REVSTAT – Statistical Journal
Volume 18, Number 2, April 2020, 153–163

SIMULTANEOUS INFERENCE OF GENE ISOFORM EXPRESSION FOR RNA SEQUENCING DATA

Author: Bo Li
– Department of Mathematical Sciences,
The Citadel, The Military College of South Carolina,
Charleston, SC 29409, South Carolina, USA
bli@citadel.edu

Received: January 2019 Revised: September 2019 Accepted: October 2019

Abstract:

- In this article, we describe simultaneous inferential methods in detecting differentially expressed gene isoforms based on the Poisson generalized linear models. We derive the joint asymptotic distribution of pivotal quantities. The sample size of RNA sequencing data is often small in practice. Using multiple comparison procedures based on large-sample approximation becomes problematic. The parametric bootstrap method based on pivotal quantities is outlined as a robust alternative. Moreover, we observe the validity of robustness of the bootstrap method when mild overdispersion presents in RNA-sequencing data. We demonstrate the validity of the proposed method in detecting differentially expressed isoforms through Monte Carlo simulation. It shows the proposed method controls the family-wise error rate for large-scale inference. Even though the proposed method can be extended to many experimental designs, we focus on factorial designs in this article.

Key-Words:

- *RNA sequencing data; simultaneous inference; parametric bootstrap.*

1. INTRODUCTION

Studies of Gene isoform expression have not only been concentrated on detecting differentially expressed genes with known gene bank ID but also their isoforms due to the development of RNA sequencing technology. RNA sequencing technology, also known as Next Generation Sequencing (NGS), counts how many copies of nucleotide sequence for hundreds to thousands of gene isoforms.

To detect which genes are differentially expressed among hundreds even thousands of genes, researchers often conduct large-scale multiple hypotheses tests simultaneously, see Dudoit *et al.* [3]. One of the major concerns of gene expression analysis is to control the family-wise error rate (FWER). When the multiplicity is overlooked, researchers may claim dozens even hundreds of genes which are differentially expressed but in fact, they are false positives. Concerted efforts have been devoted to controlling FWER for microarray gene expression analysis. Dudoit *et al.* [4] applied Westfall and Young step-down method (Westfall and Young, [13]) based on two-sample Welch's t-tests to detect differentially expressed genes in microarray experiments. Alternatively, simultaneous confidence intervals based on the linear models of Kerr *et al.* [7] are constructed, see Hsu *et al.* [6]. Li and Mansouri [8] proposed simultaneous rank tests to search differentially expressed genes when microarray data violate normality assumption and contain a large number of outliers.

Auer and Doerge [1] proposed factorial designs for RNA sequencing experiments. To account for a variety of sources of variations, the resulting observations are fit to the Poisson generalized linear models, see Auer and Doerge [1]. Under this framework, we propose the simultaneous testing procedure to detect differentially expressed gene isoforms such that it controls FWER. Simultaneous test based on large-sample approximation is outlined. The sample size for RNA sequencing study is often small. As it will be shown in Section 4 that the large-sample approximation method does not provide a satisfactory solution in terms of controlling FWER. Monte Carlo simulation of Mansouri and Li [9] shows that percentile-t bