A} = \{1, 1.01, 1.02, \cdots, 10\}$, counts the average number of selected pseudo-covariates $\bar{U}^*(\alpha)$ and the average total number selected variables $\bar{S}_P(\alpha)$, then calculate

$$\widehat{\gamma}_{ER,P}(\alpha) = \frac{\bar{U}^*_P(\alpha)}{1+S(\alpha)}.$$

3. Obtain an initial value cut-off $c^{(0)}$ from the formula

$$c^{(0)} = \frac{k_P \gamma_0}{k_T},$$

where k_T is the number of real covariates considered for selection. Define $\widehat{\alpha}_*^{(0)}$ as follows

$$\widehat{\alpha}_*^{(0)} = \min\{\alpha : \widehat{\gamma}_{ER,P}(\alpha) \leq c^{(0)},\ \alpha \in \mathcal{A}\}.$$

4. Run backward selection on the original set of covariates $\boldsymbol{X}_2$ without pseudo-covariates using the significance-level-to-leave $\widehat{\alpha}_*^{(0)}$. Denote the size of the selected model by $\widehat{k}_I^{(0)}$ and set $\widehat{k}_U^{(0)} = k_T - \widehat{k}_I^{(0)}$.

5. Update the cut-off by

$$c^{(1)} = \frac{k_P \gamma_0}{\widehat{k}_U^{(0)}},$$

and then find

$$\widehat{\alpha}_*^{(1)} = \min\{\alpha : \widehat{\gamma}_{ER,P}(\alpha) \leq c^{(1)},\ \alpha \in \mathcal{A}\}.$$

6. Go back to Step 4 and iterate until there is no change in $\widehat{k}_U^{(i)}$. The final $\widehat{\alpha}_*^{(i)}$ is used in a final backward selection on the original set of data.

3.2.4.4 Pseudo-covariate Generation

As mentioned previously, conditions (A1) and (A2) provide guidance for generating pseudo-covariates. Pseudo-covariates will be generated so that the average inclusion probabilities of relevant covariates are approximately equal with data $(\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2)$ and $(\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2, \boldsymbol{Z})$, and the average inclusion probabilities of irrelevant covariates (real and pseudo) are approximately equal with data $(\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2)$ and $(\boldsymbol{C}; \boldsymbol{X}_1; \boldsymbol{X}_2, \boldsymbol{Z})$, where $\boldsymbol{Z}$ is a set of k_P randomly generated pseudo-covariates. Based on these principles, Wu et al. (2007) proposed four different methods to generate pseudo-covariates. In the first method, entries of the $n \times k_P$ matrix $\boldsymbol{Z}$ are independently and identically distributed $N(0, 1)$; in the second method, the n rows of $\boldsymbol{Z}$ are obtained by randomly permuting

the rows and the k_P columns of $\boldsymbol{X}$. In both methods, the pseudo-covariates are stochastically uncorrelated with the original explanatory variables; while in the second method, the pseudo-covariates have the same distribution as a subset of the explanatory variables. The third and the fourth methods are variants of the first two methods in which $\boldsymbol{Z}$ is replaced by $(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}})\boldsymbol{Z}$, where $\boldsymbol{H}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. The variants are such that the pseudo-covariates have the sample means and sample correlations with the explanatory variables identically equal to 0. Note that the variants are only possible when the rank of the linear space generated by the explanatory variables (including intercept, the set of included covariates, and the covariates subject to variable selection) is smaller than the number of RNA-seq samples – the case when a backward selection strategy can be applied. If the number of explanatory variables is large, a forward selection strategy can be implemented together with FSR method, however, only the first two pseudo-covariates generating methods can be used. We call the first two pseudo-covariate generating methods WN (white noise $N(0,1)$) and RX (permuting rows of $\boldsymbol{X}$), their variant versions OWN and ORX, respectively.

3.3 Real Data Analysis

Now we apply the FSR variable selection method to the RFI RNA-seq dataset. This dataset has been used and described in Nguyen et al. (2015). For completeness, we recall the details of the RFI data set in this article.

Residual feed intake (RFI) is an important quantitative trait measuring feed efficiency. RFI is calculated as the difference between the observed feed intake and the expected feed intake considering the animal growth and size. Pigs with low RFI tend to eat more efficiently than normal pigs, while pigs with high RFI eat less efficiently than normal pigs. Researchers are interested in finding genes whose expression levels differ between two lines of pigs, high RFI line and low RFI line. The analysis is more complicated because of the presence of heterogeneity among pigs, among blood samples extracted from pigs, and among the processed and measured RNA samples derived from the blood samples. The heterogeneity is shown in associated covariates that are measured along with RNA-seq samples such as *Diet, RFI, Concb, Conca, RINa, RINb, Lymp, Neut, mono,*

Eosi, Baso, Block, and *Order* (described in detail in the Appendix). In summary, the RFI RNA-seq dataset we are analyzing consists of 11280 genes each have average read counts of at least 8 and no more than 27 zero counts out of 31 pigs. The included variable is *Line* with two levels, low RFI and high RFI. *Line* is the only variable not subject to variable selection. The available covariates subject to variable selection are *Diet* (2 levels), *Order* (8 levels), *Block* (4 levels), *RINa*, *RINb*, *Conca*, *Concb*, *Lymp*, *Mono*, *Baso*, *Eosi*, *Neut*, and *RFI* (continuous covariates). Generally speaking,

$$\boldsymbol{X}_1 = \{Line\}$$

$$\boldsymbol{X}_2 = \{Diet, Order, Block, RINa, RINb, Conca, Concb, Lymp, Mono, Baso, Eosi, Neut, RFI\}.$$

Before conducting differential expression analysis, we apply the FSR variable selection method to this dataset. Following the same recommendation by Wu et al. (2007), we generate pseudo-covariates using the variant methods, OWN and ORX. To estimate $\hat{\gamma}_\cdot(\alpha)$, we use $B = 100$ sets of $k_P = 7$ pseudo-covariates. We also consider different FSR threshold $\gamma_0 \in \{0.01, 0.05, 0.1, 0.2\}$ in the analysis. A summary of the removed covariates is given in Table 3.1 and Table 3.2. Table 3.1 shows covariates removed at each step of the backward selection procedure. For example, the covariate *RINb* was the first to be removed from the full model with $r(RINb) = 0.26$, followed in subsequent iterations by the covariates *Eosi*, *Order*, *Conca*, *Diet*, *RFI*, *Lymp*, *Baso*, *RINa*, *Block*, *Neut*, *Concb*, and *Mono*.

Table 3.1: The last 13 columns show the removed covariate and its r value at each iteration of the FSR backward selection algorithm applied to the RFI RNA-seq dataset.

Covariate	*RINb*	*Eosi*	*Order*	*Conca*	*Diet*	*RFI*	*Lymp*	*Baso*	*RINa*	*Block*	*Neut*	*Concb*	*Mono*
r	0.26	0.49	0.62	0.65	0.53	2.07	2.87	3.46	6.30	7.71	7.85	9.42	11.45

Table 3.2 shows the estimate $\hat{\alpha}_*$ and the set of selected covariates with different FSR threshold γ_0. The results are unchanged when using either FSR formulas, γ_{RE} or γ_{ER}, and when using either pseudo-covariate generating methods, OWN or ORX. Therefore, only the results when using γ_{RE} and ORX are shown in Table 3.2. The selected covariates are slightly different for each FSR threshold. In particular, when $\gamma_0 = 0.01, 0.05, 0.1$ and 0.2, the number of selected covariates is 5,

6, 7, and 8, respectively. The set of selected covariates when $\gamma_0 = 0.1$ is the same as the set of covariates selected by the backward selection strategy in Nguyen et al. (2015).

Table 3.2: The selected covariates when applying the FSR backward selection algorithm to the RFI RNA-seq dataset with $\gamma_0 \in \{0.01, 0.05, 0.1, 0.2\}$.

γ_0	$\hat{\alpha}_*$	ORX
0.01	4.244055	*RINa*, *Block*, *Neut*, *Concb*, *Mono*
0.05	3.354193	*Baso*, *RINa*, *Block*, *Neut*, *Concb*, *Mono*
0.1	2.486859	*Lymp*, *Baso*, *RINa*, *Block*, *Neut*, *Concb*, *Mono*
0.2	1.799750	*RFI*, *Lymp*, *Baso*, *RINa*, *Block*, *Neut*, *Concb*, *Mono*

Figure 3.2 shows the histograms of p-values of selected covariates when $\gamma_0 \in \{0.01, 0.05, 0.1, 0.2\}$. These histograms have decreasing shape, which shows evidence that these covariates are highly relevant.

3.4 Simulation Study

3.4.1 Simulation Description

The goal of our simulation study is twofold: 1) to evaluate the FSR method in terms of its ability to select the most relevant covariates and 2) to evaluate the model selected by FSR method in terms of its ability to identify DE genes while controlling FDR. Such evaluations require simulated datasets to contain a set of truly relevant covariates and to contain both EE and DE genes for the included variables.

First, to evaluate the FSR method's ability to select the relevant covariates, we examine how well the method can control FSR at nominal thresholds. We consider 3 different nominal FSR thresholds $\gamma_0 \in \{0.01, 0.05, 0.1\}$ and 6 different sets of relevant covariates as shown in Table 3.3. These covariates are chosen based on their levels of relevance when applying the FSR backward selection procedure to the RFI RNA-seq dataset. The first three cases represent situations where there are a small number of relevant covariates (0, 1, or 2 relevant covariates) among all 13 covariates, while

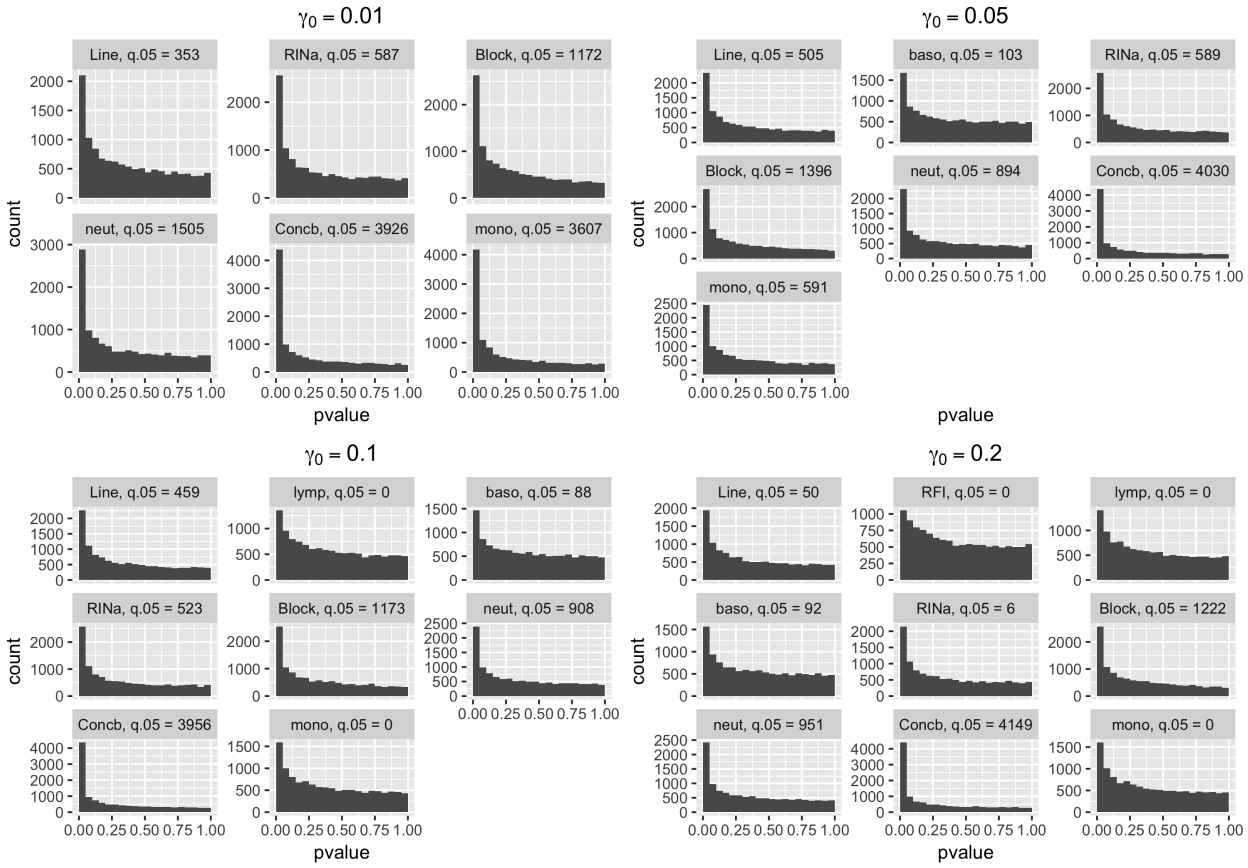

Figure 3.2: Histograms of p-values of the included variable (*Line*) and the covariates selected by our FSR backward selection algorithm with $\gamma_0 \in \{0.01, 0.05, 0.1, 0.2\}$.

the last three cases represent situations where there are a large number of relevant covariates (6, 7, or 8 relevant covariates) among all 13 covariates.

The last case with 8 relevant covariates is an example of when the relevant covariate *RFI* is strongly correlated with the included variable *Line*. The covariate *RFI* provides a continuous measure of residual feed intake for each of the 31 pigs in the study. Because the low RFI and high RFI lines were created by selecting on residual feed intake for several generations, it is not surprising that the low RFI pigs tend to have lower RFI values than the high RFI pigs in our study. This inclusion of the strongly confounding variable *RFI* makes it difficult to distinguish the direct effect of *Line* from the direct effect of *RFI* on transcript abundance levels, which may result in the failure of FDR control (Nguyen et al., 2015).

Table 3.3: Six different simulation scenarios corresponding to six different sets of truly relevant covariates.

Case	Model Size	Truly Relevant Covariates
1	0	Nothing
2	1	*Mono*
3	2	*Mono*, *Concb*
4	6	*Mono*, *Concb*, *Neut*, *Block*, *RINa*, *Baso*
5	7	*Mono*, *Concb*, *Neut*, *Block*, *RINa*, *Baso*, *Lymp*
6	8	*Mono*, *Concb*, *Neut*, *Block*, *RINa*, *Baso*, *Lymp*, *RFI*

Second, to evaluate the selected model's ability to identify DE genes while controlling FDR, we simulated datasets that contain both EE and DE genes with respect to the included variable and each of the relevant covariates. For each simulation scenario, as true parameters to simulate new data, we used the precision weights, the scaled error variances and the partial regression coefficient estimates from the fit of the corresponding model to the RFI RNA-seq data, except that we set partial regression coefficients on each variable to zero for a subset of genes to permit simulation of EE genes. More specifically, for each variable j (either relevant covariate or the included variable *Line*), the $\hat{m}_0^{(j)}$ least significant partial regression coefficients were set to zero, where $\hat{m}_0^{(j)}$ is the estimated number of the j-variable partial regression coefficients equal to zero when the method of Nettleton et al. (2006) is applied to the j-variable's p-values from the fit of the corresponding model

to the RFI RNA-seq data. This strategy yielded a parameter vector consisting of a scaled error variance, precision weights, and partial regression coefficients for each of 11280 genes. To simulate any particular dataset for a given set of truly relevant covariates (either 0, 1, 2, 6, 7, or 8 relevant covariates), we randomly sampled 2000 gene parameter vectors. The selected parameters and the explanatory variable values for the 31 pigs were used to simulate a 2000×31 dataset of read counts following the inverse steps of (3.1) and (3.2). Random selection of parameters and generation of data was independently repeated 100 times to obtain the 100 datasets for each scenario.

In addition to two goals above, we also investigate the sensitivity of the FSR approach to the number of pseudo-covariates $k_P = \{1, 3, 5, 7\}$. Furthermore, we consider 8 versions of the FSR method by combining 2 FSR formulas – γ_{ER} and γ_{RE} – and 4 pseudo-covariate generating methods – WN, RX, OWN, and ORX. We call these 8 versions WN.RE, WN.ER, RX.RE, RX.ER, OWN.RE, OWN.ER, ORX.RE, and ORX.ER.

3.4.2 Simulation Results

Using the simulated datasets, we first evaluate the ability to control FSR of 9 methods

- OldBS: the backward selection procedure with the $p.05$ measure of covariate relevance (Nguyen et al., 2015).
- WN.RE, WN.ER, RX.RE, RX.ER, OWN.RE, OWN.ER, ORX.RE, and ORX.ER: 8 versions of our FSR backward selection method.

Then, we analyzed these simulated datasets using covariates obtained from the 9 methods together with 5 other strategies handling covariates. These 5 strategies use model that includes

- all available covariates (Full)
- only the factor of primary interest (OnlyLine)
- surrogate variable analysis (sva -Leek and Storey (2007))
- direct surrogate variable analysis (dSVA -Lee et al. (2017))

- the true set of covariates used to simulate the data for each gene (Oracle).

Of course, the Oracle procedure cannot be used in practice, but its inclusion provides a useful reference measure of the performance achieved if covariate selection was perfect. In addition, sva (Leek and Storey, 2007) and dSVA (Lee et al., 2017) are the surrogate variable analysis method where the surrogate variables are constructed by ignoring all available covariates.

For these analysis strategies, the voom method in the `limma R` package was used to compute p-values for testing the significance of the partial regression coefficients corresponding to the explanatory variables. For the included variables, these p-values were converted to q-values (as described in Section 3.2.2), and genes with q-values no larger than 0.05 were declared as DE. For covariates that are subject to variable selection, these p-values were used to calculate the relevance measure r.

Figure 3.3 shows simulation results in evaluating the ability to select relevant covariates of OldBS and 8 versions of our proposed FSR method. OldBS intends to select a subset of covariates whose effects are accounted for in a model to maximize the number of DE genes with respect to the included variable *Line*. Because it is not designed to control FSR, FSR value of OldBS is unchanged for any threshold γ_0. The FSR of OldBS seems to be decreasing with respect to the number of relevant covariates k_I. In the scenario $k_I = 8$, FSR of OldBS is almost 0 because OldBS selects *Mono*, *Concb*, *Neut*, *Block*, *RNAa*, *Baso*, *Lymp* for more than 90 of the 100 simulated datasets in each scenario. This happens because in scenario $k_I = 8$ the relevant covariate *RFI* is strongly associated with the included variable *Line* due to the selection of lines as discussed in Section 3.4.1. Because OldBS always prefers model with maximum number of DE genes with respect to *Line*, *RFI* is discouraged in the selection process, which shown by the number of covariates selected in this case, $S = 7$.

Figure 3.4 shows the performance of 14 methods in identifying DE genes with respect to the included variable *Line*. As shown in Figure 3.3, our method performs best when using $k_P = 7$ pseudo-covariates. Therefore, when analyzing simulated data, our FSR method was implemented using $k_P = 7$ pseudo-covariates. Figure 3.4 shows that all 8 versions of our method control FDR

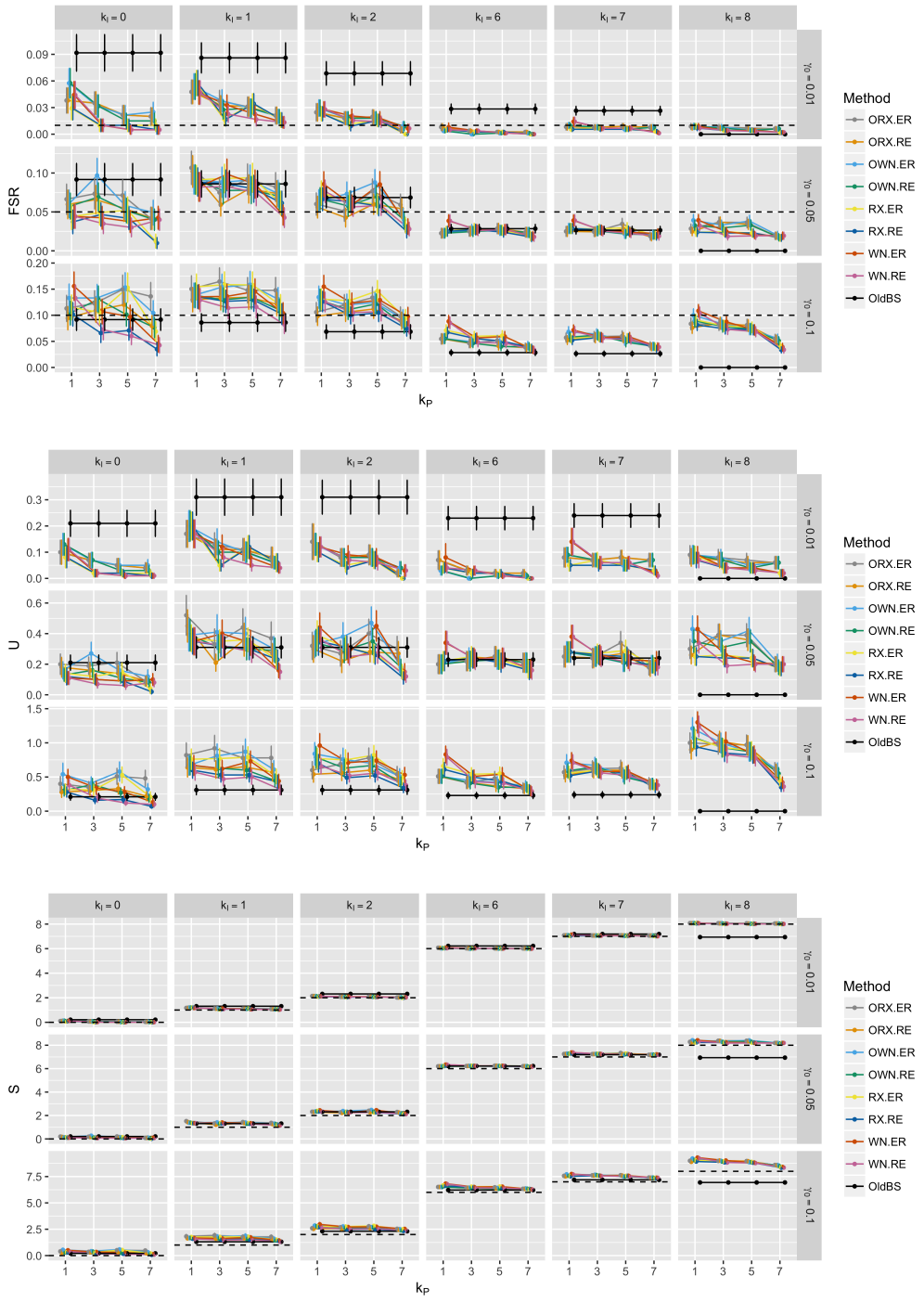

Figure 3.3: Empirical estimates of false selection rate (FSR), the average number of selected irrelevant covariates (U), the average number of selected relevant covariates (S) from 100 replications as a function of $k_P \in \{1, 3, 5, 7\}$ for OldBS, ORX.ER, ORX.RE, OWN.ER, OWN.RE, RX.ER, RX.RE, WN.ER, and WN.RE methods, three FSR thresholds $\gamma_0 \in \{0.01, 0.05, 0.1\}$, and six scenarios.

well. The OnlyLine, sva and dSVA methods fail to control FDR when $k_I = 8$, which is the case there is a relevant covariate that is strongly correlated with the included variable. OldBS performs as well as our method except when $k_I = 8$. The 8 versions of our proposed method perform well in terms of PAUC. Among all scenarios that FDR is controlled at the nominal level 0.05, the number of true positives detected by our method is very high.

3.5 Discussion

In this paper, we proposed a new covariate selection strategy in RNA-seq data analysis. We showed that our method can accurately choose the truly relevant covariates, even when there are covariates strongly associated with the included variables. As a result, our method performs very well in the downstream differential expression analysis. In particular, our method gives a reliable list of DE genes, which are shown by its ability to control FDR and its ability to distinguish EE and DE genes from one another.

We've also shown that the sva and dSVA methods suffer when there are many relevant covariates available. This suggests a careful consideration of analysis strategy needs to be taken into account under the availability of many covariates. These covariates should be checked to see if any of them is relevant before conducting further analyses.

We also want to emphasize that the proposed covariate selection strategy can be applied to the analysis of other 'omics data as well, such as microarray data because the nature of adding pseudo-covariates can be extended to any other high-dimensional data types.

3.6 Appendix: Description of Variables in the RFI Dataset

$\boldsymbol{x}_{\cdot 1} =$ ***Line*** is the categorical factor of primary scientific interest. Line has two levels, which correspond to the HRFI and LRFI selection lines. Among the 31 pigs in this study, 15 were from the LRFI line and 16 were from the HRFI line.

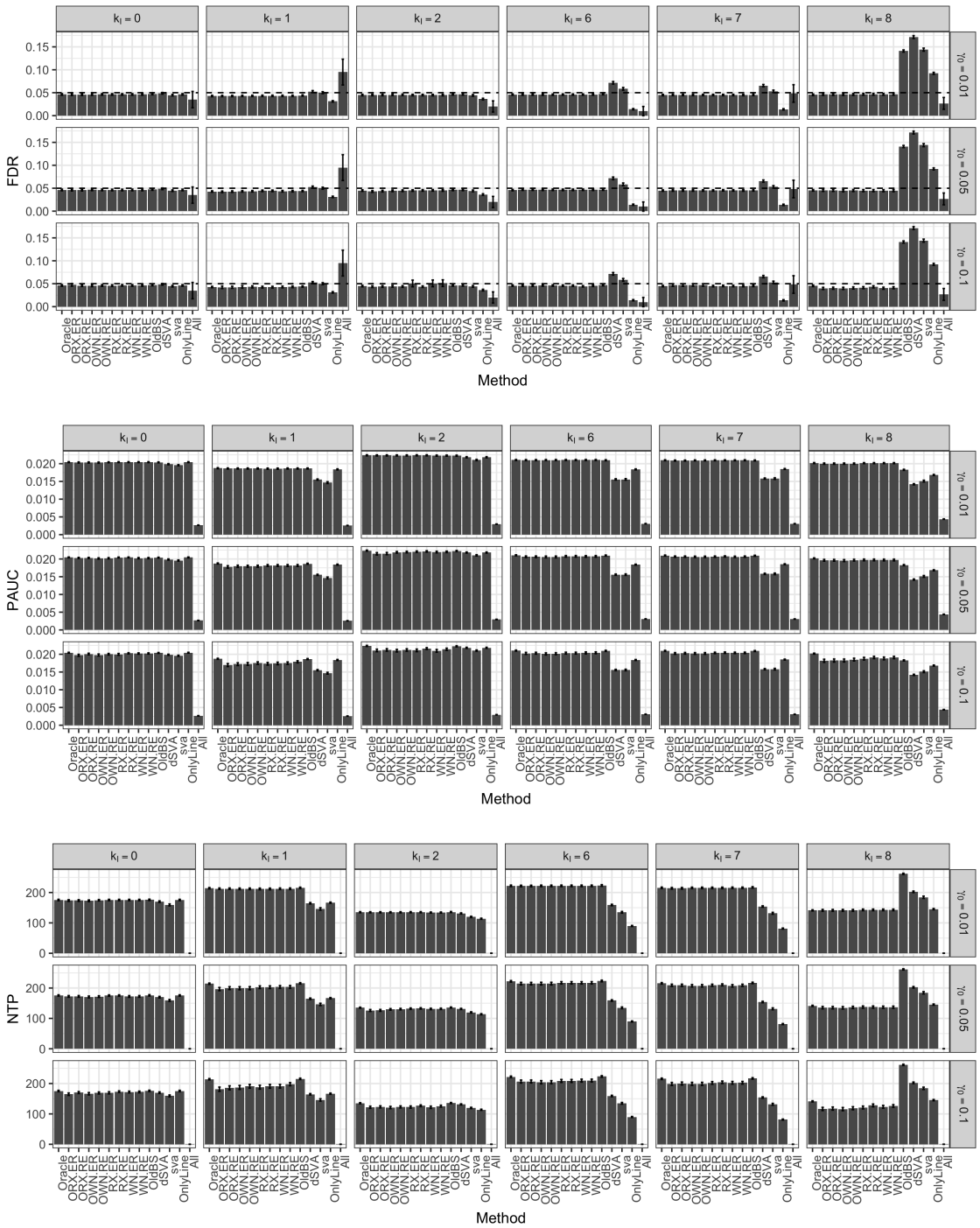

Figure 3.4: Empirical estimates of false discovery rate (FDR), the average number of true positive (NTP) detections of differential expression, and the average partial area under the receiver operating characteristic curve (PAUC) from 100 replications for Oracle, ORX.ER, ORX.RE, OWN.ER, OWN.RE, RX.ER, RX.RE, WN.ER, WN.RE, OldBS, dSVA, sva, OnlyLine, and All methods, three FSR thresholds $\gamma_0 \in \{0.01, 0.05, 0.1\}$, and six scenarios.

$\boldsymbol{x}_{\cdot 2}$ = ***RFI*** is a continuous covariate that provides a measure of the residual feed intake for each of the 31 pigs from which blood samples were drawn for RNA-seq analysis. Pigs in the HRFI line tend to have high *RFI* values, while pigs in the LRFI line tend to have low *RFI* values.

$\boldsymbol{x}_{\cdot 3}$ = ***Diet*** is a categorical factor with two levels corresponding to the two diets (high fiber, low energy vs. low fiber, high energy) that were fed to the pigs in this study. Approximately half the pigs within each line were fed each diet. Because RNA-seq analysis was performed on blood samples collected prior to the initiation of the two diets, this factor is not expected to be associated with the transcript abundance levels measured by RNA-seq.

$\boldsymbol{x}_{\cdot 4}$ = ***Baso*** is a continuous covariate that provides a measure of the concentration of basophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 5}$ = ***Eosi*** is a continuous covariate that provides a measure of the concentration of eosinophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 6}$ = ***Lymp*** is a continuous covariate that provides a measure of the concentration of lymphocyte cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 7}$ = ***Mono*** is a continuous covariate that provides a measure of the concentration of monocyte cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 8}$ = ***Neut*** is a continuous covariate that provides a measure of the concentration of neutrophil cells in the blood sample drawn from each pig.

$\boldsymbol{x}_{\cdot 9}$ = ***Concb*** is a continuous measure of the RNA concentration in each sample before globin depletion (a step that is necessary to focus sequencing efforts on messenger RNA molecules other than highly abundant globin messenger RNA in each blood sample).

$\boldsymbol{x}_{\cdot 10}$ = ***Conca*** is a continuous measure of the RNA concentration in each sample after globin depletion.

$\boldsymbol{x}_{\cdot 11}$ = ***RINb*** is a continuous measure of RNA integrity within each sample before globin depletion.

$\boldsymbol{x}_{\cdot 12}$ = ***RINa*** is a continuous measure of RNA integrity within each sample after globin depletion.

$\boldsymbol{x}_{\cdot 13}$ = ***Block*** is a categorical factor with four levels corresponding to the four blocks used to organize sample collection and processing. Initially, each block involved eight samples, two for each combination of *Line* and *Diet*. One LRFI sample from the first block was removed from the study due to low-quality RNA.

$\boldsymbol{x}_{\cdot 14}$ = ***Order*** is a categorical factor with eight levels indicating the random order samples were processed within each block.

Acknowledgments

This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), and by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) and the joint National Science Foundation (NSF)/NIGMS Mathematical Biology Program under award number R01GM109458. The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of USDA, NSF, or NIH.

Bibliography

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11(1):94.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2):125–153.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Lee, S., Sun, W., Wright, F. A., and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika*, 104(2):303–316.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lun, A. T. L., Chen, Y., and Smyth, G. K. (2016). It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. In Mathé, E. and Davis, S., editors, *Statistical Genomics: Methods and Protocols*, pages 391–416. Springer New York, New York, NY.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5):1544–6115.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628.

Nettleton, D., Hwang, J. T. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337.

Nguyen, Y., Nettleton, D., Liu, H., and Tuggle, C. K. (2015). Detecting differentially expressed genes with RNA-seq data using backward selection to account for the effects of relevant covariates. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(4):577–597.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014a). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014b). The role of spike-in standards in the normalization of RNA-seq. In *Statistical Analysis of Next Generation Sequencing Data*, pages 169–190. Springer.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):479–498.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Wu, Y. (2004). *Controlling Variable Selection By the Addition of Pseudo-Variables*. PhD dissertation, Department of Statistics, North Carolina State University.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102(477):235–243.

CHAPTER 4. RNA-SEQ ANALYSIS FOR REPEATED-MEASURES DATA

A paper in preparation

Yet Nguyen and Dan Nettleton

Abstract

With the reduction in price of next generation sequencing technologies, gene expression profiling using RNA-seq has increased the scope of sequencing experiments to include more complex designs, such as repeated-measures. In such designs, RNA samples are extracted from each experimental unit at multiple time points. The read counts that result from RNA sequencing of the samples extracted from the same experimental unit tend to be temporally correlated. Although there are many methods for RNA-seq differential expression analysis, existing methods do not properly account for within-unit correlations that arise in repeated-measures designs. We address this shortcoming by using normalized log-counts and associated precision weights in a general linear model pipeline with continuous autoregressive structure to account for the correlation among observations within each experimental unit. We then utilize parametric bootstrap to conduct differential expression inference. Simulation studies show the advantages of our method over alternatives that do not account for the correlation among observations within experimental units.

4.1 Introduction

One of the goals of transcriptomics data analysis is to identify genes whose mean transcript abundance levels differ across the levels of one or more categorical factors of interest. Such genes are typically referred to as differentially expressed (DE). Genes that are not DE are referred to as equivalently expressed (EE). Over the past decade, RNA sequencing (RNA-seq) technologies have emerged as a powerful and increasingly popular tool for expression profiling and differential

expression analysis (Oshlack et al., 2010). In a typical RNA-seq experiment, messenger RNA is extracted from each biological sample of interest. RNA sample is converted to complementary DNA (cDNA) which in turn is sequenced with high-throughput sequencing technology. This process generates millions of short reads from one or both ends of cDNA fragments. These short reads are mapped to the reference genome and the number of mapped short reads for a gene represents a measurement of the transcript abundance level of that gene in a given sample.

With the decreasing in price and increasing use of next generation sequencing technologies, RNA-seq experimental designs have become more complex. As a motivating example, we consider an RNA-seq experiment conducted on eight pigs, four from a high residual feed intake line (HRFI) and four from a low residual feed intake line (LRFI). Researchers wanted to evaluate how pigs from different lines respond to a treatment designed to stimulate the immune system, and how the responses change over time at the molecular genetic level. They used RNA-seq technology to measure transcript abundances in blood samples from each pig at four times after treatment: 0, 2, 6, and 24 hours. The experiment is explained in greater detail in Section 4.3 of this paper. A statistical model for these data should consider the within-unit correlation expected due to repeated measurements on each pig.

Many general purpose RNA-seq differential expression analysis methods have been developed, such as edgeR (Robinson et al., 2010), QuasiSeq (Lund et al., 2012), DESeq and DESeq2 (Anders and Huber, 2010; Love et al., 2014) among many others. These methods use negative binomial generalized linear models to analyze RNA-seq data and are appropriate for designs providing uncorrelated measurements within each gene. Furthermore, several methods have been developed for time-course designs, such as NextmaSigPro (Nueda et al., 2014), DyNB (Äijö et al., 2014), TRAP (Jo et al., 2014), SMARTS (Wise and Bar-Joseph, 2015), and EBSeq-HMM (Leng et al., 2015), which were collectively reviewed by Spies and Ciaudo (2015). However, these methods do not take within-unit correlation of transcript abundance measurements into account, which may result in many false discoveries or failure to distinguish EE and DE genes. Theoretically, a generalized linear mixed model (GLMM) approach can be used to account for random effects and general correla-

tion structure, but the approach suffers from convergence issues for many genes because RNA-seq experiments usually have a small sample size and many zero counts for many genes (Cui et al., 2016). Therefore, a new statistical method that is stable numerically under small sample size circumstances and, at the same time, controls type I error rate well is desirable. One approach that addresses numerical instability when analyzing repeated-measures RNA-seq data is to use normal-error linear modeling for log-transformed counts instead of using discrete probability distributions, such as the negative binomial distribution.

Recently, Law et al. (2014) have proposed the voom approach to use normal-based methods for analyzing log-transformed RNA-seq data with linear models that explicitly account for heteroscedasticity by the use of precision weights. They showed that correctly capturing the mean-variance relationship in the transformed data is more important than assuming a probability model that acknowledges the discrete characteristics of the original counts. In particular, by estimating precision weights for observations of transformed counts and including them into a general linear model framework, Law et al. (2014) showed that the log-transformed-based linear model approach performs better than methods based on negative binomial models. Furthermore, the voom approach facilitates more complex analyses, such as the variance component score test for gene set testing in longitudinal RNA-seq data recently proposed by Agniel and Hejblum (2017).

In our paper, we will take advantage of the voom approach together with a parametric bootstrap method to detect DE genes with repeated-measures RNA-seq data. For each gene, we model the correlation among observations taken at unequally-spaced time points by a continuous autoregressive correlation structure in a general linear model framework. Parameters are estimated by residual maximum likelihood (REML) using the `gls` function in the `nlme` R package (Pinheiro et al., 2017). We conduct hypothesis testing using a parametric bootstrap method. Simulation studies show the advantages of our method over alternatives that do not account for the correlation among observations within each gene in terms of false discovery rate (FDR) control and the ability to distinguish EE and DE genes. Although, we focus on repeated-measures analysis in this paper, our method can also be easily extended to other complex designs.

The remainder of the paper is organized as follows. We formally define our proposed method in Section 4.2, first by revisiting the voom procedure and then specifying the bootstrap strategy for inference. In Section 4.3, we apply our proposed method as well as several other alternative methods to analyze the repeated-measures RNA-seq dataset that motivates our work. We compare the performance of our method with that of alternative methods by a simulation study in Section 4.4. The paper concludes with a discussion in Section 4.5.

4.2 Methods

4.2.1 Notations and Preliminaries

Consider the analysis of m genes using RNA-seq read count data from n subjects and T time points. For $g = 1, \ldots, m$, $i = 1, \ldots, n$, and $t = 1, \ldots, T$, let r_{git} be the read count for gene g from subject i at time d_t. Let $\boldsymbol{x}_{it} = (\boldsymbol{x}'_{it1}, \ldots, \boldsymbol{x}'_{itk})'$ be a vector encoding information on k explanatory variables for subject i at time d_t. The k explanatory variables may include multilevel factors of primary scientific interest and other continuous or multilevel categorical covariates. Let $\boldsymbol{X} = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1T}, \ldots, \boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nT})'$ and suppose that $\boldsymbol{X}$ has full column rank with $\text{rank}(\boldsymbol{X}) = u$. Law et al. (2014) defined the following transformation to obtain the log-counts per million (log-cpm) for each count

$$y_{git} = \log_2\left(\frac{r_{git} + 0.5}{R_{it} + 1} \times 10^6\right), \ \boldsymbol{y}_g = (y_{g11}, \ldots, y_{g1T}, \ldots, y_{gn1}, \ldots, y_{gnT})', \tag{4.1}$$

where R_{it} is a normalization offset computed for subject i at time d_t. The normalization offsets account for differences in read counts across the RNA-seq samples. Many normalization procedures have been proposed in the literature (see, e.g., Marioni et al. (2008), Mortazavi et al. (2008), Robinson and Oshlack (2010), Anders and Huber (2010), Bullard et al. (2010), Risso et al. (2014a), Risso et al. (2014b), and references therein). Throughout this paper, we set R_{it} to be the 0.75 quantile of RNA-seq sample read counts from subject i at time d_t according to the recommendation of Bullard et al. (2010). With this choice for the normalization factor, the y_{git} values are no longer

"counts per million mapped reads" on the log scale, but this interpretation is irrelevant for the differential expression analysis that is the focus of our work.

4.2.2 The voom Procedure

The voom procedure (Law et al., 2014) estimates the mean-variance relationship of the log-counts and generates a precision weight for each observation according to the following algorithm:

1. For each gene g, initially assume the linear model

$$y_{git} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_g + \varepsilon_{git},\ \varepsilon_{git} \sim \mathcal{N}(0, \sigma_g^2), \quad g = 1, \ldots, m;\ i = 1, \ldots, n;\ t = 1, \ldots, T.$$

2. Let $\widetilde{\boldsymbol{\beta}}_g = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_g$ and $\tilde{\sigma}_g = \sqrt{\frac{(\boldsymbol{y}_g - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_g)'(\boldsymbol{y}_g - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_g)}{nT-u}}$ be the maximum likelihood (ML) and REML estimates of $\boldsymbol{\beta}_g$ and σ_g, respectively.

3. Let $\tilde{r}_g = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{git} + \frac{1}{nT}\log_2\left(\prod_{i=1}^{n}\prod_{t=1}^{T}(R_{it}+1)\right) - \log_2(10^6)$ be the mean log-count value for gene g.

4. Let $\mathrm{lo}(\cdot)$ be the predictor obtained by fitting a LOWESS regression (Cleveland, 1979) of $\tilde{\sigma}_g^{1/2}$ on $\tilde{r}_g$. The voom precision weight for y_{git} is calculated by

$$w_{git} = \left[\mathrm{lo}\left(\boldsymbol{x}_{it}^T\widehat{\boldsymbol{\beta}}_g + \log_2(R_{it}+1) - \log_2(10^6)\right)\right]^{-4}.$$

4.2.3 Modeling for Repeated Measure RNA-seq Data

To account for the correlation among observations within the g^{th} gene, we assume the Gaussian general linear model

$$\boldsymbol{y}_g = \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g, \quad \boldsymbol{\varepsilon}_g \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2\boldsymbol{V}_g), \quad \boldsymbol{V}_g = \boldsymbol{W}_g^{-1/2}\boldsymbol{D}_g\boldsymbol{W}_g^{-1/2}, \tag{4.2}$$

where

$$\boldsymbol{W}_g = \begin{bmatrix} w_{g11} & 0 & \ldots & 0 \\ 0 & w_{g12} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_{gnT} \end{bmatrix}$$

is the matrix of precision weights and $\boldsymbol{D}_g$ is an $nT \times nT$ block-diagonal correlation matrix with blocks of the form

$$\boldsymbol{A}_g = \begin{bmatrix} 1 & \rho_g^{|d_2-d_1|} & \dots & \rho_g^{|d_T-d_1|} \\ \rho_g^{|d_1-d_2|} & 1 & \dots & \rho_g^{|d_T-d_2|} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_g^{|d_1-d_T|} & \rho_g^{|d_2-d_T|} & \dots & 1 \end{bmatrix}, \quad 0 \le \rho_g < 1.$$

This is a continuous autoregressive correlation structure, denoted as $CAR(1)$ (Pinheiro and Bates, 2000), which we will use to model the dependence among within-unit observations. We employ the function `gls` in the `nlme` R package (Pinheiro et al., 2017) to fit model (4.2), resulting in the REML estimators $\widehat{\sigma}_g^2$ and $\widehat{\rho}_g$ of σ_g^2 and ρ_g, respectively, as well as the plug-in estimator $\widehat{\boldsymbol{V}}_g = \boldsymbol{W}_g^{-1/2}\widehat{\boldsymbol{D}}_g\boldsymbol{W}_g^{-1/2}$ of $\boldsymbol{V}_g$ where ρ_g in $\boldsymbol{D}_g$ is substituted by $\widehat{\rho}_g$, and $\widehat{\boldsymbol{\beta}}_g = (\boldsymbol{X}'\widehat{\boldsymbol{V}}_g^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{V}}_g^{-1}\boldsymbol{y}_g$ as an estimator of $\boldsymbol{\beta}_g$.

4.2.4 Shrinkage Estimators of Error Variances

In microarray analysis, Smyth (2004) showed that using the shrinkage of the estimated error variances toward a pooled estimate can stabilize inference when the number of arrays is small. We follow the same procedure to obtain the shrinkage estimator of the error variance σ_g^2 for each gene. Particularly, we assume that

$$\widehat{\sigma}_g^2|\sigma_g^2 \sim \sigma_g^2 \frac{\chi^2_{nT-u}}{nT-u} \tag{4.3}$$

and, for some parameters s_0^2 and u_0,

$$\frac{u_0 s_0^2}{\sigma_g^2} \sim \chi^2_{u_0},$$

which together with (4.3) implies an inverse-gamma conditional distribution for σ_g^2 specified by

$$\frac{1}{\sigma_g^2}\Big|\widehat{\sigma}_g^2 \sim \text{Gamma}\left(\frac{nT-u+u_0}{2}, \frac{(nT-u)\widehat{\sigma}_g^2 + u_0 s_0^2}{2(nT-u+u_0)}\right).$$

A shrinkage estimator of σ_g^2 is given by

$$s_g^2 = \widehat{E}^{-1}(\sigma_g^{-2}|\widehat{\sigma}_g^2) = \frac{(nT-u)\widehat{\sigma}_g^2 + \widehat{u}_0\widehat{s}_0^2}{nT-u+\widehat{u}_0}, \tag{4.4}$$

where $\widehat{u}_0$ and $\widehat{s}_0^2$ are the estimators of the hyperparameters u_0 and s_0^2 obtained from the theoretical marginal distribution of $\widehat{\sigma}_g^2$ using a method of moments approach (Smyth, 2004). The shrinkage estimator s_g^2 will be used in our inference strategy instead of the unshrunken REML estimator $\widehat{\sigma}_g^2$.

4.2.5 General Hypothesis Testing of Regression Coefficients Using Moderated F-Statistics

Suppose for each gene g we are interested in testing a null hypothesis of the form

$$H_{0g} : \boldsymbol{C}\boldsymbol{\beta}_g = \mathbf{0} \quad \text{vs.} \quad H_{ag} : \boldsymbol{C}\boldsymbol{\beta}_g \neq \mathbf{0},$$

where $\boldsymbol{C}$ is an $l \times u$ contrast matrix of rank l. An extension of the moderated F-statistic of Smyth (2004) for gene g is defined as

$$K_g = (\boldsymbol{C}\widehat{\boldsymbol{\beta}}_g)'(\boldsymbol{C}(s_g^2 \boldsymbol{X}'\widehat{\boldsymbol{V}}_g^{-1}\boldsymbol{X})^{-1}\boldsymbol{C}')^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_g)/l. \tag{4.5}$$

In general $\boldsymbol{C}\widehat{\boldsymbol{\beta}}_g$ is a non-linear function of $\boldsymbol{y}_g$, and the exact distribution of K_g is unknown even when model (4.2) holds exactly. Because RNA-seq experiments often have small sample size, we cannot rely on asymptotic approximations and instead will approximate the distribution of K_g using a parametric bootstrap approach (Efron and Tibshirani, 1993). For all $g = 1, \ldots, m$, carry out the following steps:

1. Simulate $\boldsymbol{\varepsilon}_g^* \sim \mathcal{N}(\mathbf{0}, s_g^2 \widehat{\boldsymbol{V}}_g)$ and calculate $\boldsymbol{y}_g^* = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_g + \boldsymbol{\varepsilon}_g^*$.

2. Calculate r_{git}^* using y_{git}^* according to (4.1), i.e.,

$$r_{git}^* = \max\{2^{y_{git}^*} \times (R_{it} + 1)/10^6 - 0.5, 0\}.$$

3. Apply the voom procedure described in Section 4.2.3 and the shrinkage procedure described in Section 4.2.4 to compute $\widehat{\boldsymbol{\beta}}_g^*, s_g^{2*}, \widehat{\rho}_g^*$, and $\widehat{\boldsymbol{V}}_g^*$ from $\{r_{git}^*\}$ and $\boldsymbol{X}$ just as $\widehat{\boldsymbol{\beta}}_g, s_g^2, \widehat{\rho}_g$, and $\widehat{\boldsymbol{V}}_g$ were obtained from $\{r_{git}\}$ and $\boldsymbol{X}$.

4. Compute $K_g^* = (\boldsymbol{C}\widehat{\boldsymbol{\beta}}_g^* - \boldsymbol{C}\widehat{\boldsymbol{\beta}}_g)'(\boldsymbol{C}(s_g^{2*}\boldsymbol{X}'\widehat{\boldsymbol{V}}_g^{*-1}\boldsymbol{X})^{-1}\boldsymbol{C}')^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_g^* - \boldsymbol{C}\widehat{\boldsymbol{\beta}}_g)/l$.

5. Repeat steps 1 through 4 B times to obtain null statistics $K^*_{g1}, \ldots, K^*_{gB}$.

Taking advantage of the parallel structure in which the same model is fitted for each of many genes, we combine the bootstrap null statistics for all genes to calculate a p-value for each gene. Numerically, the p-value for gene g is calculated by the proportion of all bootstrap null statistics $\{K^*_{g1}, \ldots, K^*_{gB} : g = 1, \ldots, m\}$ that match or exceed the observed statistic K_g, i.e.,

$$p_g = \frac{1}{mB} \sum_{j=1}^{m} \sum_{b=1}^{B} \mathbb{1}(K^*_{jb} \geq K_g), \tag{4.6}$$

where $\mathbb{1}$ is an indicator function. These p-values are converted to q-values (Storey, 2002). To approximately control FDR at any desired level α, a null hypothesis is rejected if and only if its q-value is less than or equal to α. When calculating q-values by the method of Storey (2002), we need an estimate of m_0, the number of true null hypotheses among all m null hypotheses tested. In this paper, m_0 is estimated by the histogram-based method of Nettleton et al. (2006). Desirable theoretical properties of a closely related histogram-based approach were illustrated by Liang and Nettleton (2012).

The same idea of pooling used in (4.6) has been used by Storey et al. (2005) in a time-course microarray analysis. Even if the test statistics for all genes do not follow the same null distribution, p-values computed via pooling can be valid for use in q-value estimation (Storey et al., 2004). Particularly, Storey et al. (2004) showed that a sufficient condition for valid q-value estimation is that the collection of p-values from tests with a true null hypothesis have an empirical distribution that is stochastically smaller than or equal to a uniform distribution. Results from the analysis of simulated data in Section 4.4 show that our approach to p-value calculation satisfies this sufficient condition and thus provides valid p-values for the calculation of q-values that can be used to control FDR.

4.3 Analysis of an LPS RNA-Seq Dataset

Lipopolysaccharide (LPS) is extensively used to study acute inflammatory and immune response in humans and animals. In this section, we apply our proposed method and three other methods –

DESeq2 (Love et al., 2014), voom (Law et al., 2014), and edgeR (Robinson et al., 2010; Lun et al., 2016) – to analyze an RNA-seq dataset from a study of the inflammatory response in pigs triggered by LPS at the transcription level (Liu, 2017, Chapter 2). The experiment design is described as follows. Four pigs of each residual feed intake line, HRFI and LRFI, were injected LPS from *E. coli* 05:B5 bacteria. Blood samples were collected from eight pigs immediately before the injection (called time point 0 in the following), 2, 6, and 24 hours after the injection. An RNA sample was extracted and sequenced from each blood sample after globin depletion. In total, there were 4 (pigs) $\times$ 2 (lines) $\times$ 4 (time points) = 32 RNA-seq libraries. Researchers wanted to understand the molecular mechanism of LPS response by identifying genes differentially expressed between lines (Line), across time points (Time), or through interactions among lines and time points (Line $\times$ Time).

This is an example of a repeated-measures design, where RNA samples were extracted from each pig at four different unequally-spaced time points. The RNA-seq dataset consists of read counts for 11911 genes for each of 32 RNA samples. Following standard practice, this dataset excludes genes with mostly low read counts because such genes contain little information about differential expression. In particular, the 11911 genes analyzed in this study each have average read counts of at least 8 and no more than 28 zero counts across 32 RNA samples. The same threshold for gene inclusion was used throughout the simulation studies described in Section 4.4.

A special characteristic of this experiment is the potential for circadian rhythm effects that may induce the correlation between observations taken at the same time of day. Thus, although times 0 and 24 are farthest apart when time is considered to unfold on a linear axis, the correlation between the time 0 and 24 observations may be large because these observations are taken at the same time of day. To evaluate this possibility, we conducted a preliminary analysis of the LPS RNA-seq dataset by applying the voom procedure and model (4.2) as in Section 4.2.3, where $\boldsymbol{D}_g$ is an $nT \times nT$ block-diagonal correlation matrix with blocks of the unstructured form

$$\boldsymbol{A}_g = \begin{bmatrix} 1 & \rho_{g,1} & \rho_{g,2} & \rho_{g,3} \\ \rho_{g,1} & 1 & \rho_{g,4} & \rho_{g,5} \\ \rho_{g,2} & \rho_{g,4} & 1 & \rho_{g,6} \\ \rho_{g,3} & \rho_{g,5} & \rho_{g,6} & 1 \end{bmatrix}, \quad 0 \leq \rho_{g,1}, \ldots, \rho_{g,6} \leq 1,$$

instead of the $CAR(1)$ form described in Section 4.2.3. The mean structure of the data is modeled by $\boldsymbol{X}\boldsymbol{\beta}_g$, where the design matrix $\boldsymbol{X}$ is constructed by two factors Time and Line so that there are eight different means, one for each combination of Time and Line.

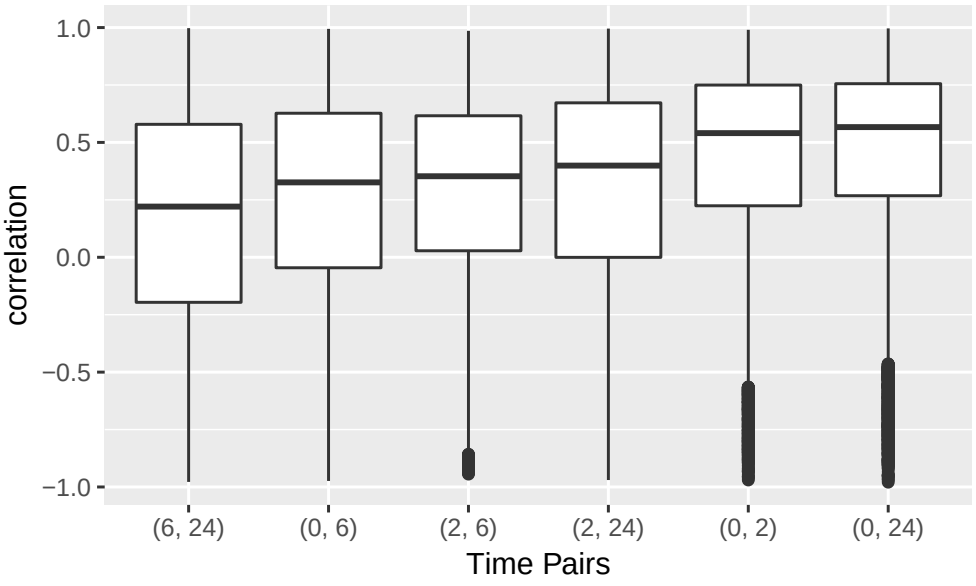

Figure 4.1: Estimated correlations across all 11911 genes for each pair of time points. The correlation for each gene was estimated by REML using the function `gls` in the `nlme` R package applied to the log-transformed LPS RNA-seq data and their precision weights according to the model (4.2).

Figure 4.1 shows boxplots of correlations across all 11911 genes for each pair of time points. Both the average and median correlations increase across the time pair sequence (6, 24), (0, 6), (2, 6), (2, 24), (0, 2), (0, 24). This evidence suggests that the circadian rhythm effects on correlations may be relevant. In particular, this empirical evidence shows the correlation between time 6 and time 24 observations tends to be smallest and the correlation between time 0 and time 24 observations tends to be largest. To account for correlations that are not monotone with time difference on

the original time scale, we propose a remapping procedure of the linear time points to a new time coordinate system so that time 6 and time 24 are farthest apart and other times positioned between them in accordance with the empirical correlation patterns apparent in the data. Without loss of generality, we consider a mapping of the time points $\{0, 2, 6, 24\}$ to the following new points

$$24 \mapsto d_{24} \equiv 0,\ 6 \mapsto d_6 \equiv 1;\ 0 \mapsto d_0;\ 2 \mapsto d_2,\ 0 < d_0, d_2 < 1.$$

The diagonal block $\boldsymbol{A}_g$ now is

$$\boldsymbol{A}_g = \begin{bmatrix} 1 & \rho_g^{|d_0-d_2|} & \rho_g^{|d_0-d_6|} & \rho_g^{|d_0-d_{24}|} \\ \rho_g^{|d_2-d_0|} & 1 & \rho_g^{|d_2-d_6|} & \rho_g^{|d_2-d_{24}|} \\ \rho_g^{|d_6-d_0|} & \rho_g^{|d_6-d_2|} & 1 & \rho_g^{|d_6-d_{24}|} \\ \rho_g^{|d_{24}-d_0|} & \rho_g^{|d_{24}-d_2|} & \rho_g^{|d_{24}-d_6|} & 1 \end{bmatrix} \tag{4.7}$$

$$= \begin{bmatrix} 1 & \rho_g^{|d_0-d_2|} & \rho_g^{1-d_0} & \rho_g^{d_0} \\ \rho_g^{|d_2-d_0|} & 1 & \rho_g^{1-d_2} & \rho_g^{d_2} \\ \rho_g^{1-d_0} & \rho_g^{1-d_2} & 1 & \rho_g \\ \rho_g^{d_0} & \rho_g^{d_2} & \rho_g & 1 \end{bmatrix}, \quad 0 \le \rho_g < 1. \tag{4.8}$$

To estimate appropriate values for d_0 and d_2, we consider values best supported by REML log likelihood across all genes. Let $\ell_g(\sigma_g^2, \rho_g | \boldsymbol{d} := (d_0, d_2, 1, 0))$ be the REML log likelihood function for data from gene g according to model (4.2) with $\boldsymbol{A}_g$ as defined in (4.7). We choose d_0 and d_2 to maximize

$$h(\boldsymbol{d}) = \sum_{g=1}^{m} \ell_g(\widehat{\sigma}_g^2, \widehat{\rho}_g | \boldsymbol{d}),$$

where $\widehat{\sigma}_g^2$ and $\widehat{\rho}_g$ are REML estimates of σ_g^2 and ρ_g, respectively. Using the function `constrOptim` in the `base` R package, we can easily obtain an approximate maximizer of $h(\boldsymbol{d})$ at

$$\hat{d}_0 = 0.26, \hat{d}_2 = 0.52.$$

In terms of AIC, AIC of the model (4.2) for our choice of $\boldsymbol{A}_g$ using the new time points is smaller than AIC of that for $\boldsymbol{A}_g = \boldsymbol{I}$, $\boldsymbol{A}_g = Symm$, $\boldsymbol{A}_g = CompSymm$, $\boldsymbol{A}_g = AR(1)$ with original time points, $\boldsymbol{A}_g = AR(1)$ with new time points and $\boldsymbol{A}_g = CAR(1)$ with original time points on

average 66%, 72%, 53%, 68%, 54% and 76% of the genes, respectively. Even though AIC does not guarantee to lead us to the correct correlation structure (Keselman et al., 1998), it still provides useful evidence for choosing a reasonable correlation structure. In this sense, AIC seems to suggest that our choice of correlation structure is superior than other common correlation structures.

Now we apply our proposed method to the LPS RNA-seq data using the new time points instead of the original time points. We also compare our results to those obtained by the popular RNA-seq analysis methods – voom, DESeq2 and edgeR – which ignore correlation among observations. Fig. 4.2 summarizes the analysis results of these methods when FDR is nominally controlled at 5%. Recall that both DESeq2 and edgeR methods utilize negative binomial generalized linear models. DESeq2 uses shrinkage estimation for dispersion parameters and fold changes to improve the stability and interpretability of estimates, while edgeR employs its own version of shrinkage estimation for dispersion parameters and does not shrink log fold change estimates. To conduct inference about the contrasts of interest, we use likelihood ratio test in DESeq2 and the quasi-likelihood F-test in edgeR. It is clear from the Venn diagrams in Figure 4.2 that our proposed method, voomboot, detects the smallest number of DE genes with respect to the Line main effect, and detects the largest number of DE genes for the tests that involve the time factor. The differences between our proposed method and the others can be explained due to the fact that voom, DESeq2 and edgeR tend to underestimate the covariances between observations measured at different time points, and therefore overestimate the variances of differences between these observations, as well as underestimate the variances of averages of these observations. This leads us to the situation that the three methods voom, DESeq2, and edgeR may have an excessive number of large values of test statistics for the Line main effect, while have an inadequate number of small values of test statistics that involve the time factor.

4.4 Simulation Study

We considered three simulation scenarios described in detail in Sections 4.4.1, 4.4.2, and 4.4.3. In each scenario, voomboot, voom, egdeR, and DESeq2 are compared in terms of their ability to

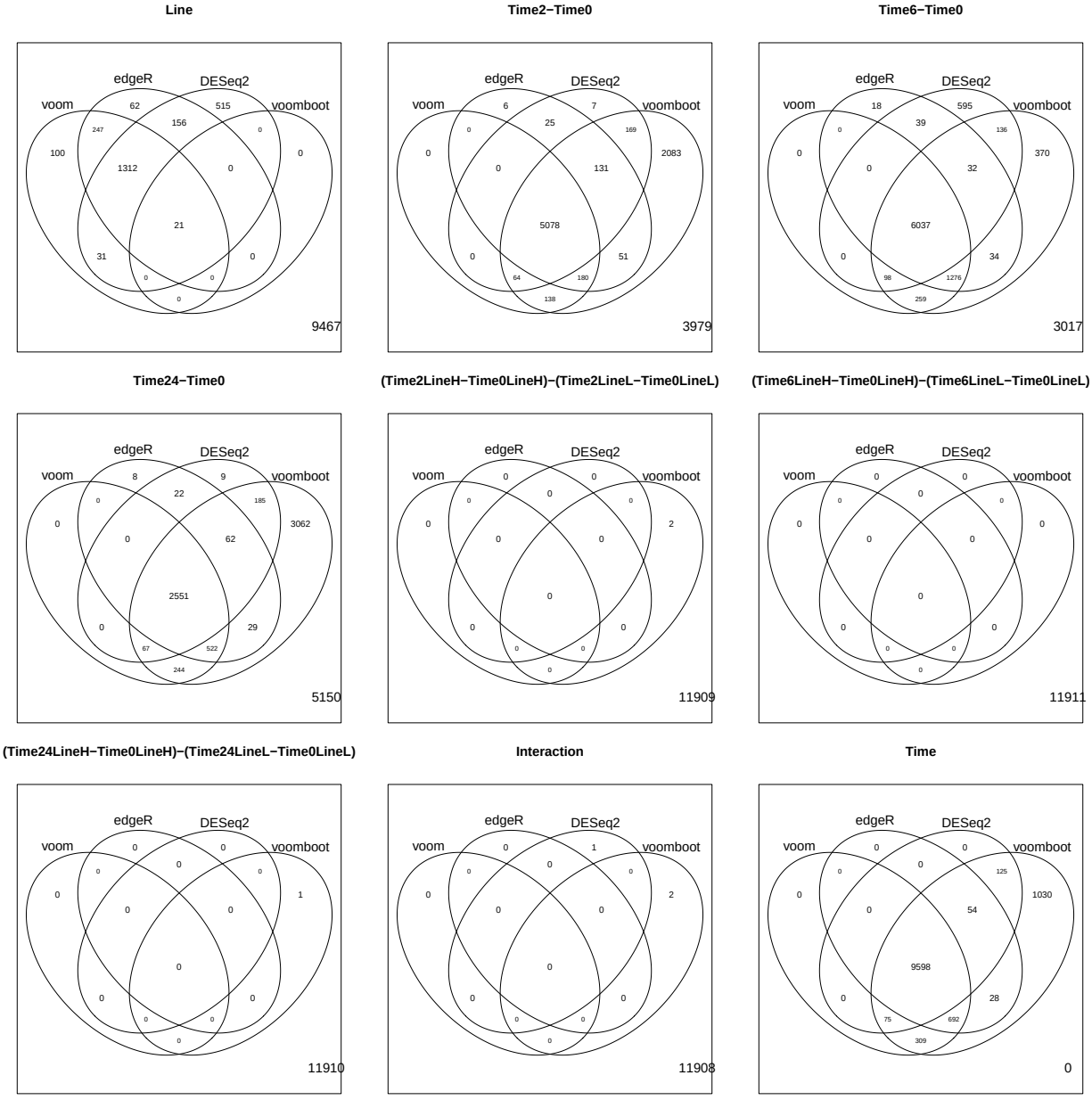

Figure 4.2: Venn diagrams showing numbers of DE genes (FDR is nominally controlled at 0.05) with respect to nine effects when analyzing the LPS RNA-seq dataset using four methods: voom, edgeR, DESeq2, and voomboot.

identify DE genes while controlling FDR. Such comparisons require simulated datasets to contain both EE and DE genes with respect to a contrast of interest. Within each scenario, we consider four contrasts: 1) `Line`: the main effect of Line factor, 2) `Time`: the main effect of Time factor, 3) `Time2-Time0`: the difference between time 2 and time 0, and 4) `Interaction`: the interaction between line and time factors.

For each scenario and contrast, we simulated 100 datasets. Each dataset included read counts for 8 pigs at 4 time points and 5000 genes. The read counts were simulated based on (4.1) and (4.2) for scenarios described in Sections 4.4.1 and 4.4.2. For the third scenario described in Section 4.4.3, the read counts were simulated from a negative binomial generalized linear mixed-effects model.

We want to emphasize that our simulation study considers four contrasts of interest. This is different from most simulation studies where only two-group comparison is considered to evaluate the performance of a differential expression method. Our simulation setting allows us to fully investigate effects of within-unit correlation on the inference of within-subject and between-subject contrasts. The analysis in Section 4.3 showed that a method ignoring within-unit correlation tends to overestimate the variance of a within-subject contrast and underestimate the variance of a between-subject contrast, which may be inefficient for inference of both within-subject and between-subject contrasts.

4.4.1 Simulation Scenario 1: Ideal Case, $\rho_g \neq 0$

The first simulation scenario provides a favorable case for our proposed method in which the read counts were simulated from the working model assumptions (4.1) and (4.2). As true parameter values for simulating new data, for each gene, we used the normalization offsets R_{it}, the estimates of the precision weight matrix $\boldsymbol{W}_g$, the correlation parameter ρ_g, and the regression coefficients $\boldsymbol{\beta}_g$ from the fit of the model (4.2) to the LPS RNA-seq dataset, except that we set partial regression coefficients corresponding to the contrast of interest to zero for a subset of genes to permit simulation of EE genes with respect to the contrast of interest. More specifically, 5955 least significant partial regression coefficients for the contrast of interest were set to zero. This strategy yielded a parameter

set (consisting of the normalization offsets R_{it}, the precision weight matrix $\boldsymbol{W}_g$, the correlation parameter $\widehat{\rho}_g$, and regression coefficients $\boldsymbol{\beta}_g$) for each of 5955 EE genes, 11911 - 5955 = 5956 DE genes and a given contrast. To simulate any particular dataset for a given contrast of interest, we randomly sampled 4000 parameter sets from the EE genes and 1000 parameter sets from the DE genes. The selected parameter sets and the design matrix constructed by the linear combination of Time, Line, and Line × Time for 32 samples were used to simulate a 5000 × 32 dataset of read counts by first simulating log-transformed data using formula (4.2), then converting the log-transformed data back to read counts using formula (4.1). Random selection of parameter sets and generation of data was independently repeated 100 times to obtain 100 datasets for each one of the four contrasts of interest: `Line`, `Time`, `Interaction`, `Time2-Time0`.

4.4.2 Simulation Scenario 2: Model Misspecification Case, $\rho_g = 0$

The second simulation scenario is designed to evaluate our proposed method when the observations within each gene are independent. This scenario slightly violates our working model assumptions and is less favorable for our method than alternatives such as voom, edgeR2 and DESeq2 that do not take within-gene correlation into account. In this scenario, each dataset was simulated using exactly the same procedure described in Section 4.4.1, except that the within-gene correlation was set to zero instead of using the estimate $\widehat{\rho}_g$ from LPS RNA-seq data.

4.4.3 Simulation Scenario 3: Model Misspecification, Negative Binomial Generalized Linear Mixed Effect Model (NB_GLIMMIX)

The third simulation scenario is designed to evaluate our proposed method when, contrary to our working model assumptions, read counts were generated from a negative binomial GLMM. First each gene of the LPS RNA-seq data was analyzed using the SAS GLIMMIX procedure with the negative binomial distribution and a log link function including the linear combination of Line, Time, and Line × Time. The offset parameters in the GLIMMIX procedure were set to be $\log(R_{it})$, the log of the upper quartiles of LPS RNA-seq samples. The $\boldsymbol{R}$-side random effect

was set by default as $\boldsymbol{R}_g = \phi_g \boldsymbol{I}$, where ϕ_g is the negative binomial dispersion parameter. The $\boldsymbol{G}$-side random effect was chosen as $SP(POW)$ structure with respect to time factor, which is the same as $CAR(1)$ structure in our working model. The pseudo-likelihood technique (Wolfinger and O'Connell, 1993; Breslow and Clayton, 1993) was employed in estimation. The estimates of covariance matrix of the fixed-effect parameter and denominator degrees of freedom for t- and F-tests were adjusted using Kenward-Roger method (Kenward and Roger, 1997). When analyzing the LPS RNA-seq data using the GLIMMIX procedure, we found that it failed to converge for many genes no matter which estimation algorithm was used, for example, algorithms based on linearization, Laplace approximation, or adaptive quadrature. A possible reason could be due to the small sample size of RNA-seq data with many zero counts. The same numerical instability has been observed in literature, for example, see Cui et al. (2016). Therefore, we did not incorporate negative binomial GLMM in simulation study and in real data analysis. We only used GLIMMIX to obtain parameter sets which in turn were used to simulate data for evaluating our method when the data-generating model is extremely misspecified. Also, we only used the GLIMMIX results for the genes that the estimation algorithm converged. In this scenario, each dataset were generated following the same procedure in Section 4.4.1, except the parameter sets come from the output of the GLIMMIX procedure applying to the LPS RNA-seq data, and read counts were simulated from a NB GLMM model instead of our model (4.1) and (4.2).

4.4.4 Simulation Results

We analyzed simulated datasets from three simulation scenarios using voomboot, voom, edgeR, DESeq2, oracle – the method that uses true correlation and unshrunken error variance, and oracle_shrunken – the method that uses true correlation and shrunken error variance. Of course, the two oracle procedures cannot be used in practice, but their inclusion provides a useful reference measure of the performance achieved if the within-gene correlations were known. Due to the numerical instability of GLMM, we do not have oracle and oracle_shrunken for the third simulation

scenario, therefore, the oracle methods are only available for the first two simulation scenarios when data were generated using our working model (4.1) and (4.2).

For all six analysis methods, p-value for testing the significance of the partial regression coefficients on the contrast of interest was calculated for each gene. These p-values were converted to q-values as described in Section 4.2.5, and genes with q-values no larger than 0.05 were declared to be DE. Using these p-values and q-values, we evaluated each method's performance based on four criteria: the relationship between empirical distribution of true null p-values and the uniform(0,1) distribution, the incurred FDR when FDR is nominally controlled at 5%, the number of true positive (NTP) detections of differential expression, and the partial area under the receiver operating characteristic curve (PAUC) corresponding to false positive rates less than or equal to 0.05. These performance criteria assess the validity of p-values, FDR control, power, and the ability to distinguish EE and DE genes from one another.

All simulation results in terms of the first criterion are displayed in Figures 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8. In simulation scenario 1 when data were generated using our working model with non-zero correlations, the empirical quantiles of the null p-values of three methods voomboot, oracle, and oracle_shrunken are very similar to the uniform(0,1) quantiles in all four contrasts of interest. On the other hand, the null p-values of voom, edgeR, and DESeq2 are very liberal for `Line` main effect, among the three methods, DESeq2 results in the most liberal null p-values. For the other three contrasts `Time`, `Time2-Time0`, and `Interaction`, voom and edgeR give very conservative p-values, while DESeq2 gives conservative p-values in the cases `Time2-Time0`, and liberal ones in the case `Time` and `Interaction`. In simulation scenario 2 when data were generated using our working model with zero correlations, as expected, voom, voomboot, oracle, and oracle_shrunken have the null p-values close to the uniform(0,1) distribution, while the other methods give liberal null p-values in all four contrasts, but the level of liberty of edgeR and DESeq2 is not as severe as those in simulation scenario 1. In the simulation scenario 3 when data were generated using a negative binomial GLMM with non-zero correlation incorporating in $\boldsymbol{G}$-side effect of the SAS procedure GLIMMIX, the null p-values from all methods depart from the uniform(0,1) distribution. The null

p-values of voomboot are slightly liberal in all contrasts, and closer to the uniform(0,1) distribution than the null p-values of the other methods. voom, edgeR, and DESeq2 still give very liberal null p-values for `Line` main effect, and very conservative null p-values in the other three contrasts.

The behavior of null p-values of all methods can be explained as follows. When within-gene correlations exist, as demonstrated in simulation scenario 1, the methods voom, edgeR and DESeq2 do not take within-unit correlation into account, therefore, they tend to underestimate the variances of within-subject contrasts such as `Line` main effect, therefore, inflate the corresponding test statistics values, resulting in liberal p-values. These methods also overestimate the variances of between-subject contrasts such as `Time`, `Time2-Time0`, `Interaction`, therefore deflate these test statistics values, resulting in conservative p-values. In simulation scenario 2, the empirical distribution of the null p-values of edgeR and DESeq2 slightly deviates from the uniform(0,1) distribution due to the non-existence of within-unit correlations among observations of simulated data. But both edgeR and DESeq2 still suffer from model misspecification because they use negative binomial generalized linear model. In simulation scenario 3, voomboot faces a severe model misspecification when the simulated data were generated from a negative binomial GLMM. Even though, because voomboot takes within-gene correlation into account, its null p-values deviate from the uniform(0,1) distribution less than the other three methods do.

The simulation results in terms of FDR control are summarized in Figure 4.9. In simulation scenario 1, voomboot is able to control FDR well for all four contrasts; while voom, edgeR and DESeq2 fail to control FDR for `Line` main effect with extremely high incurred FDR. For the other three contrasts `Time, Time2-Time0` and `Interaction`, both voom and edgeR are able to control FDR conservatively; meanwhile DESeq2 fails to control FDR for all four contrasts except `Time2-Time0` effect. In simulation scenario 2, voomboot and voom can control FDR; while edgeR and DESeq2 fail to do so. In simulation scenario 3, most methods fail to control FDR in all cases, except that voom and edgeR control FDR for the contrasts `Time, Time2-Time0` and `Interaction`; DESeq2 controls FDR for the contrasts `Time` and `Time2-Time0`. In all simulation scenarios, among the three methods voom, edgeR and DESeq2, DESeq2 gives the most liberal incurred FDR.

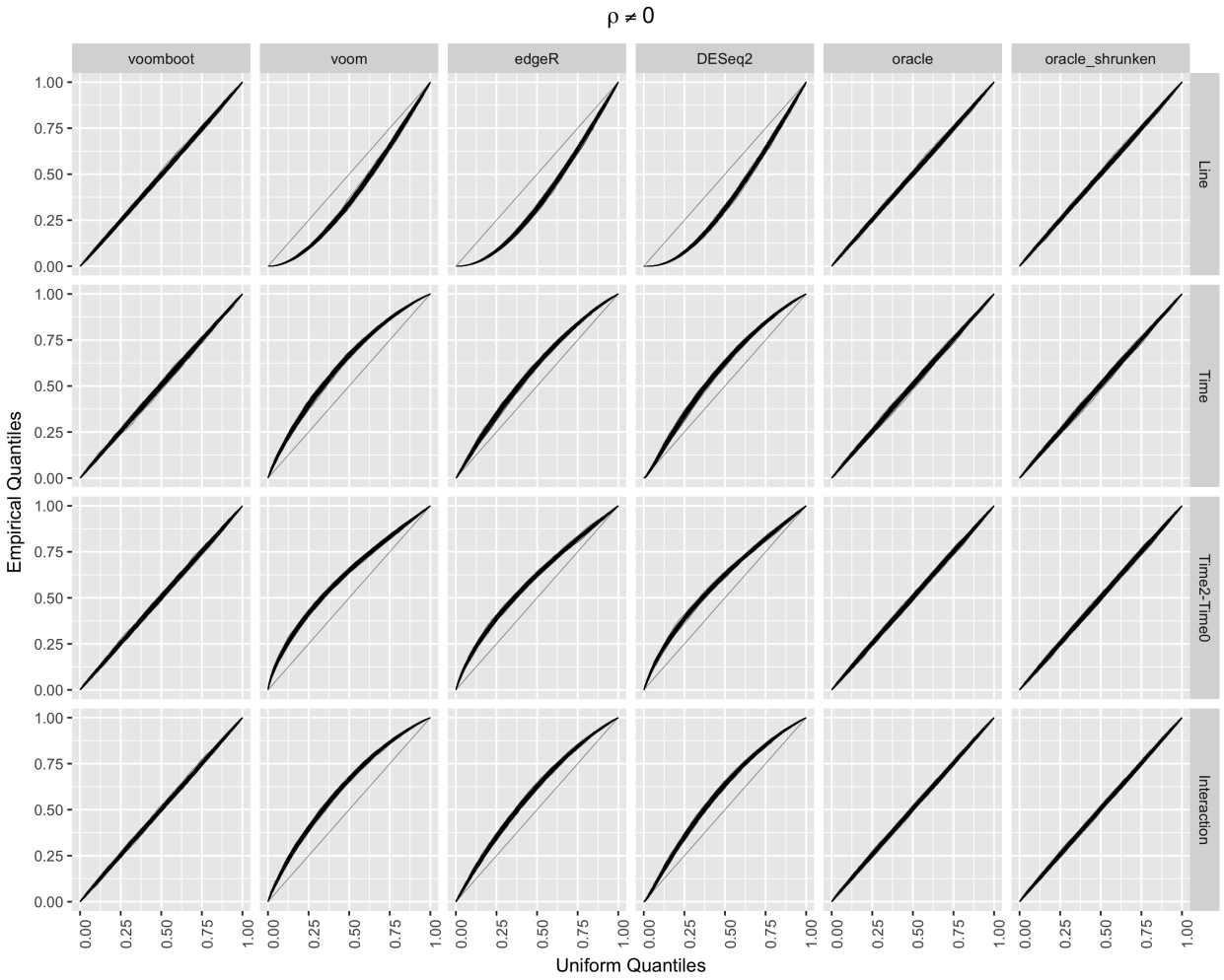

Figure 4.3: A plot of quantiles of null p-values versus quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 1. Each line represents the quantiles from a single simulation, the diagonal line represents the quantiles of the uniform(0,1) distribution.

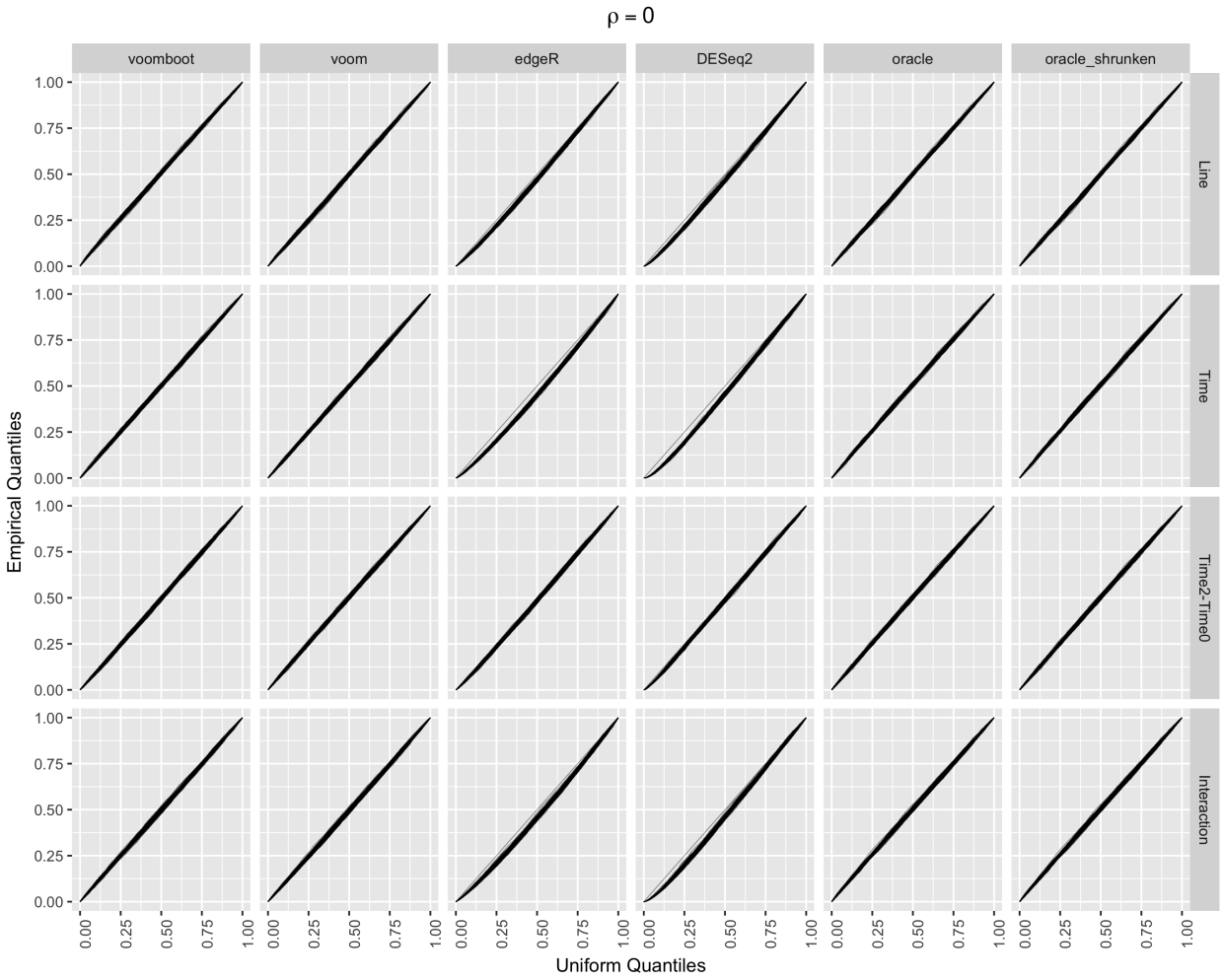

Figure 4.4: A plot of quantiles of null p-values versus quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 2. Each line represents the quantiles from a single simulation, the diagonal line represents the quantiles of the uniform(0,1) distribution.

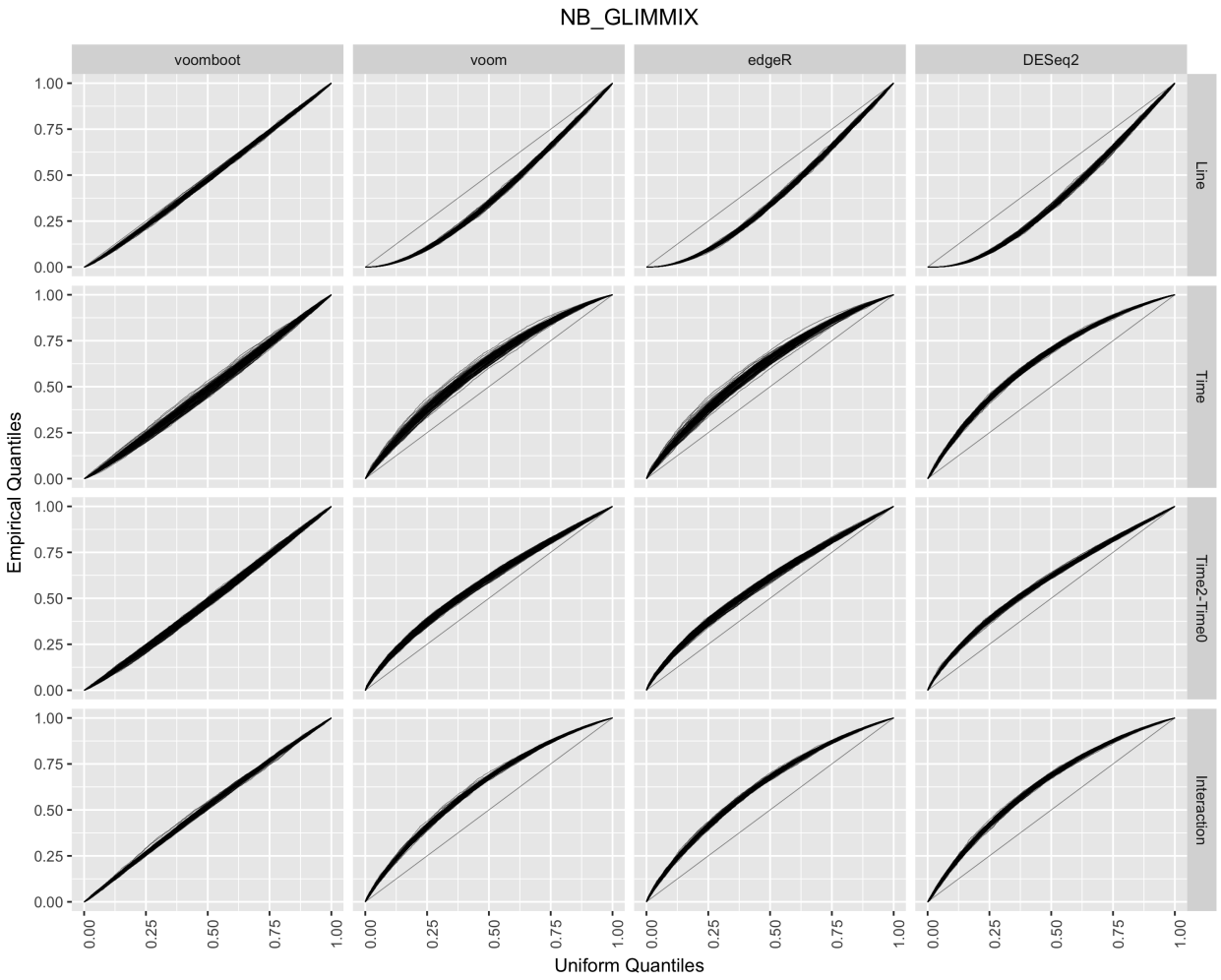

Figure 4.5: A plot of quantiles of null p-values versus quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 3. Each line represents the quantiles from a single simulation, the diagonal line represents the quantiles of the uniform(0,1) distribution.

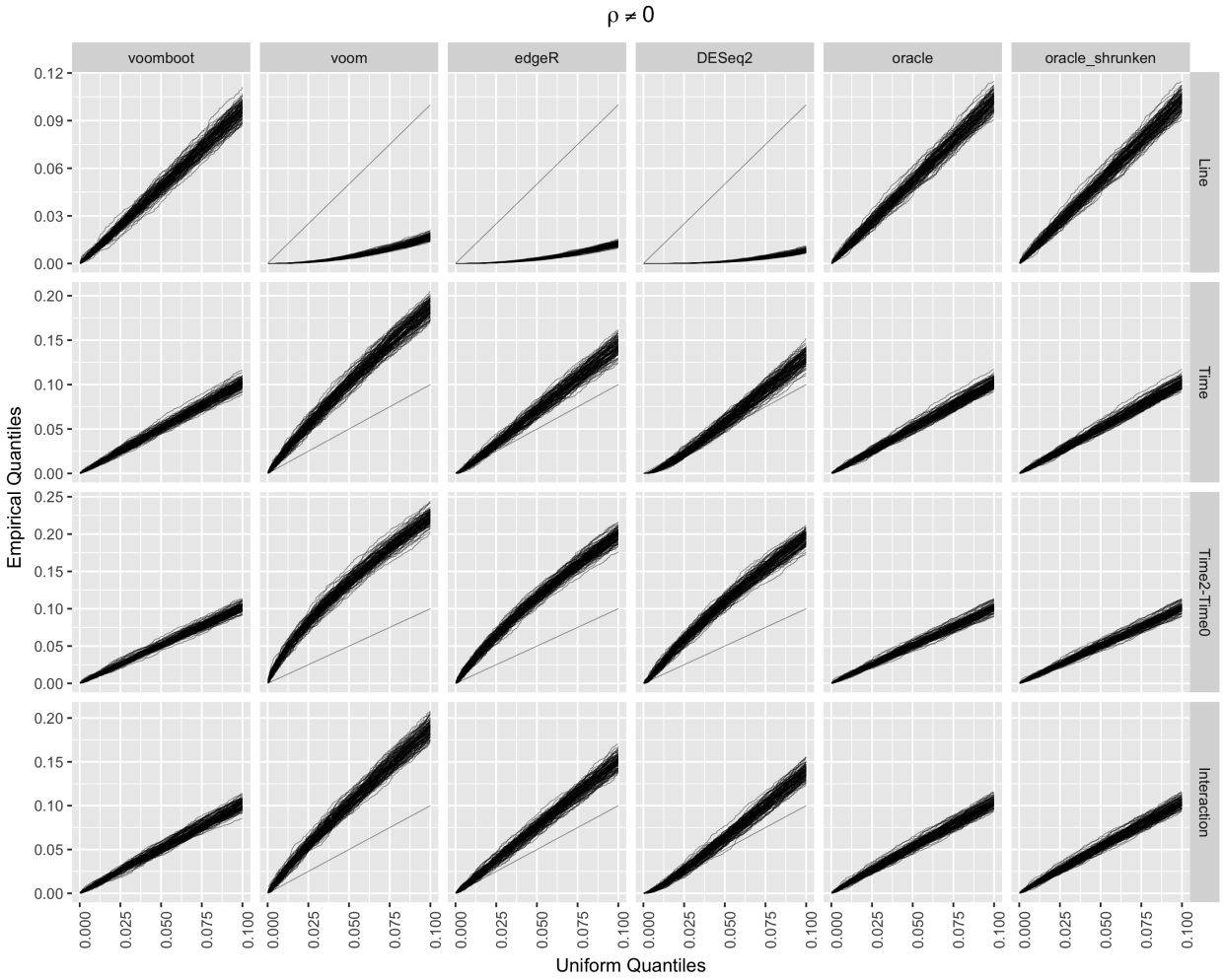

Figure 4.6: A plot of the less-than-10% quantiles of null p-values versus the less-than-10% quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 1. Each line represents the less-than-10% quantiles from a single simulation, the diagonal line represents the the less-than-10% quantiles of the uniform(0,1) distribution.

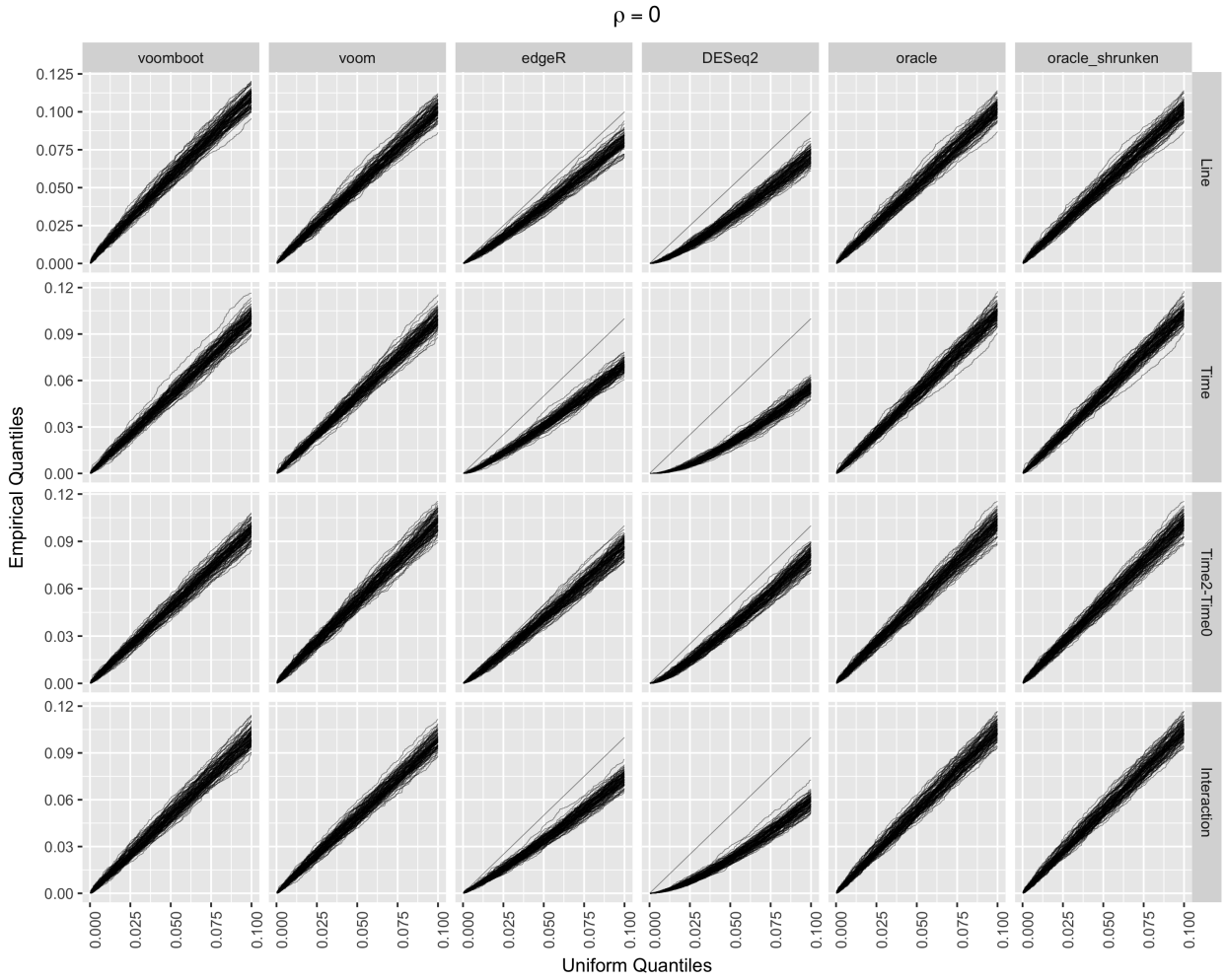

Figure 4.7: A plot of the less-than-10% quantiles of null p-values versus the less-than-10% quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 2. Each line represents the less-than-10% quantiles from a single simulation, the diagonal line represents the the less-than-10% quantiles of the uniform(0,1) distribution.

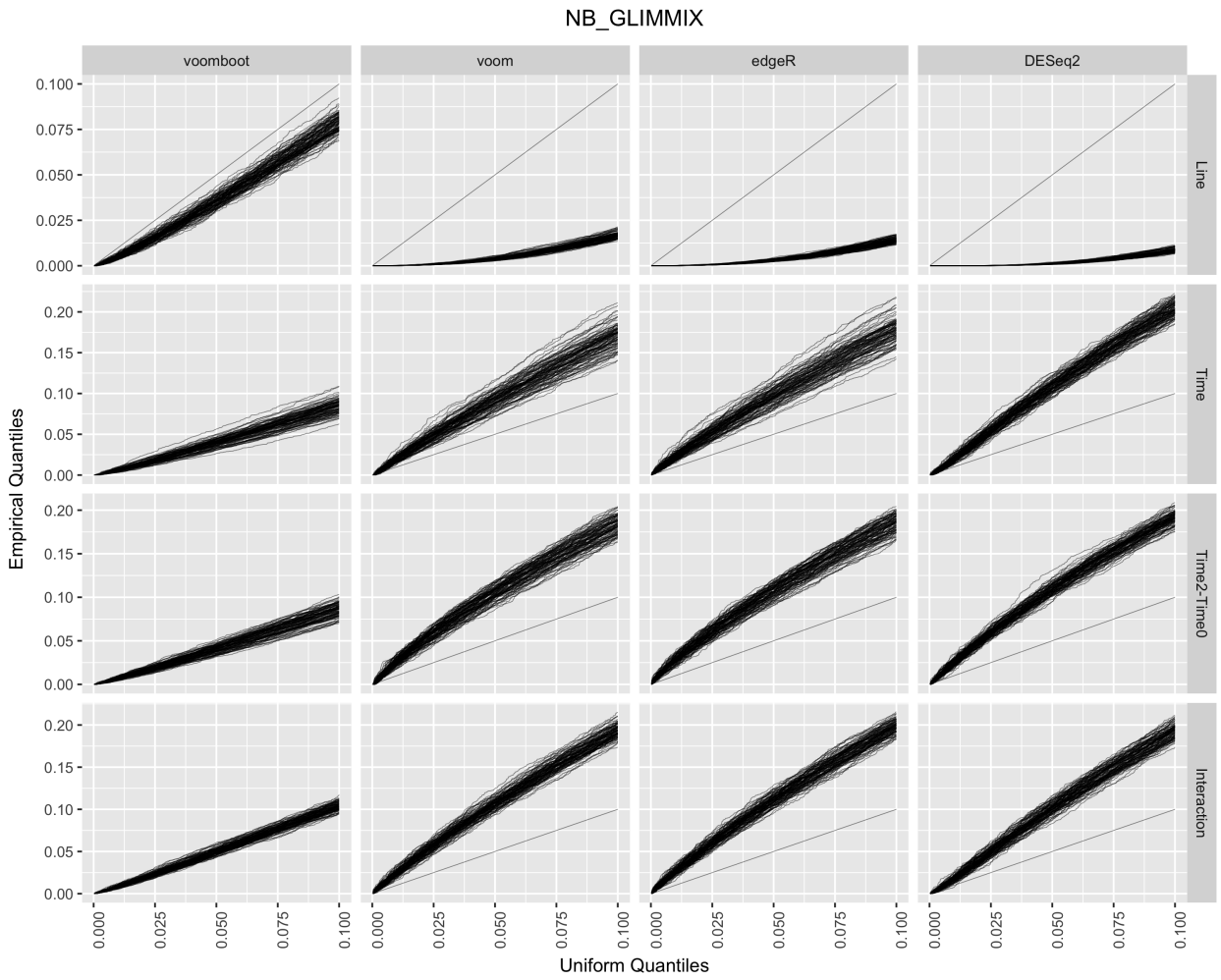

Figure 4.8: A plot of the less-than-10% quantiles of null p-values versus the less-than-10% quantiles of the uniform(0,1) distribution for all methods and contrasts in simulation scenario 3. Each line represents the less-than-10% quantiles from a single simulation, the diagonal line represents the the less-than-10% quantiles of the uniform(0,1) distribution.

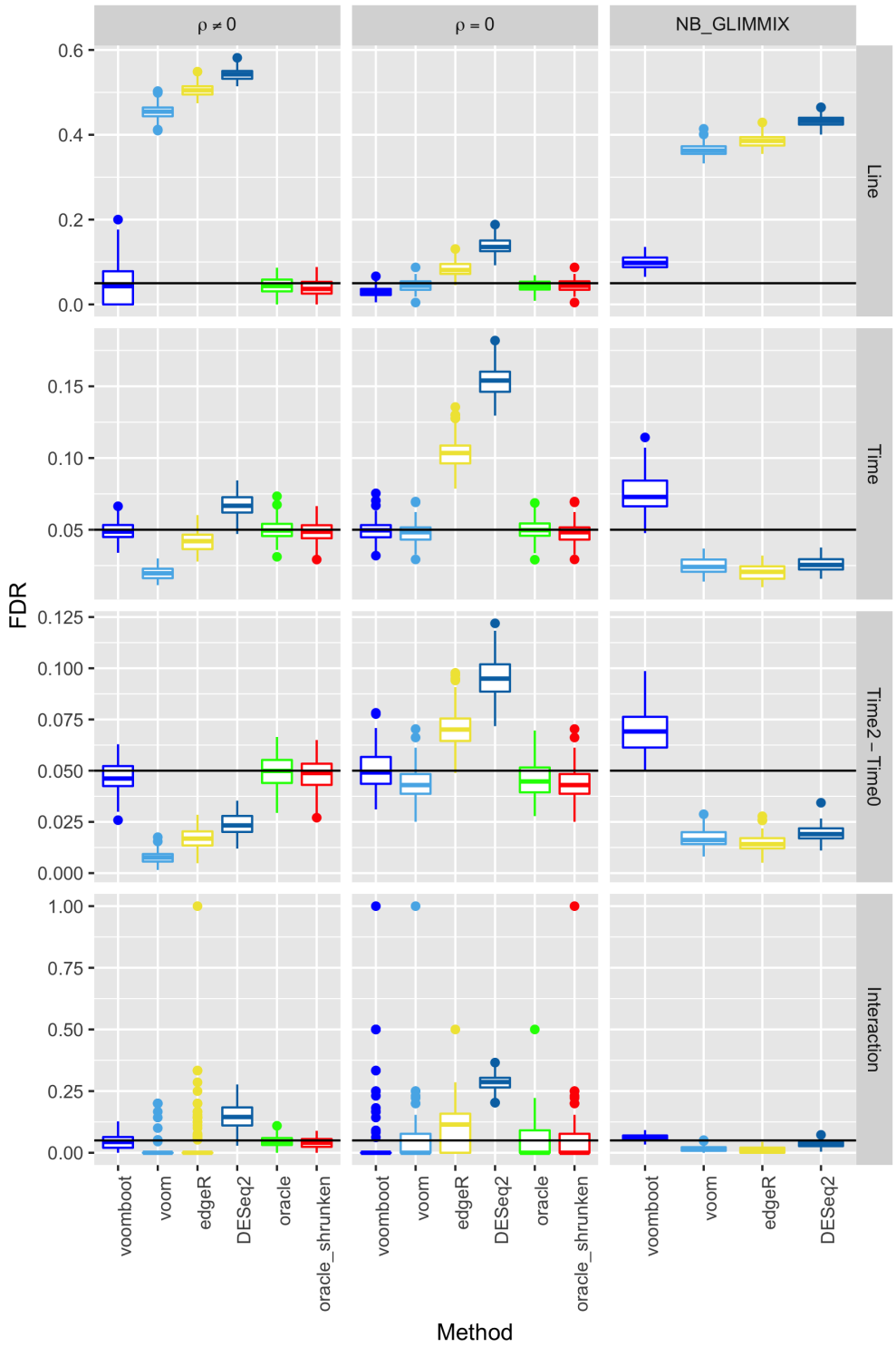

Figure 4.9: Boxplots of the incurred FDR when FDR is nominally controlled at 0.05 for all methods and all contrasts in 3 simulation scenarios. Each boxplot has 100 data points representing 100 simulated datasets.

The simulation results in terms of PAUC, the ability to distinguish DE and EE genes from one another, are presented in Figure 4.10. For all contrasts, voomboot outperforms all alternatives except voom in simulation scenario 2. This is obvious because in simulation scenario 2, voom is exactly oracle_shrunken. oracle_shrunken performs best as expected, so does voom. In all simulation scenarios, DESeq2 is the worst method in terms of PAUC among three methods voom, edgeR and DESeq2.

The simulation results in terms of power are shown in Figure 4.11. Since many methods fail to control FDR in many cases, it is hard to evaluate their power in all cases. For the contrast `Time2-Time0` in the simulation scenario 1 when voomboot, voom, edgeR and DESeq2 control FDR, it is clear that voomboot is the most powerful method.

4.5 Discussion

The proposed method voomboot provides a practical tool for identifying DE genes using RNA-seq data from repeated-measures designs. The idea is to use normalized log-counts and their associated precision weights in a general linear model pipeline for estimation, and then employ a parametric bootstrap procedure for hypothesis testing. Correlation among observations within each gene is accounted for using the continuous autoregressive correlation structure $CAR(1)$. Under our working model assumptions, simulation studies show the advantages of our method compared to the alternatives that do not account for the within-gene correlation induced by the repeated-measures structure. In particular, our method outperforms the alternatives that do not consider correlation among observations within gene in terms of FDR control and the ability to distinguish EE and DE genes from one another. Our method suffers when the model is extremely misspecified, such as when the true data-generating model follows a negative binomial GLMM. Our method is implemented in an R package available at https://github.com/ntyet/tcrmrnaseq.

The parametric bootstrap inference approach proposed in our paper can be easily extended to other RNA-seq designs that may contain factors whose effects are best modeled as random thanks

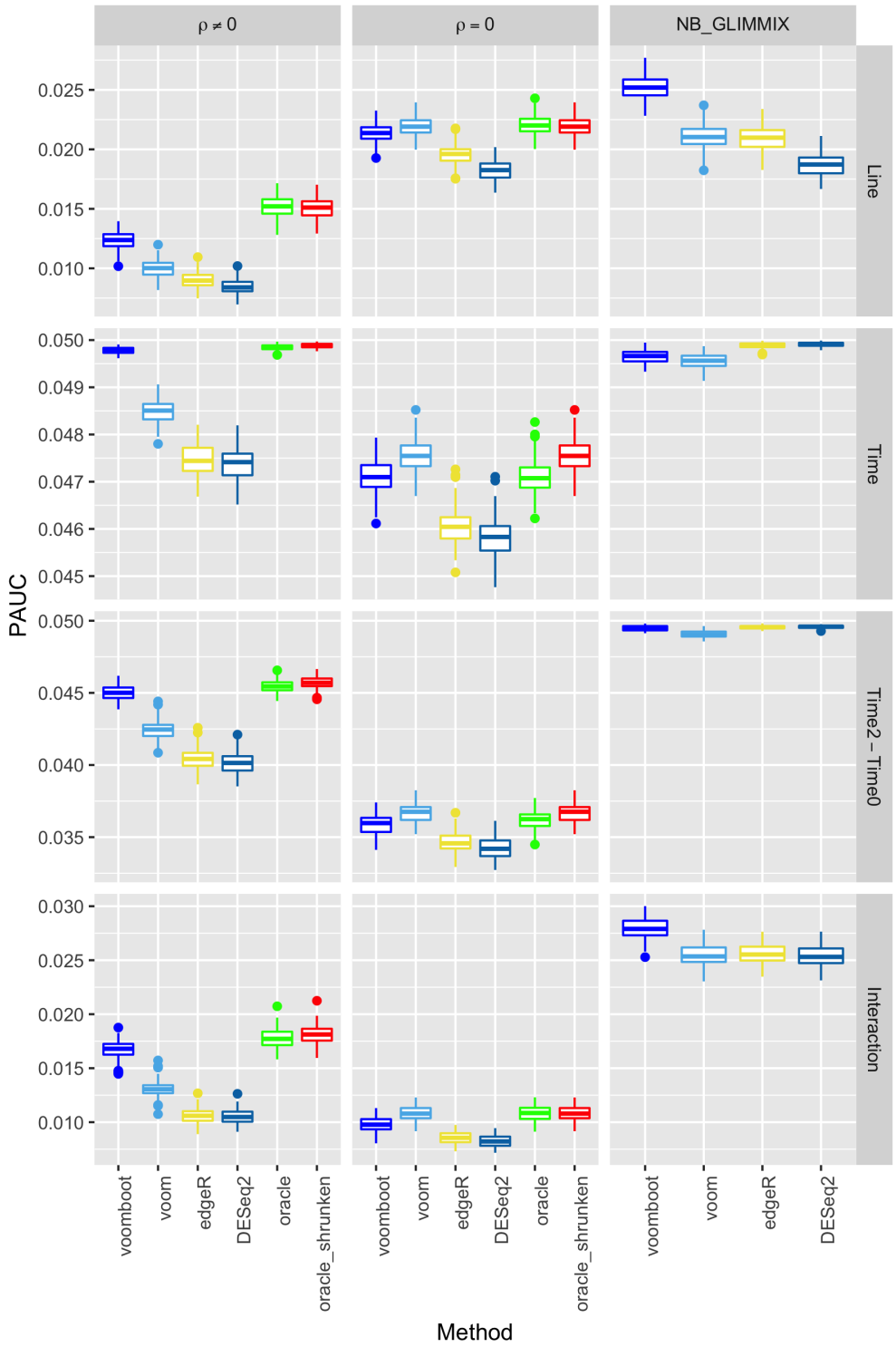

Figure 4.10: Boxplots of the partial area under the receiver operating characteristic curve (PAUC) when false positive rate is less than or equal to 0.05 for all methods and all contrasts in 3 simulation scenarios. Each boxplot has 100 data points representing 100 simulated datasets.

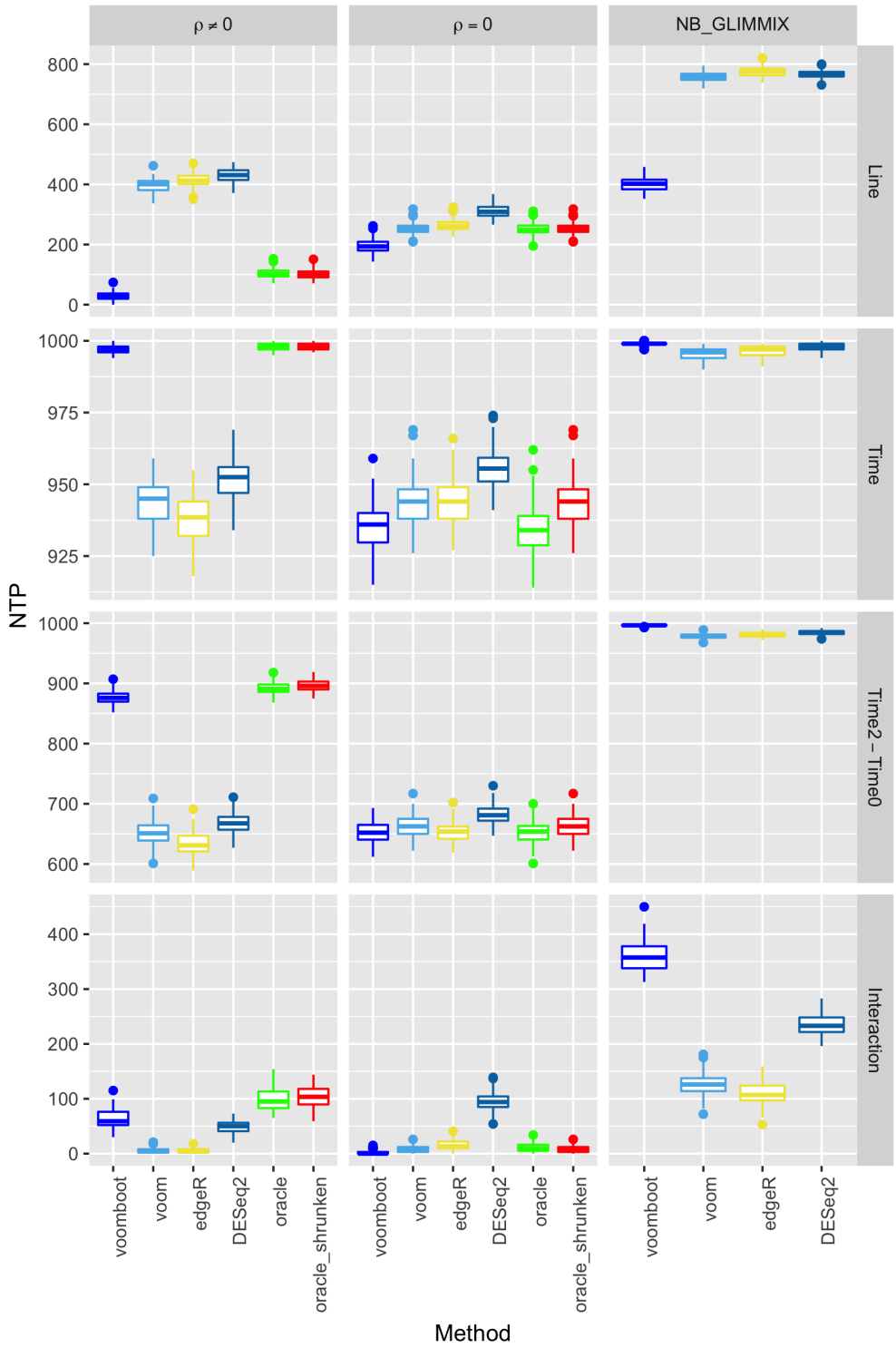

Figure 4.11: Boxplots of number of true positive (NTP) detections when FDR is nominally controlled at 0.05 for all methods and all effects in 3 simulation scenarios. Each boxplot has 100 data points representing 100 simulated datasets.

to the simple and straightforward application of linear model using normalized log-counts data. We also expect that the inference approach behaves well in such situations.

Our method is computationally intensive due to its utilization of a parametric bootstrap procedure to make inference. Our implementation of voomboot method in the `R` package `tcrmrnaseq` use parallelization to speed up the algorithm. Using 16 cores computer in parallel, it takes about 65 minutes to analyze 11911 genes of the LPS RNA-seq dataset. In a personal laptop with 4 cores, it takes about 4 hours and 20 minutes for such analysis.

It is worth to recall that our proposed method is not the only one that can account for the within-gene correlation among observations in the analysis of RNA-seq data or any other omic-count data. There are several other options that within-gene correlations can be handled.

First, one may use negative binomial or Poisson generalized linear mixed model, for examples, see Sun et al. (2016), Zhang et al. (2017). Sun et al. (2016) developed a negative binomial GLMM framework to analyze a time-course RNA-seq experiment at exon level, where they used smoothing spline to model time effect and group effect, and a random effect to model time dependency. However, the random effect does not reflect the general unequally spaced time point situation as shown in our motivating data example. On the other hand, Zhang et al. (2017) also proposed a negative binomial GLMM for microbiome data to detect significant taxa with respect to a factor of interest accounting for correlation among samples. The sample size in their working dataset is about several hundreds samples, which is not a typical sample size in RNA-seq experiments. Also, from our experience, a regular GLMM fit in RNA-seq context with the autoregressive correlation structure as in our motivating data example has shown to be numerically unstable, because it fails to converge for many genes.

Second, other approach is to use Kenward-Roger's method (KR) for normalized log-counts data in a general linear model framework. However, our extra simulation studies show KR method does not work well in RNA-seq data with the considered modeling assumption in terms of FDR control.

In conclusion, our proposed method works well under the general linear model framework compared to other alternative approaches. Moreover, our approach can be extended to other complex designs that may contain factors whose effects are best modeled as random.

Acknowledgments

This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), and by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) and the joint National Science Foundation (NSF)/NIGMS Mathematical Biology Program under award number R01GM109458. The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of USDA, NSF, or NIH.

Bibliography

Agniel, D. and Hejblum, B. P. (2017). Variance component score test for time-course gene set analysis of longitudinal RNA-seq data. *Biostatistics*, 18(4):589–604.

Äijö, T., Butty, V., Chen, Z., Salo, V., Tripathi, S., Burge, C. B., Lahesmaa, R., and Lähdesmäki, H. (2014). Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics*, 30(12):i113–i120.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, 11(1):94.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cui, S., Ji, T., Li, J., Cheng, J., and Qiu, J. (2016). What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Statistical Applications in Genetics and Molecular Biology*, 15(2):87–105.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to Bootstrap*. Chapman & Hall/CRC.

Jo, K., Kwon, H.-B., and Kim, S. (2014). Time-series RNA-seq analysis package (TRAP) and its application to the analysis of rice, Oryza sativa L. ssp. Japonica, upon drought stress. *Methods*, 67(3):364–372. Systems Biology with Omics Data.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997.

Keselman, H. J., Algina, J., Kowalchuk, R. K., and Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics - Simulation and Computation*, 27(3):591–604.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Leng, N., Li, Y., McIntosh, B. E., Nguyen, B. K., Duffin, B., Tian, S., Thomson, J. A., Dewey, C. N., Stewart, R., and Kendziorski, C. (2015). EBSeq-HMM: a Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics*, 31(16):2614–2622.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Liu, H. (2017). *Swine blood transcriptomics: Application and advancement*. PhD thesis, Graduate Theses and Dissertations, Iowa State University, Ames, IA, 50011.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lun, A. T. L., Chen, Y., and Smyth, G. K. (2016). It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. In Mathé, E. and Davis, S., editors, *Statistical Genomics: Methods and Protocols*, pages 391–416. Springer New York, New York, NY.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5):1544–6115.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628.

Nettleton, D., Hwang, J. T. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337.

Nueda, M. J., Tarazona, S., and Conesa, A. (2014). Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, 30(18):2598–2602.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014a). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014b). The role of spike-in standards in the normalization of RNA-seq. In *Statistical Analysis of Next Generation Sequencing Data*, pages 169–190. Springer.

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

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Spies, D. and Ciaudo, C. (2015). Dynamics in transcriptomics: Advancements in RNA-seq time course and downstream analysis. *Computational and Structural Biotechnology Journal*, 13:469–477.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):187–205.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842.

Sun, X., Dalpiaz, D., Wu, D., S. Liu, J., Zhong, W., and Ma, P. (2016). Statistical inference for time course RNA-seq data using a negative binomial mixed-effect model. *BMC Bioinformatics*, 17(1):324.

Wise, A. and Bar-Joseph, Z. (2015). SMARTS: reconstructing disease response networks from multiple individuals using time series gene expression data. *Bioinformatics*, 31(8):1250–1257.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 4:233–243.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4.

CHAPTER 5. A HISTOGRAM-BASED METHOD FOR FALSE DISCOVERY RATE CONTROL IN TWO INDEPENDENT EXPERIMENTS

A paper in preparation

Yet Nguyen, Megan Orr, Peng Liu, and Dan Nettleton

Abstract

This paper presents a new method to estimate and control false discovery rate (FDR) when identifying simultaneous signals in two independent experiments. In one experiment, thousands or millions features are tested for significance with respect to some test of interest. In a second experiment, the same features are also tested for significance. Researchers are interested in identifying simultaneous signals, i.e., features that are significant in both experiments. We develop an FDR estimation and control procedure that is a generalization of the histogram-based FDR estimation and control procedure for one experiment proposed by Nettleton et al. (2006); Liang and Nettleton (2012). We show that our method performs well and better than other existing methods.

5.1 Introduction

Multiple hypothesis testing is a popular topic due to the demand from many scientific fields, such as genomics, where researchers are interested in determining which of thousands or millions genetic features are affected by a treatment or predictive of a trait. False discovery rate (FDR) is a well-known error rate to consider in such applications. The most heavily cited papers on statistical theory and methods for FDR include Benjamini and Hochberg (1995), Storey (2002), and Storey et al. (2004). These and most other FDR methods were developed for multiple hypothesis testing conducted within a single experiment.

Currently, there is considerable interest in multiple testing for two or more independent experiments. For example, in genomics, there is a phenomenon called pleiotropy (Grüneberg, 1938) in which a single gene influences two or more seemingly unrelated phenotypic traits. Much research about evidence for pleiotropy has been presented in literature; see, e.g., Sivakumaran et al. (2011), Cross-Disorder Group of the Psychiatric Genomics Consortium et al. (2013a,b); Andreassen et al. (2013); Chung et al. (2014) and references therein. Another example is in replicability research where the goal is to find significant features/signals that are replicable in two or more experiments (Orr, 2012; Bogomolov and Heller, 2013; Heller et al., 2014).

There are several publications addressing the challenges of FDR estimation and control in multiple experiments; see, e.g., Phillips and Ghosh (2014); Chung et al. (2014); Heller and Yekutieli (2014) and Zhao and Nguyen (2017) (under review, and hereafter referred to as ZN). The FDR estimation and control procedures in Phillips and Ghosh (2014) and Chung et al. (2014) use p-values as inputs, while in Heller and Yekutieli (2014) and ZN, the inputs can be either p-values or general test statistics. The performance as well as limitation of these methods are discussed in ZN.

ZN showed that their method controls FDR well and outperforms methods of Chung et al. (2014) and Heller and Yekutieli (2014) in the setting of sparse simultaneous signals. However, the method of ZN also has its own limitations. Even though their method works well when simultaneous signals are sparse, there is no guarantee that the method performs satisfactorily in non-sparse cases.

In this paper, we propose a new method for FDR estimation and control when identifying simultaneous signals in two independent experiments. Our method can overcome the limitations of existing methods. The inputs of our method are p-values from these two experiments, with one pair of p-values for each feature common to both experiments. We employ an extension of the histogram-based method for a single experiment (Nettleton et al., 2006; Liang and Nettleton, 2012) to estimate FDR and propose a new q-value calculation in two dimensions, similar to the q-value calculation in one dimension (Storey, 2002), which can be readily used to control FDR. Our method does not require sparsity of simultaneous signals. Therefore, it can be used in a wider array of applications. We also show that, asymptotically, our method estimates FDR better than

the method of ZN. In particular, we show that, asymptotically the bias of our FDR estimator is less than that of ZN's method. The asymptotic results are also supported by a simulation study.

Our paper is organized as follows. Section 5.2 is devoted to the description of our proposed method, including an FDR estimation procedure and a procedure to identify simultaneous signals that controls FDR. Section 5.3 presents asymptotic results for our method and a comparison with the method of ZN and is followed by a simulation study in Section 5.4. Some discussion and concluding remarks are presented in Section 5.5.

5.2 Methods

5.2.1 Statistical setting

For $j = 1, \ldots, m$ and $k = 1, 2$, let P_{kj} be the p-value for feature j in experiment k. Let $I_{kj} = 1$ if feature j from the experiment k is a false null, otherwise, $I_{kj} = 0$. For now, we assume all elements of the sequence $\{P_{kj}\}_{j=1}^{m}$ are independent.

Let

$$\mathcal{A}_{ab} = \{j : I_{1j} = a, I_{2j} = b\}$$
$$V_{ab}(t_1, t_2) = \sum_{j \in \mathcal{A}_{ab}} \mathbb{1}(P_{1j} \leq t_1, P_{2j} \leq t_2)$$
$$m_{ab} = \text{Card}\{\mathcal{A}_{ab}\}.$$

It is assumed that P_{kj} are stochastically smaller when $I_{kj} = 1$ than when $I_{kj} = 0$, so that

$$P_{kj}|I_{kj} = 0 \sim F_k^0; \quad P_{kj}|I_{kj} = 1 \sim F_{kj}^1; \quad F_{kj}^1 \geq F_k^0, \tag{5.1}$$

where the null distributions F_k^0 are uniform distributions while the alternative distributions F_{kj} are unknown and may be different for different features.

Furthermore, we assume that the following conditions hold. A related assumption is also used in ZN. We will revisit these conditions in the discussion section. In fact, the main results of our paper still hold under some weaker assumptions.

Assumption 5.2.1 *There exist continuous distribution functions* $F_1^1(t_1), F_2^1(t_2)$ *and* $G^1(t_1, t_2)$ *such that as* $p \to \infty$, *for* $a, b = 0, 1$,

$$\frac{1}{m_{a1}} \sum_{j \in \mathcal{A}_{a1}} F_{2j}^1(t_2) \to F_2^1(t_2),$$
$$\frac{1}{m_{1b}} \sum_{j \in \mathcal{A}_{1b}} F_{1j}^1(t_1) \to F_1^1(t_1),$$
$$\frac{1}{m_{11}} \sum_{j \in \mathcal{A}_{11}} F_{1j}^1(t_1) F_{2j}^1(t_2) \to G^1(t_1, t_2) = F_1^1(t_1) F_2^1(t_2)$$

uniformly in $t_1, t_2 \in [0, 1]$. *There also exist* π_{ab} *such that* $m_{ab}/m \to \pi_{ab}$ *for* $a, b = 0, 1$ *as* $p \to \infty$.

From Assumption 5.2.1, it is easy to see that $\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$. Next, let

$$\hat{F}_k(t_k) = m^{-1} \sum_{j=1}^{m} \mathbb{1}(P_{kj} \le t_k)$$

$$F_k(t_k) = \pi_k^0 F_k^0(t_k) + \pi_k^1 F_k^1(t_k), \quad \text{where } \pi_k^0 + \pi_k^1 = 1,\ k = 1, 2,\ \pi_1^0 = \pi_{00} + \pi_{01},\ \pi_2^0 = \pi_{00} + \pi_{10}$$

$$\hat{G}^0(t_1, t_2) = m^{-1}(V_{00}(t_1, t_2) + V_{10}(t_1, t_2) + V_{01}(t_1, t_2)) = m^{-1} \sum_{j \notin \mathcal{A}_{11}}^{m} \mathbb{1}(P_{1j} \le t_1, P_{2j} \le t_2)$$

$$G^0(t_1, t_2) = \pi_{00} F_1^0(t_1) F_2^0(t_2) + \pi_{10} F_1^1(t_1) F_2^0(t_2) + \pi_{01} F_1^0(t_1) F_2^1(t_2)$$

$$\hat{G}^1(t_1, t_2) = m_{11}^{-1} \sum_{j \in \mathcal{A}_{11}} \mathbb{1}(P_{1j} \le t_1, P_{2j} \le t_2)$$

$$\hat{G}(t_1, t_2) = m^{-1} R(t_1, t_2) = m^{-1} \sum_{j=1}^{m} \mathbb{1}(P_{1j} \le t_1, P_{2j} \le t_2)$$

$$\begin{aligned} G(t_1, t_2) &= \pi_{00} F_1^0(t_1) F_2^0(t_2) + \pi_{10} F_1^1(t_1) F_2^0(t_2) + \pi_{01} F_1^0(t_1) F_2^1(t_2) + \pi_{11} G^1(t_1, t_2) \\ &= \pi_{00} F_1^0(t_1) F_2^0(t_2) + \pi_{10} F_1^1(t_1) F_2^0(t_2) + \pi_{01} F_1^0(t_1) F_2^1(t_2) + \pi_{11} F_1^1(t_1) F_2^1(t_2). \end{aligned}$$

F_k, G^0, G^1 and G are the limit distributions of the empirical functions $\hat{F}_k, \hat{G}^0, \hat{G}^1$ and $\hat{G}$, respectively.

5.2.2 False Discovery Proportion and False Discovery Rate

When conducting multiple tests in two independent experiments, depending on the significant thresholds t_1, t_2, there are different possible outcomes in terms of true discoveries and false discoveries in accordance to the true status of the features. A summary of the outcomes is presented in Table 5.1.

Table 5.1: Possible outcome of a hypothesis testing procedure for two independent experiments.

(I_{1j}, I_{2j})	Not discovered	Discovered	Total
$(0,0)$	U_{00}	V_{00}	m_{00}
$(1,0)$	U_{10}	V_{10}	m_{10}
$(0,1)$	U_{01}	V_{01}	m_{01}
$(1,1)$	U_{11}	V_{11}	m_{11}
Total	$m-R$	R	m

Different from one experiment, a false discovery in two independent experiments occurs when a feature is declared to be a discovery in both experiments, when in reality, that feature is null in one or both experiments. The false discovery proportion, denoted as FDP, is defined as the proportion of false discoveries among all discoveries:

$$\mathrm{FDP}(t_1,t_2) := \frac{V_{00}(t_1,t_2)+V_{10}(t_1,t_2)+V_{01}(t_1,t_2)}{1 \vee R(t_1,t_2)} = \frac{\hat{G}^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)}. \tag{5.2}$$

Then, the false discovery rate, denoted as FDR, is defined as the expected value of the false discovery proportion:

$$\mathrm{FDR}(t_1,t_2) := \mathrm{E}\left(\mathrm{FDP}(t_1,t_2)\right) = \mathrm{E}\left(\frac{\hat{G}^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)}\right), \tag{5.3}$$

where

$$m^{-1} \vee \hat{G}(t_1,t_2) := \max\{m^{-1}, \hat{G}(t_1,t_2)\}.$$

Both FDP and FDR are unknown, therefore, we need a good estimation procedure for these quantities, which is one of the focuses in this paper.

5.2.3 False Discovery Rate Estimation Procedure

From (5.3), a natural estimator of FDR is a ratio whose the denominator is $m^{-1} \vee \hat{G}(t_1,t_2)$ and the numerator is an estimator of $E(\hat{G}^0(t_1,t_2))$. Using Assumption 5.2.1, a simple calculation of this expectation gives us

$$E(\hat{G}^0(t_1,t_2)) = m^{-1}\left[E(V_{00}(t_1,t_2)+E(V_{10}(t_1,t_2)+E(V_{01}(t_1,t_2)\right]$$

$$\to \pi_{00} m_{00}^{-1} E(V_{00}(t_1, t_2)) + \pi_{10} m_{10}^{-1} E(V_{10}(t_1, t_2)) + \pi_{01} m_{01}^{-1} E(V_{01}(t_1, t_2))$$
$$\to \pi_{00} F_1^0(t_1) F_2^0(t_2) + \pi_{10} F_1^1(t_1) F_2^0(t_2) + \pi_{01} F_1^0(t_1) F_2^1(t_2) = G^0(t_1, t_2)$$

as $m \to \infty$. Therefore, an estimate of $G^0(t_1, t_2)$ can be served as the estimate for $E(\hat{G}^0(t_1, t_2))$. To do this, we need to estimate $\pi_{00}, \pi_{10}, \pi_{01}, F_1^1(t_1)$ and $F_2^1(t_2)$. Motivated by the histogram-based FDR estimation procedure (Nettleton et al., 2006; Liang and Nettleton, 2012), these quantities can be estimated as follows

- First, we estimate π_k^0 by:

$$\begin{aligned}
\hat{\pi}_k^0(\lambda_k) &= \frac{m^{-1} \sum_{j=1}^m \mathbb{1}(P_{kj} > \lambda_k)}{1 - \lambda_k} \\
&= \frac{m^{-1} \sum_{j=1}^m (1 - \mathbb{1}(P_{kj} \leq \lambda_k))}{1 - \lambda_k} \\
&= \frac{1 - \hat{F}_1(\lambda_k)}{1 - F_k^0(\lambda_k)} \quad \text{for some } \lambda_k \in [0, 1) \quad k = 1, 2.
\end{aligned} \tag{5.4}$$

  The reasoning behind this estimator is that a p-value from a false null tends to be small, therefore, if λ_k is large enough, a feature with p-value larger than λ_k is more likely to be a true null. As a consequence, most p-values exceeding λ_k are more likely to be true nulls, having uniform (0,1) distribution. This estimator depends on a parameter $\lambda_k \in [0, 1)$ for $k = 1, 2$. There are different ways to select λ_k. In this paper, we use the histogram-based method of Nettleton et al. (2006) to choose λ_k.

- Estimator of π_{00}: Using the same reasoning as above, we propose an estimator of π_{00} as follows

$$\begin{aligned}
\hat{\pi}_{00}(\lambda_1, \lambda_2) &= \frac{m^{-1} \sum_{j=1}^m \mathbb{1}(P_{1j} > \lambda_1, P_{2j} > \lambda_2)}{(1 - \lambda_1)(1 - \lambda_2)} \\
&= \frac{m^{-1} \sum_{j=1}^m \mathbb{1}(P_{1j} > \lambda_1) \mathbb{1}(P_{2j} > \lambda_2)}{(1 - \lambda_1)(1 - \lambda_2)} \\
&= \frac{m^{-1} \sum_{j=1}^m (1 - \mathbb{1}(P_{1j} \leq \lambda_1))(1 - \mathbb{1}(P_{2j} \leq \lambda_2))}{(1 - \lambda_1)(1 - \lambda_2)} \\
&= \frac{1 - \hat{F}_1(\lambda_1) - \hat{F}_2(\lambda_2) + \hat{G}(\lambda_1, \lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))}.
\end{aligned} \tag{5.5}$$

- Estimator of π_{01}: Since $\pi_{01} = \pi_1^0 - \pi_{00}$, a natural estimator of π_{01} is

$$\hat{\pi}_{01}(\lambda_1, \lambda_2) = \hat{\pi}_1^0(\lambda_1) - \hat{\pi}_{00}(\lambda_1, \lambda_2). \tag{5.6}$$

- Estimator of π_{10}: Similarly, since $\pi_{10} = \pi_2^0 - \pi_{00}$, we estimate π_{10} by

$$\hat{\pi}_{10}(\lambda_1, \lambda_2) = \hat{\pi}_2^0(\lambda_2) - \hat{\pi}_{00}(\lambda_1, \lambda_2). \tag{5.7}$$

- Estimator of $F_k^1(t_k)$: Since $F_k(t_k) = \pi_k^0 F_k^0(t_k) + (1 - \pi_k^0)F_k^1(t_k)$, we can estimate $F_k^1(t_k)$ by

$$\hat{F}_k^1(t_k, \lambda_k) = \frac{\hat{F}_k(t_k) - F_k^0(t_k)\hat{\pi}_k^0(\lambda_k)}{1 - \hat{\pi}_k^0(\lambda_k)}. \tag{5.8}$$

Finally, we estimate FDR by

$$\widehat{\mathrm{FDR}}^{\lambda_1,\lambda_2}(t_1, t_2) := \frac{\hat{\pi}_{00}(\lambda_1, \lambda_2)F_1^0(t_1)F_2^0(t_2) + \hat{\pi}_{10}(\lambda_1, \lambda_2)\hat{F}_1^1(t_1, \lambda_1)F_2^0(t_2) + \hat{\pi}_{01}(\lambda_1, \lambda_2)F_1^0(t_1)\hat{F}_2^1(t_2, \lambda_2)}{m^{-1} \vee \hat{G}(t_1, t_2)}. \tag{5.9}$$

Note again that, in general, λ_k can be any value in $[0, 1)$. In practice, λ_k are chosen by the histogram-based method of Nettleton et al. (2006). Then one of the main results of our paper is the following theorem.

Theorem 5.2.1 *For any* $0 \leq \lambda_1, \lambda_2 < 1$, *and* $0 < \delta_1, \delta_2 \leq 1$,

$$\lim_{m\to\infty} \inf_{t_1 \geq \delta_1, t_2 \geq \delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1, t_2) - \mathrm{FDP}(t_1, t_2) \right) \geq 0 \tag{5.10}$$

and

$$\lim_{m\to\infty} \inf_{t_1 \geq \delta_1, t_2 \geq \delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1, t_2) - \mathrm{FDR}(t_1, t_2) \right) \geq 0. \tag{5.11}$$

Theorem 5.2.1 states that our FDR estimation procedure is asymptotically conservative in estimating both FDR and FDR. This result is similar to the result of ZN. This result is also similar to the result in Storey (2002), Storey et al. (2004), and Liang and Nettleton (2012) for one experiment.

5.2.4 A Procedure to Identify Simultaneous Signals

In this section, we present a simultaneous signals identification procedure that can control FDR. First, suppose that $\{Q_{kj}\}_{j=1}^m$ are the corresponding q-values of $\{P_{kj}\}_{j=1}^m$. Here, q-values $\{Q_{kj}\}_{j=1}^m$ are calculated by the method of Storey (2002). When computing these q-values for experiment k, an estimate of the proportion of true nulls π_k^0 is needed, and again, we use $\hat{\pi}_k^0(\lambda_k)$ as an estimate of π_k^0 with λ_k estimated by the histogram-based method of Nettleton et al. (2006). Let $Q_j^{max} = \max\{Q_{1j}, Q_{2j}\}$. Without loss of generality, assume $Q_j^{max} \leq Q_{j+1}^{max}$ for $j = 1, \ldots, m-1$. Define

$$t_{kj} = \max\{P_{ki} : i \in \mathcal{Q}(k,j)\}, \quad k = 1,2;\ j = 1, \ldots, m, \tag{5.12}$$

where

$$\mathcal{Q}(k,j) = \left\{i \in \{1, \ldots, m\} : Q_{ki} \leq Q_j^{max}\right\}. \tag{5.13}$$

There are several nice properties of (t_{1j}, t_{2j}) as shown in the following lemma.

Lemma 5.2.1 *Suppose $(t_{1j}, t_{2j})_{j=1}^m$ are defined as in (5.12). Then*

1. *$(t_{1j}, t_{2j})_{j=1}^m$ are well-ordered, i.e., $t_{kj} \leq t_{kj'}$ for $k = 1,2$; $1 \leq j \leq j' \leq m$.*
2. *$P_{kj} \leq t_{kj}$ for all $k = 1,2; j = 1, \ldots, m$.*
3. *If $Q_j^{max} < Q_{j'}^{max}$ for some $1 \leq j < j' \leq m$ then $P_{1j'} > t_{1j}$ or $P_{2j'} > t_{2j}$.*

Proof:

1. For $1 \leq j \leq j' \leq m$, since $Q_j^{max} \leq Q_{j'}^{max}$ and by the definition (5.13) of $\mathcal{Q}(k,j)$, we have

$$\mathcal{Q}_{kj} \subset \mathcal{Q}_{kj'}.$$

Therefore

$$\max\{P_{ki} : i \in \mathcal{Q}_{kj}\} \leq \max\left\{P_{ki} : i \in \mathcal{Q}_{kj'}\right\},$$

which together with the definition (5.12) of t_{kj} implies

$$t_{kj} \leq t_{kj'} \quad \text{for} \quad k = 1,2.$$

2. For all $k = 1, 2; j = 1, \ldots, m$, $j \in \mathcal{Q}_{kj}$, which together with (5.12) implies $P_{kj} \leq t_{kj}$.

3. We prove this by contradiction, i.e., suppose that $P_{kj'} \leq t_{kj}$ for $k = 1, 2$. For each k, by the definition of q-values (Storey, 2002),

$$\begin{aligned} Q_{kj'} &= \inf_{t \geq P_{kj'}} \{\text{FDR}(t)\} \\ &\leq \inf_{t \geq t_{kj}} \{\text{FDR}(t)\} \\ &= \inf_{t \geq P_{ki}} \{\text{FDR}(t)\} \text{ for some } i \in \mathcal{Q}(k, j) \text{ (by (5.12))} \\ &= Q_{ki} \text{ for some } i \in \mathcal{Q}(k, j) \\ &\leq Q_j^{max}. \end{aligned}$$

   Therefore,

$$Q_{j'}^{max} = \max\{Q_{1j'}, Q_{2j'}\} \leq Q_j^{max}$$

   which contradicts $Q_j^{max} < Q_{j'}^{max}$. Hence, $P_{1j'} > t_{1j}$ or $P_{2j'} > t_{2j}$.

□

Now, for a given FDR level α, let

$$j^* = \max\left\{j : \widehat{\text{FDR}}^{(\lambda_1, \lambda_2)}(t_{1j}, t_{2j}) \leq \alpha\right\}. \tag{5.14}$$

Then, a feature j is declared as a simultaneous signal if

$$P_{1j} \leq t_{1j^*} \quad \text{and} \quad P_{2j} \leq t_{2j^*}.$$

The second main result of our paper is the following theorem stating that under some conditions, this decision rule controls FDR asymptotically at level α.

Theorem 5.2.2 *Suppose that Assumption 5.2.1 holds. Assume further that there exist $\delta_1, \delta_2 > 0$ such that*

$$\lim \inf_{m \to \infty} t_{1j^*} = \delta_1; \lim \inf_{m \to \infty} t_{2j^*} = \delta_2.$$

Then

$$\lim \sup_{m \to \infty} \text{FDR}(t_{1j*}, t_{2j*}) \leq \alpha.$$

5.3 Proofs of Asymptotic Results

Before proving Theorem 5.2.1, we need the following lemmas.

Lemma 5.3.1 *Given Assumption 5.2.1, the followings hold.*

1. $\lim_{m\to\infty} \sup_{t\in[0,1]} |\hat{F}_k(t) - F_k(t)| = 0$ *a.s.* $k = 1, 2$.

2. $\lim_{m\to\infty} \sup_{t_1,t_2\in[0,1]} |\hat{G}(t_1, t_2) - G(t_1, t_2)| = 0$ *a.s.*

3. $\lim_{m\to\infty} \sup_{t_1,t_2\in[0,1]} |\hat{G}^0(t_1, t_2) - G^0(t_1, t_2)| = 0$ *a.s.*

4. $$\lim_{m\to\infty} \sup_{t_1\geq\delta_1,t_2\geq\delta_2} \left|\mathrm{FDP}(t_1, t_2) - \frac{G^0(t_1, t_2)}{m^{-1} \vee \hat{G}(t_1, t_2)}\right| = \lim_{m\to\infty} \sup_{t_1\geq\delta_1,t_2\geq\delta_2} \left|\frac{\hat{G}^0(t_1, t_2) - G^0(t_1, t_2)}{m^{-1} \vee \hat{G}(t_1, t_2)}\right| = 0$$

   a.s.

5. $$\lim_{m\to\infty} \sup_{t_1\geq\delta_1,t_2\geq\delta_2} |\mathrm{FDP}(t_1, t_2) - \mathrm{FDR}(t_1, t_2)| = \lim_{m\to\infty} \sup_{t_1\geq\delta_1,t_2\geq\delta_2} \left|\frac{\hat{G}^0(t_1, t_2)}{m^{-1} \vee \hat{G}(t_1, t_2)} - \mathrm{FDR}(t_1, t_2)\right| = 0$$

   a.s.

Proof:

1. Fix k to be either 1 or 2. Let $X_j = P_{kj}$ for $j = 1, \ldots, m$. Then $\hat{F}_k$ is the empirical distribution function of the sequence $\{X_j\}_{j=1,\ldots,m}$ whose the average distribution function is

$$\bar{F}_{kp} = m^{-1} \sum_{j=1}^{m} F_{kj}(t) = m^{-1} \left(\sum_{j:I_{kj}=0} F_k^0(t) + \sum_{j:I_{kj}=1} F_k^1(t) \right),$$

   which converges to $F_k(t)$ uniformly by Assumption 5.2.1. This implies that the corresponding sequence of probability measures $\{\mu_{\bar{F}_{kp}}\}_m$ converges weakly to μ_{F_k}, i.e.,

$$\lim_{m\to\infty} \rho(\mu_{\bar{F}_{kp}}, \mu_{F_k}) = 0, \tag{5.15}$$

as a result, $\{\mu_{\bar{F}_{kp}}\}_m$ is tight by Lemma 5.6.4, which in turn together with Lemma 5.6.3 implies that the sequence of empirical probability measures $\{\mu_{\hat{F}_k}\}_m$ is equivalent to $\{\mu_{\bar{F}_{kp}}\}_m$ in ρ-metric, more precisely,

$$\lim_{m\to\infty} \rho(\mu_{\hat{F}_k}, \mu_{\bar{F}_{kp}}) = 0 \quad a.s. \tag{5.16}$$

Combining (5.15) and (5.16), we obtain

$$\lim_{m\to\infty} \rho(\mu_{\hat{F}_k}, \mu_{F_k}) = 0 \quad a.s.,$$

which together with Lemma 5.6.2 (or Lemma 5.6.1) implies

$$\lim_{m\to\infty} \sup_{t\in[0,1]} |\hat{F}_k(t) - F_k(t)| = 0 \quad a.s.$$

2. This property is proven using the same arguments used in the proof of Lemma 5.3.1 item 1.

3. This property is proven using the same arguments used in the proof of Lemma 5.3.1 item 1.

4. Since $\hat{G}(t_1,t_2) \geq \hat{G}(\delta_1,\delta_2)$ for all $t_1 \geq \delta_1, t_2 \geq \delta_2$, $\lim_{m\to\infty} m^{-1} = 0$ and $\lim_{m\to\infty} \hat{G}(\delta_1,\delta_2) = G(\delta_1,\delta_2)$ a.s., therefore

$$\begin{aligned}
\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left| \frac{\hat{G}^0(t_1,t_2) - G^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} \right| &\leq \frac{1}{G(\delta_1,\delta_2)} \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left| \hat{G}^0(t_1,t_2) - G^0(t_1,t_2) \right| \\
&\leq \frac{1}{G(\delta_1,\delta_2)} \lim_{m\to\infty} \sup_{t_1,t_2\in[0,1]} \left| \hat{G}^0(t_1,t_2) - G^0(t_1,t_2) \right| \\
&= 0 \quad a.s.
\end{aligned}$$

5. To prove this, it is sufficient to show that

$$\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left| \frac{\hat{G}^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right| = 0 \quad a.s. \tag{5.17}$$

Because if (5.17) holds, then by the dominated convergence theorem,

$$\begin{aligned}
0 &= E\left(\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left| \frac{\hat{G}^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right| \right) \\
&= \lim_{m\to\infty} E\left(\sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left| \frac{\hat{G}^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right| \right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} E\left|\frac{\hat{G}^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right| \\
&\geq \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|E\left(\frac{\hat{G}^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right)\right| \\
&= \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\mathrm{FDR}(t_1,t_2) - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right|. \qquad (5.18)
\end{aligned}
$$

Combining (5.17) and (5.18) gives us

$$
\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{\hat{G}^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)} - \mathrm{FDR}(t_1,t_2)\right| = 0 \quad a.s.
$$

Now (5.17) is true since

$$
\begin{aligned}
&\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{\hat{G}^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right| \\
\leq\ &\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{\hat{G}^0(t_1,t_2)-G^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)}\right| + \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{G^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right| \\
\leq\ &0 + \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|G^0(t_1,t_2)\right| \left|\frac{1}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{1}{G(t_1,t_2)}\right| \\
\leq\ &\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{1}{m^{-1}\vee\hat{G}(t_1,t_2)} - \frac{1}{G(t_1,t_2)}\right| \\
\leq\ &\lim_{m\to\infty} \frac{1}{m^{-1}\vee G(\delta_1,\delta_2)} \frac{1}{G(\delta_1,\delta_2)} \lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|m^{-1}\vee\hat{G}(t_1,t_2) - G(t_1,t_2)\right| \\
\leq\ &\left(\frac{1}{G(\delta_1,\delta_2)}\right)^2 \lim_{m\to\infty} \sup_{t_1,t_2\in[0,1]} \left|m^{-1}\vee\hat{G}(t_1,t_2) - G(t_1,t_2)\right| = 0.
\end{aligned}
$$

This completes the proof.

□

The following lemma is a summary of several asymptotic results of the estimators $\hat{\pi}_{00}(\lambda_1,\lambda_2)$, $\hat{\pi}^0_k(\lambda_k)$, $\hat{\pi}_{10}(\lambda_1,\lambda_2)$, $\hat{\pi}_{01}(\lambda_1,\lambda_2)$ and $\hat{F}^1_k(t_k,\lambda_k)$.

Lemma 5.3.2 *Suppose Assumption 5.2.1 holds. Let* $g_k = \frac{1-F^1_k(\lambda_k)}{1-F^0_k(\lambda_k)}$ *for* $k = 1, 2$. *Then, for any* $0 \leq \lambda_1, \lambda_2 < 1$, *the followings hold almost surely as* $m \to \infty$

1. $\hat{\pi}_{00}(\lambda_1,\lambda_2) = \pi_{00} + \pi_{10}g_1 + \pi_{01}g_2 + \pi_{11}g_1g_2 + o(1)$
2. $\hat{\pi}^0_k(\lambda_k) = \pi^0_k + \pi^1_k g_k + o(1)$

3. $\hat{\pi}_{01}(\lambda_1, \lambda_2) = (\pi_{01} + \pi_{11} g_1)(1 - g_2) + o(1)$

4. $\hat{\pi}_{10}(\lambda_1, \lambda_2) = (\pi_{10} + \pi_{11} g_2)(1 - g_1) + o(1)$

5. $\hat{F}_k^1(t_k, \lambda_k) = \frac{F_k^1(t_k) - g_k F_k^0(t_k)}{1 - g_k} + o(1)$,

where $x_p = o(y_p)$ *means that* $\lim_{m \to \infty} x_p / y_p = 0$.

Proof:

1. Using Lemma 5.3.1 items 1 and 2 above, i.e.,

$$\lim_{p \to \infty} \sup_{t \in [0,1]} |\hat{F}_k(t) - F_k(t)| = 0 \quad a.s.$$

$$\lim_{p \to \infty} \sup_{t_1, t_2 \in [0,1]} |\hat{G}(t_1, t_2) - G(t_1, t_2)| = 0 \quad a.s.,$$

and by the definition of $\hat{\pi}_{00}(\lambda_1, \lambda_2)$ as in (5.5), with probability 1,

$$\begin{aligned}
\hat{\pi}_{00}(\lambda_1, \lambda_2) &= \frac{1 - \hat{F}_1(\lambda_1) - \hat{F}_2(\lambda_2) + \hat{G}(\lambda_1, \lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&= \frac{1 - F_1(\lambda_1) - F_2(\lambda_2) + G(\lambda_1, \lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&\quad - \frac{(\hat{F}_1(\lambda_1) - F_1(\lambda_1)) + (\hat{F}_2(\lambda_2) - F_2(\lambda_2)) - (\hat{G}(\lambda_1, \lambda_2) - G(\lambda_1, \lambda_2))}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&= \frac{1 - F_1(\lambda_1) - F_2(\lambda_2) + G(\lambda_1, \lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} + o(1) \\
&= \frac{\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} - (\pi_{00} + \pi_{01}) F_1^0(\lambda_1) - (\pi_{10} + \pi_{11}) F_1^1(\lambda_1)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&\quad - \frac{(\pi_{00} + \pi_{10}) F_2^0(\lambda_2) + (\pi_{01} + \pi_{11}) F_2^1(\lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&\quad + \frac{\pi_{00} F_1(\lambda_1)^0 F_2^0(\lambda_2) + \pi_{10} F_1^1(\lambda_1) F_2^0(\lambda_2) + \pi_{01} F_1^0(\lambda_1) F_2^1(\lambda_2) + \pi_{11} F_1^1(\lambda_1) F_2^1(\lambda_2)}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&\quad + o(1) \\
&= \pi_{00} \frac{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} + \pi_{10} \frac{(1 - F_1^1(\lambda_1))(1 - F_2^0(\lambda_2))}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} \\
&\quad + \pi_{01} \frac{(1 - F_1^0(\lambda_1))(1 - F_2^1(\lambda_2))}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} + \pi_{11} \frac{(1 - F_1^1(\lambda_1))(1 - F_2^1(\lambda_2))}{(1 - F_1^0(\lambda_1))(1 - F_2^0(\lambda_2))} + o(1) \\
&= \pi_{00} + \pi_{10} g_1 + \pi_{01} g_2 + \pi_{11} g_1 g_2 + o(1),
\end{aligned}$$

which completes the proof.

2. From the definition of $\hat{\pi}_k^0(\lambda_k)$ as in (5.4), we have

$$\begin{aligned}
\hat{\pi}_k^0(\lambda_k) &= \frac{1-\hat{F}_k(\lambda_k)}{1-F_k^0(\lambda_k)} \\
&= \frac{1-F_k(\lambda_k)}{1-F_k^0(\lambda_k)} - \frac{\hat{F}_k(\lambda_k)-F_k(\lambda_k)}{1-F_k^0(\lambda_k)} \\
&= \frac{1-F_k(\lambda_k)}{1-F_k^0(\lambda_k)} + o(1) \\
&= \frac{\pi_k^0+\pi_k^1-\pi_k^0F_k^0(\lambda_k)-\pi_k^1F_k^1(\lambda_k)}{1-F_k^0(\lambda_k)} + o(1) \\
&= \frac{\pi_k^0(1-F_k^0(\lambda_k))+\pi_k^1(1-F_k^1(\lambda_k))}{1-F_k^0(\lambda_k)} + o(1) \\
&= \pi_k^0+\pi_k^1 g_k + o(1)
\end{aligned}$$

as $m \to \infty$ a.s.

3. From two equalities above, we have

$$\begin{aligned}
\hat{\pi}_{01}(\lambda_1,\lambda_2) &= \hat{\pi}_1^0(\lambda_1) - \hat{\pi}_{00}(\lambda_1,\lambda_2) \\
&= \pi_1^0+\pi_1^1 g_1 - (\pi_{00}+\pi_{10}g_1+\pi_{01}g_2+\pi_{11}g_1g_2) + o(1) \\
&= \pi_1^0+(\pi_{10}+\pi_{11})g_1 - (\pi_{00}+\pi_{10}g_1+\pi_{01}g_2+\pi_{11}g_1g_2) + o(1) \\
&= \pi_1^0-\pi_{00}+\pi_{11}g_1(1-g_2)-\pi_{01}g_2 + o(1) \\
&= (\pi_{01}+\pi_{11}g_1)(1-g_2) + o(1)
\end{aligned}$$

as $m \to \infty$ a.s.

4. Similarly, we have

$$\begin{aligned}
\hat{\pi}_{10}(\lambda_1,\lambda_2) &= \hat{\pi}_2^0(\lambda_2) - \hat{\pi}_{00}(\lambda_1,\lambda_2) \\
&= \pi_2^0+\pi_2^1 g_2 - (\pi_{00}+\pi_{10}g_1+\pi_{01}g_2+\pi_{11}g_1g_2) + o(1) \\
&= \pi_2^0-\pi_{00}+(\pi_{01}+\pi_{11})g_2 - (\pi_{00}+\pi_{10}g_1+\pi_{01}g_2+\pi_{11}g_1g_2) + o(1) \\
&= (\pi_{10}+\pi_{11}g_2)(1-g_1) + o(1)
\end{aligned}$$

as $m \to \infty$ a.s.

5. Using (5.8) and item 2 in this lemma, we have

$$\begin{aligned}
\hat{F}_k^1(t_k,\lambda_k) &= \frac{\hat{F}_k(t_k)-F_k^0(t_k)\hat{\pi}_k^0(\lambda_k)}{1-\hat{\pi}_k^0(\lambda_0)} \\
&= \frac{\hat{F}_k(t_k)-F_k^0(t_k)(\pi_k^0+\pi_k^1 g_k+o(1))}{1-(\pi_k^0+\pi_k^1 g_k+o(1))} \\
&= \frac{F_k(t_k)-F_k^0(t_k)(\pi_k^0+\pi_k^1 g_k+o(1))}{1-(\pi_k^0+\pi_k^1 g_k+o(1))}+\frac{\hat{F}_k(t_k)-F_k(t_k)}{1-(\pi_k^0+\pi_k^1 g_k+o(1))} \\
&= \frac{F_k(t_k)-F_k^0(t_k)(\pi_k^0+\pi_k^1 g_k)}{1-(\pi_k^0+\pi_k^1 g_k)}+o(1) \\
&= \frac{(F_k(t_k)-\pi_k^0 F_k^0(t_k))-\pi_k^1 F_k^0(t_k)g_k}{(1-\pi_k^0)-\pi_k^1 g_k}+o(1) \\
&= \frac{\pi_k^1 F_k^1(t_k)-\pi_k^1 F_k^0(t_k)g_k}{\pi_k^1-\pi_k^1 g_k}+o(1) \\
&= \frac{F_k^1(t_k)-F_k^0(t_k)g_k}{1-g_k}+o(1)
\end{aligned}$$

as $m\to\infty$ a.s.

□

Lemma 5.3.3 *Suppose Assumption 5.2.1 holds, then*

$$\lim_{m\to\infty}\inf_{t_1\geq\delta_1,t_2\geq\delta_2}\left(\hat{\pi}_{00}F_1^0(t_1)F_2^0(t_2)+\hat{\pi}_{10}\hat{F}_1^1(t_1,\lambda_1)F_2^0(t_2)+\hat{\pi}_{01}F_1^0(t_1)\hat{F}_2^1(t_2,\lambda_2)-G^0(t_1,t_2)\right)\geq 0$$

as $p\to\infty$ a.s. As a consequence,

$$\lim_{m\to\infty}\inf_{t_1\geq\delta_1,t_2\geq\delta_2}\left(\widehat{\mathrm{FDR}}^{\lambda_1,\lambda_2}(t_1,t_2)-\frac{G^0(t_1,t_2)}{m^{-1}\vee\hat{G}(t_1,t_2)}\right)\geq 0.$$

Proof: For simplicity and without loss of generality, from now, we use F_k^a and $\hat{F}_k^a$ instead of $F_k^a(t_k)$ and $\hat{F}_k(t_k,\lambda_k)$ for $a=0,1$; $k=1,2$. From Lemma 5.3.2 (items 4 & 5), we have

$$\begin{aligned}
\hat{\pi}_{10}\hat{F}_1^1F_2^0 &= (\pi_{10}+\pi_{11}g_2)(1-g_1)\left(\frac{F_1^1-g_1F_1^0}{1-g_1}+o(1)\right)F_2^0 \\
&= (\pi_{10}+\pi_{11}g_2)(F_1^1-g_1F_1^0)F_2^0+o(1) \qquad (5.19)\\
&= \pi_{10}F_1^1F_2^0-\pi_{10}g_1F_1^0F_2^0+\pi_{11}g_2(F_1^1-g_1F_1^0)F_2^0+o(1). \qquad (5.20)
\end{aligned}$$

Similarly,

$$\hat{\pi}_{01}F_1^0\hat{F}_2^1=\pi_{01}F_1^0F_2^1-\pi_{01}g_2F_1^0F_2^0+\pi_{11}g_1F_1^0(F_2^1-g_2F_2^0)+o(1). \qquad (5.21)$$

Also, from Lemma 5.3.2 item 1, we have

$$\hat{\pi}_{00}F_1^0F_2^0 = \pi_{00}F_1^0F_2^0 + \pi_{10}g_1F_1^0F_2^0 + \pi_{01}g_2F_1^0F_2^0 + \pi_{11}g_1g_2F_1^0F_2^0 + o(1) \tag{5.22}$$

Combining (5.19), (5.21), and (5.22) gives us

$$\begin{aligned}
\hat{\pi}_{00}F_1^0F_2^0 + \hat{\pi}_{10}\hat{F}_1^1F_2^0 + \hat{\pi}_{01}F_1^0\hat{F}_2^1 &= \pi_{00}F_1^0F_2^0 + \pi_{10}g_1F_1^0F_2^0 + \pi_{01}g_2F_1^0F_2^0 + \pi_{11}g_1g_2F_1^0F_2^0 \\
&\quad + \pi_{10}F_1^1F_2^0 - \pi_{10}g_1F_1^0F_2^0 + \pi_{11}g_2(F_1^1 - g_1F_1^0)F_2^0 \\
&\quad + \pi_{01}F_1^0F_2^1 - \pi_{01}g_2F_1^0F_2^0 + \pi_{11}g_1F_1^0(F_2^1 - g_2F_2^0) \\
&= \pi_{00}F_1^0F_2^0 + \pi_{10}F_1^1F_2^0 + \pi_{01}F_1^0F_2^1 \\
&\quad + \pi_{11}g_2(F_1^1 - g_1F_1^0)F_2^0 + \pi_{11}g_1F_1^0F_2^1 + o(1) \\
&= G^0(t_1,t_2) + \pi_{11}g_2(F_1^1 - g_1F_1^0)F_2^0 \\
&\quad + \pi_{11}g_1F_1^0F_2^1 + o(1)
\end{aligned} \tag{5.23}$$

which together with the fact that $\pi_{11}g_2(F_1^1 - g_1F_1^0)F_2^0 + \pi_{11}g_1F_1^0F_2^1 \geq 0$ (since $F_1^1 \geq F_1^0, 0 \leq g_1 \leq 1$) completes the proof. □

Now we are in a position to prove Theorem 5.2.1.

Proof of Theorem 5.2.1: First, we have

$$\begin{aligned}
\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDP}(t_1,t_2) &= \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \frac{G^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)}\right) \\
&\quad + \left(\frac{G^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} - \mathrm{FDP}(t_1,t_2)\right).
\end{aligned}$$

This together with Lemma 5.3.3

$$\lim_{p\to\infty} \inf_{t_1\geq\delta_1, t_2\geq\delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \frac{G^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)}\right) \geq 0$$

and Lemma 5.3.1 item 4

$$\lim_{p\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} \left|\frac{G^0(t_1,t_2)}{m^{-1} \vee \hat{G}(t_1,t_2)} - \mathrm{FDP}(t_1,t_2)\right| = 0$$

implies that

$$\lim_{p\to\infty} \inf_{t_1\geq\delta_1, t_2\geq\delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDP}(t_1,t_2)\right) \geq 0. \tag{5.24}$$

Combining (5.24) and Lemma 5.3.1 item 5

$$\lim_{m\to\infty} \sup_{t_1\geq\delta_1, t_2\geq\delta_2} |\mathrm{FDP}(t_1,t_2) - \mathrm{FDR}(t_1,t_2)| = 0$$

together with

$$\begin{aligned}\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDR}(t_1,t_2) &= \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDP}(t_1,t_2)\right) \\ &\quad + (\mathrm{FDP}(t_1,t_2) - \mathrm{FDR}(t_1,t_2))\end{aligned}$$

implies

$$\lim_{m\to\infty} \inf_{t_1\geq\delta_1, t_2\geq\delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDR}(t_1,t_2)\right) = 0,$$

which completes the proof of Theorem 5.2.1. □

Using Theorem 5.2.1, we can now prove Theorem 5.2.2.

Proof of Theorem 5.2.2: From the assumption of Theorem 5.2.2 and Theorem 5.2.1, we have

$$\begin{aligned}&\lim \inf_{m\to\infty} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_{1j^*},t_{2j^*}) - \mathrm{FDP}(t_{1j^*},t_{2j^*})\right) \\ \geq\ &\lim_{m\to\infty} \inf_{t_1\geq\delta_1, t_2\geq\delta_2} \left(\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_1,t_2) - \mathrm{FDP}(t_1,t_2)\right) \\ \geq\ &0 \qquad a.s.\end{aligned}$$

This together with the definition of j^* as in (5.14)

$$\widehat{\mathrm{FDR}}^{(\lambda_1,\lambda_2)}(t_{1j^*},t_{2j^*}) \leq \alpha$$

implies that

$$\lim \sup_{m\to\infty} \mathrm{FDP}(t_{1j^*},t_{2j^*}) \leq \alpha,$$

which in turn together with the Fatou lemma gives us

$$\lim \sup_{m\to\infty} \mathrm{FDR}(t_{1j^*},t_{2j^*}) = \lim \sup_{m\to\infty} E(\mathrm{FDP}(t_{1j^*},t_{2j^*})) \leq E\left(\lim \sup_{m\to\infty} \mathrm{FDP}(t_{1j^*},t_{2j^*})\right) \leq \alpha.$$

□

5.3.1 Comparison with the FDR Estimation Method of ZN

ZN proposed two FDR estimation procedures when identifying simultaneous signals. One procedure is parametric when the null distribution of test statistic is known. The other procedure is non-parametric when the null distribution of test statistic is unknown. Both parametric and non-parametric versions of ZN are based on the following inequality.

Proposition 5.3.1 (ZN) *Under model* (5.1) *and Assumption 5.2.1,*

$$G^0(t_1,t_2) \leq F_1(t_1)F_2^0(t_2) + F_1^0(t_1)F_2(t_2) - F_1^0(t_1)F_2^0(t_2).$$

If the null distribution of the test statistic is known, they proposed the following parametric FDR estimation procedure

$$\widehat{\text{FDR}}^U(t_1,t_2) = \begin{cases} \max\left\{ \dfrac{\hat{F}_1(t_1)F_2^0(t_2) + F_1^0(t_1)\hat{F}_2(t_2) - F_1^0(t_1)F_2^0(t_2)}{\hat{G}(t_1,t_2)}, 0 \right\}, & \hat{G}(t_1,t_2) > 0, \\ 0, & \hat{G}(t_1,t_2) = 0. \end{cases} \tag{5.25}$$

To identify simultaneous signals, first consider the set of optimal rejection regions

$$\mathcal{T} = \underset{(t_1,t_2\in\Pi)}{\arg\max} \left\{ \hat{G}(t_1,t_2) : \widehat{\text{FDR}}^U(t_1,t_2) \leq \alpha \right\},$$

where

$$\Pi = \{(0,0)\} \cup \left\{ (T_{1j}, T_{2j'}) : 1 \leq j, j' \leq m \right\}.$$

Then the optimal thresholds will be taken as the region in $\mathcal{T}$ that maximizes the attained FDR

$$(\hat{t}_1, \hat{t}_2) = \underset{(t_1,t_2)\in\mathcal{T}}{\arg\max} \widehat{\text{FDR}}^U(t_1,t_2).$$

On the other hand, in the case that the null distribution of test statistic is unknown, ZN proposed a non-parametric FDR estimation procedure

$$\widehat{\text{FDR}}^U_{NP}(t_1,t_2) = \begin{cases} \dfrac{\hat{F}_1(t_1)\hat{F}_2(t_2)}{\hat{G}(t_1,t_2)}, & \hat{G}(t_1,t_2) > 0, \\ 0, & \hat{G}(t_1,t_2) = 0. \end{cases} \tag{5.26}$$

As before, the set of rejection regions is defined as

$$\mathcal{T}_{NP} = \underset{(t_1,t_2)\in\Pi}{\arg\max} \left\{ \hat{G}(t_1,t_2) : \widehat{\mathrm{FDR}}^{U}_{NP}(t_1,t_2) \leq \alpha \right\}.$$

Then the optimal thresholds will be selected such that its components have the smallest product

$$(\hat{t}_1,\hat{t}_2) = \underset{(t_1,t_2)\in\mathcal{T}_{NP}}{\arg\max}\ t_1 t_2.$$

In both FDR estimation procedures, a feature j is declared as a simultaneous signal if both its p-values are less than or equal to the respective optimal thresholds, i.e., $P_{1j} \leq \hat{t}_1, P_{2j} \leq \hat{t}_2$.

ZN proved that, asymptotically, both versions of their method conservatively estimate FDR if the simultaneous signals are sparse and the test statistics from two experiments are positively dependent. In this section, we show that asymptotically our FDR estimation procedure is less conservative than method of ZN.

Indeed, from (5.23) in the proof of Lemma 5.3.3, the asymptotic difference between the numerator of our FDR estimator and that of the theoretical FDR is

$$d_1 = \pi_{11}\left(g_1 F_1^0 F_2^1 + g_2 F_2^0 (F_1^1 - g_1 F_1^0)\right). \tag{5.27}$$

By a simple but long calculation, we can show that the asymptotic difference between the numerator of the non-parametric FDR estimator proposed by ZN and that ofthe theoretical FDR is

$$d_2 = F_1(t_1)F_2(t_2) - G^0(t_1,t_2) = \pi_{11}(F_1^0 F_2^1 + F_2^0(F_1^1 - F_1^0)) + \pi_1^1\pi_2^1(F_1^1 - F_1^0)(F_2^1 - F_2^0). \tag{5.28}$$

From (5.27) and (5.28), we have

$$\begin{aligned}
d_1 - d_2 &= \pi_{11}\{g_1 F_1^0 F_2^1 + g_2 F_2^0(F_1^1 - g_1 F_1^0) - \left(F_1^0 F_2^1 + F_2^0(F_1^1 - F_1^0)\right)\} \\
&= -\pi_{11}\{(1-g_1)F_1^0 F_2^1 + F_2^0(F_1^1 - F_1^0) - g_2 F_2^0(F_1^1 - F_1^0 + (1-g_1)F_1^0)\} \\
&= -\pi_{11}\{(1-g_1)F_1^0 F_2^1 - g_2 F_2^0(1-g_1)F_1^0 + F_2^0(F_1^1 - F_1^0) - g_2 F_2^0(F_1^1 - F_1^0)\} \\
&= -\pi_{11}\{(1-g_1)F_1^0(F_2^1 - g_2 F_2^0) + (1-g_2)F_2^0(F_1^1 - F_1^0)\} \geq 0,
\end{aligned}$$

which means that our FDR estimator is less conservative than the non-parametric FDR method of ZN. The difference between the two methods is more significant if the number of simultaneous signals is large.

On the other hand, the asymptotic difference between the numerator of the parametric FDR estimator proposed by ZN and that of the theoretical FDR is

$$\begin{aligned} d_3 = F_1F_2^0 + F_1^0F_2 - F_1^0F_2^0 - G^0(t_1,t_2) &= F_1F_2 - \pi_1\pi_2(F_1^1 - F_1^0)(F_2^1 - F_2^0) \\ &\quad - (F_1^1F_2^0 + F_2^1F_1^0 + F_1^0F_2^0) \\ &= \pi_{11}(F_1^0F_2^1 + F_1^1F_2^0 - F_1^0F_2^0). \end{aligned} \tag{5.29}$$

From (5.27) and (5.29), we have

$$\begin{aligned} d_1 - d_3 =& \pi_{11}\left\{g_1F_1^0F_2^1 + g_2F_2^0(F_1^1 - g_1F_1^0) - (F_1^0F_2^1 + F_1^1F_2^0 - F_1^0F_2^0)\right\} \\ &= \pi_{11}\left\{g_1F_1^0(F_2^1 - g_2F_2^0) - F_1^0(F_2^1 - F_2^0) - F_1^1F_2^0(1-g_2)\right\} \\ &= \pi_{11}\left\{g_1F_1^0\left(F_2^1 - \frac{1-F_2^1}{1-F_2^0}F_2^0\right) - F_1^0\left(F_2^1 - F_2^0\right) - F_1^1F_2^0\left(1 - \frac{1-F_2^1}{1-F_2^0}\right)\right\} \\ &= \pi_{11}\left\{g_1F_1^0\frac{F_2^1 - F_2^0}{1-F_2^0} - F_1^0\left(F_2^1 - F_2^0\right) - F_1^1F_2^0\frac{F_2^1 - F_2^0}{1-F_2^0}\right\} \\ &= \pi_{11}\left\{\frac{F_2^1 - F_2^0}{1-F_2^0}\left[g_1F_1^0 - F_1^0(1-F_2^0) - F_1^1F_2^0\right]\right\} \\ &= \pi_{11}\left\{\frac{F_2^1 - F_2^0}{1-F_2^0}\left[g_1F_1^0 - F_1^0 + F_1^0F_2^0 - F_1^1F_2^0\right]\right\} \\ &= -\pi_{11}\left\{\frac{F_2^1 - F_2^0}{1-F_2^0}\left[F_1^0(1-g_1) + F_2^0(F_1^1 - F_1^0)\right]\right\} < 0. \end{aligned}$$

Therefore, asymptotically, our FDR estimation procedure is less bias than both versions of FDR estimation procedure of ZN.

5.4 Simulation Study

5.4.1 Simulation Setting

In this section, we present a simulation study to investigate performance of our method compared to the method of ZN in terms of FDR control and power of detecting true simultaneous signals. In all simulation settings, we consider $m = 10000$ features. The indicators I_{kj} of true

status of the features were generated and then fixed across 100 replications, with the sparsity levels varying across the simulation settings. For a feature j in an experiment k, a test statistic T_{kj} was generated from a t_4 distribution with non-centrality parameter $\mu_{kj} = 0$ if $I_{kj} = 0$, otherwise, μ_{kj} was drawn from $N(6, 1)$ if $I_{kj} = 1$ then fixed across 100 replications. Each simulation setting corresponds to a set of values of $m_{01}, m_{10}, m_{11}, m_{00}$. In particular, we consider the following simulation settings.

- $m_{01} = m_{10} = 25; m_{11} = 50$: This is the case of sparse simultaneous signals, and the positive dependence condition $\pi_{11} > \pi_1\pi_2$ in ZN is satisfied because $\pi_{11} = 50/10^4 = 0.005 > 5.625e-05 = ((25 + 50)/10^4)^2$.

- $m_{01} = m_{10} = 1000; m_{11} = 50$: This is the case sparse simultaneous signals, but $\pi_{11} = 0.005 < 0.011025 = ((1000 + 50)/10^4)^2 = \pi_1\pi_2$, therefore, the positive dependence condition in ZN is violated.

- $m_{01} = m_{10} = 25; m_{11} = 1000$: This is the case there are a moderate number of simultaneous signals and the positive dependence condition in ZN is satisfied, because $\pi_{11} = 1000/10^4 = 0.1 > 0.01050625 = ((1000 + 25)/10^4)^2 = \pi_1\pi_2$.

- $m_{01} = m_{10} = 1000; m_{11} = 1000$: This is the case there are a moderate number of simultaneous signals and the positive dependence condition in ZN is satisfied, since $\pi_{11} = 1000/10^4 = 0.1 > 0.04 = ((1000 + 1000)/10^4)^2 = \pi_1\pi_2$.

5.4.2 Simulation Results

The simulation results in terms of FDR control (the nominal FDR level is 5%) and the power (quantified by the number of true discoveries) of each method are summarized in Figure 5.1.

Figure 5.1 shows that our method controls FDR well and is more powerful than both versions of the method of ZN. The parametric version of the method of ZN is failed to control FDR when the positive dependence condition is violated.

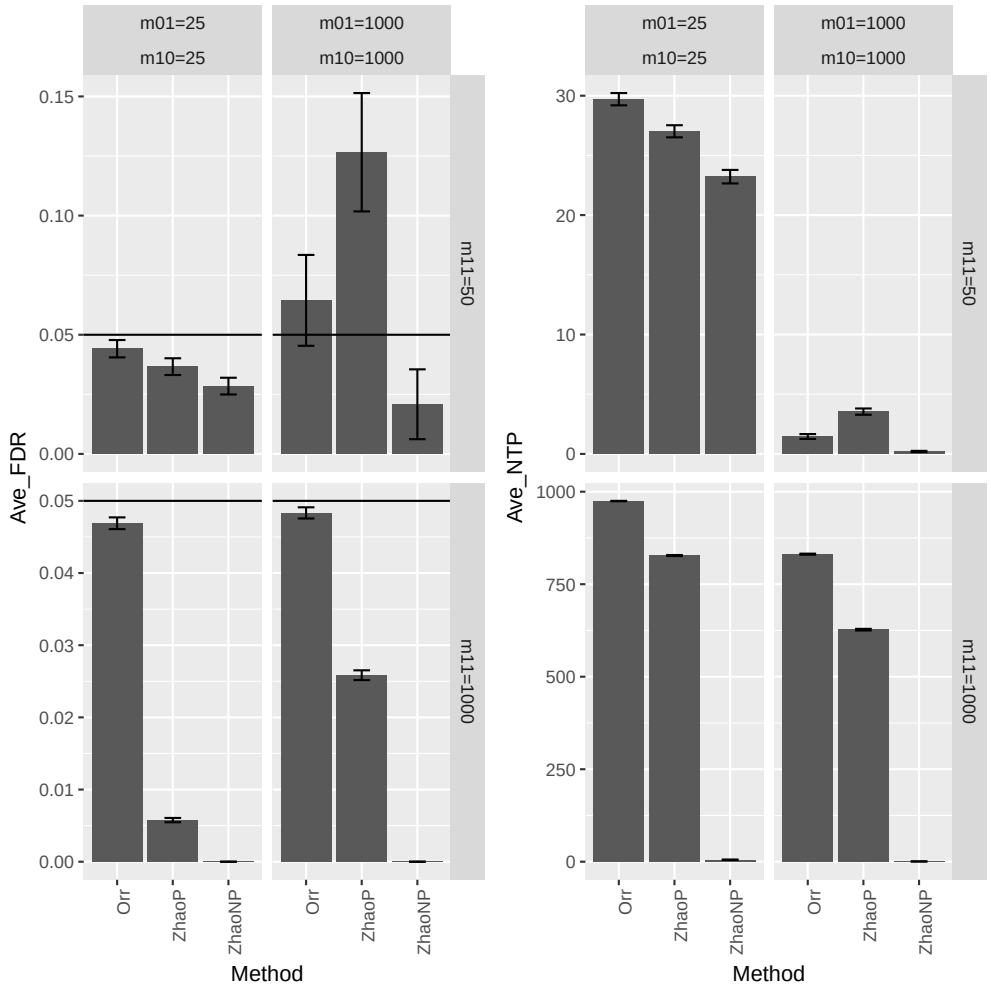

Figure 5.1: The incurred false discovery rate (FDR) and the number of true positive detections NTP averaging over 100 simulations. Orr: our method; ZhaoP and ZhaoNP are parametric and non-parametric versions of the method of ZN, respectively.

5.5 Discussion

5.5.1 About Assumption 5.2.1

To prove Theorem 5.2.1, we need results of Lemma 5.3.1 items 4, 5 and Lemma 5.3.3. These results only require the results of Lemma 5.3.1 .1, .2, .3. In facts, weaker (but equivalent) conditions similar to Lemma 5.3.1 items 1, 2, 3 needed, in particular, we only need the following conditions: There exist continuous functions $G^0_{t_1,t_2}, G^1(t_1,t_2), F^0_k(t_k), F^1_k(t_k)$ such that

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{m_{00}+m_{01}} \sum_{j\in\mathcal{A}_{00}\cup\mathcal{A}_{01}} \mathbb{1}(P_{1j}\le t_1) = F^0_1(t_1)\\
&\lim_{m\to\infty} \frac{1}{m_{10}+m_{11}} \sum_{j\in\mathcal{A}_{10}\cup\mathcal{A}_{11}} \mathbb{1}(P_{1j}\le t_1) = F^1_1(t_1)\\
&\lim_{m\to\infty} \frac{1}{m_{00}+m_{10}} \sum_{j\in\mathcal{A}_{00}\cup\mathcal{A}_{10}} \mathbb{1}(P_{2j}\le t_2) = F^0_2(t_2)\\
&\lim_{m\to\infty} \frac{1}{m_{01}+m_{11}} \sum_{j\in\mathcal{A}_{01}\cup\mathcal{A}_{11}} \mathbb{1}(P_{2j}\le t_2) = F^1_2(t_2)\\
&\lim_{m\to\infty} \hat{G}^0(t_1,t_2) = G^0(t_1,t_2)
\end{aligned}$$

a.s for each $t_1, t_2 \in [0,1]$. It can be shown that a sequence of random variables possessing ergodic property satisfies these conditions, therefore, Theorem 5.2.1 holds even in the case that p-values are dependent such as p-values possessing ergodic property (Rao, 1962, Theorem 4.2 and Theorem 6.2).

5.5.2 Extension to more than Two Independent Experiments

In this paper, we proposed a new method to estimate FDR when identifying simultaneous signals in two independent experiments. More work is needed to extend our method to more than two experiments.

5.6 Appendix: Useful Results in Probability Theory and Measure Theory

This section contains the necessary background in probability theory and measure theory that we use to prove the asymptotic results of our method.

Lemma 5.6.1 (Lemma 3.2, Rao (1962)) *Let $F(x_1, x_2, \ldots, x_k)$ be a distribution function on the Euclidean space $\mathbb{R}^k$ such that each marginal distribution function is continuous. Then a sequence of distribution functions $F_n(x_1, \ldots, x_k)$ converges weakly to $F(x_1, \ldots, x_k)$ if and only if*

$$\lim_{n\to\infty} \sup_{\boldsymbol{x}} |F_n(x_1, \ldots, x_k) - F(x_1, \ldots, x_k)| = 0, \tag{5.30}$$

where the supremum is taken over all the vectors $\boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$.

Lemma 5.6.1 is a multivariate version of Polya theorem about the relationship between uniform convergence and weak convergence of a sequence of distribution functions on real line $\mathbb{R}$. An extension of Lemma 5.6.1 is the following lemma.

Lemma 5.6.2 (Theorem 4.2, Rao (1962)) *Suppose μ is a measure on $\mathbb{R}^k$ such that every convex subset of $\mathbb{R}^k$ has μ-null boundary (i.e., each convex set is a continuity set for μ). Then μ_n converges weakly to μ ($\mu_n \Rightarrow \mu$) if and only if*

$$\sup\{|\mu_n(C) - \mu(C)| : C \in \mathcal{C}\} \to 0, \tag{5.31}$$

where $\mathcal{C}$ denotes the class of all measurable convex sets. In particular, (5.31) is valid if the measure μ is absolutely continuous with respect to Lebesgue measure.

The next result is a Glivenko-Cantelli theorem for empirical measures of independent but non-identically distributed random variables.

Lemma 5.6.3 (Theorem 1, Wellner (1981)) *Let $(\mathbf{S}, d)$ be a separable metric space. Let $\mathscr{P}(\mathbf{S})$ be the set of all Borel probability measures on $\mathbf{S}$. Let $X_1, \ldots, X_n$ be independent $\mathbf{S}$-valued random variables with distributions $P_1, \ldots, P_n$, where all $P_i \in \mathscr{P}(\mathbf{S})$. For $n \geq 1$, define the empirical measure $\mathbb{P}_n$ by*

$$\mathbb{P}_n \equiv (\delta_{X_1} + \cdots + \delta_{X_n})/n,$$

where $\delta_{X_i}(x) = 1$ if $X_i = x$ and 0 otherwise. Define the average measure $\bar{P}_n$ by

$$\bar{P}_n = (P_1 + \cdots + P_n)/n.$$

Let ρ and β denote the Prohorov and dual-bounded-Lipschitz metrics on $\mathscr{P}(\mathbf{S})$, respectively; i.e, for $P, Q \in \mathscr{P}(\mathbf{S})$,

$$\rho(P,Q) = \inf\{\varepsilon > 0 : P(A) \leq \varepsilon + Q(A^{\varepsilon}) \text{ for all Borel set } A\}$$

where

$$A^{\varepsilon} = \{y \in \mathbf{S} : d(x,y) < \varepsilon \text{ for some } x \in A\},$$

and

$$\beta(P,Q) = ||P-Q||^{*}_{BL} \equiv \sup\left\{\left|\int_x f d(Q-P)\right| : ||f||_{BL} \leq 1\right\}$$

with $||f||_{\infty} \equiv \sup_x |f(x)|, ||f||_L = \sup_{x \neq y} |f(x) - f(y)|/d(x,y)$ and $||f||_{BL} \equiv ||f||_{\infty} + ||f||_L$. If $\{\bar{P}_n\}_{n \geq 1}$ is tight,i.e. , for every $\varepsilon > 0$, there exists a compact set $K \subset \mathbf{S}$ such that $\bar{P}_n(K) > 1 - \varepsilon$ for all $n \geq 1$, then $\rho(\mathbf{P}_n, \bar{P}_n) \to 0$ a.s., and $\beta(\mathbb{P}_n, \bar{P}_n) \to 0$ a.s. as $n \to \infty$.

A sufficient condition for the tightness of a sequence of probability measure is given by the following lemma.

Lemma 5.6.4 (Proposition 9.3.4, Dudley (2002)) *Every weakly converging sequence on $\mathbb{R}^k$ is tight.*

Acknowledgments

This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) and the joint National Science Foundation (NSF)/NIGMS Mathematical Biology Program under award number R01GM109458. The opinions, findings, and conclusions stated herein are those of the authors and do not necessarily reflect those of USDA, NSF, or NIH.

Bibliography

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., et al. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics*, 9(4):e1003455.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, 108(504):1480–1492.

Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics*, 10(11):e1004787.

Cross-Disorder Group of the Psychiatric Genomics Consortium et al. (2013a). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9):984–994.

Cross-Disorder Group of the Psychiatric Genomics Consortium et al. (2013b). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1371.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

Grüneberg, H. (1938). An analysis of the "pleiotropic" effects of a new lethal mutation in the rat (mus norvegicus). *Proceedings of the Royal Society of London B: Biological Sciences*, 125(838):123–144.

Heller, R., Bogomolov, M., and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111(46):16262–16267.

Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8(1):481–498.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182.

Nettleton, D., Hwang, J. T. G., Caldo, R. A., and Wise, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337.

Orr, M. (2012). *Assessing differential expression when the distribution of effect sizes is asymmetric and evaluating concordance of differential expression across multiple gene expression experiments*. PhD thesis, Graduate Theses and Dissertations, Iowa State University, Ames, IA, 50011.

Phillips, D. and Ghosh, D. (2014). Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics. *The Annals of Applied Statistics*, 8(2):801–823.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 33(2):659–680.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, pages 607–618.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):187–205.

Wellner, J. A. (1981). A Glivenko-Cantelli theorem for empirical measures of independent but non-identically distributed random variables. *Stochastic Processes and their Applications*, 11(3):309–312.

Zhao, S., D. and Nguyen, Y., T. (2017). False discovery rate control for identifying simultaneous signals. *Under Review*.

CHAPTER 6. GENERAL CONCLUSION

This dissertation is composed of four papers addressing three statistical challenges in RNA-seq analysis and multiple hypothesis testing: 1) selecting relevant covariates in RNA-seq analysis, 2) modeling within-gene dependence in RNA-seq analysis from repeated-measures designs, and 3) estimating and controlling FDR when identifying simultaneous signals in two independent experiments. In this chapter, we summarize our findings and suggest directions for future work.

6.1 Summary

Chapter 2 presents a backward selection strategy to choose the most relevant covariates when conducting RNA-seq analysis with many available covariates. The QuasiSeq method is used to analyze these data. Using the vector of p-values for each covariate, we propose two simple covariate relevance measures: 1) the number of p-values less than 0.05, and 2) the Kolmogorov-Smirnov statistic measuring the discrepancy between the uniform(0, 1) distribution and the Grenander estimate of a non-increasing distribution computed from the elements in the vector of p-values. These two measures perform similarly in identifying the relevant covariates. As a result of accounting for relevant covariates, simulation study shows that our method outperforms methods that do not take covariate selection into account.

The method in Chapter 2 performs well except when one or more covariates are strongly correlated with the main factor of interest. We develop another covariate selection strategy in Chapter 3 that overcomes this limitation. Our method in Chapter 3 is an extension of the variable selection method hat is intended to control false selection rate (FSR) by the introduction and evaluation of pseudo-covariates that, by design, are uncorrelated with response values. We propose a simple and intuitive covariate relevance measure, which may be informally described as the ratio of the number of small p-values to the number of large p-values. Simulation study shows that our new

method performs similar to the method in Chapter 2 when the covariates are weakly correlated or uncorrelated with the main factor of interest. When there are one or more covariates strongly correlated with the main factor of interest, the new method outperforms the method in Chapter 2.

Chapter 4 introduces a statistical method to analyze RNA-seq data from repeated-measures designs. In repeated-measures experiments, observations taken at different time points from the same subject tends to be correlated. Existing RNA-seq analysis methods do not consider this correlation. Our method is based on normalized log-counts and associated precision weights in a general linear model pipeline with continuous auto-regressive structure to account for correlation among observations within each subject. We then utilize a parametric bootstrap procedure to conduct differential expression inference. Simulation study shows the advantages of our method over alternatives that do not account for correlation among observations within subjects.

Chapter 5 provides an FDR estimation and control procedure when identifying simultaneous signal in two independent experiments. Different from one experiment, a false discovery in two independent experiments occurs when a feature is declared to be a discovery in both experiments, when in reality, that feature is null in one or both experiments. FDR estimation in two independent experiments is therefore more challenging because the null for the simultaneous signal test is a composite null. We address this challenge by extending the histogram-based FDR estimation procedure for one experiment. We also propose an FDR control procedure similar to the procedure based on q-values. The desired theoretical properties of our FDR estimation and control procedure are provided. Additional simulation study also shows that our method outperforms existing methods.

6.2 Future Work

The work presented in this dissertation suggests a few directions for future research. Chapter 2 and Chapter 3 deal with variable selection when all covariates are available. An initial analysis in Chapter 2 shows that either adjusting or not adjusting for hidden covariates has no effects on our backward selection procedure. Chapter 3 further presents a simulation study when the method ignoring all available covariates and using only hidden covariates performs poorly in differential

expression analysis if some of the available covariates are strongly correlated with the main effect of interest. Therefore, a study on a unified approach to account for both relevant measured covariates and unknown artifacts could be desirable. Such a study could give a comprehensive solution to RNA-seq analysis under the effect of both known and unknown covariates.

Chapter 4 presents a general solution to differential expression analysis using RNA-seq from repeated-measures designs. Our approach could also be extended for other designs, such as split-splot designs or for other models that include random effects. Even remaining within the context of repeated-measures designs, there is still an open question on how to determine which correlation structure would be the most appropriate for a particular dataset. Furthermore, how to improve estimation of correlation parameters is also an interesting question. Additionally, an RNA-seq experiment from a repeated-measures design may also include other covariates. Select the most relevant covariates in such cases can be challenging. Additional research on complex RNA-seq designs that include both covariates and dependence among observations would be of value in practice.

Chapter 5 proposes a histogram-based FDR estimation and control procedure when identifying simultaneous signals in two independent experiments. A natural question for future work is how to extend this approach to three or more independent experiments. More work is needed to obtain satisfactory results in such extensions.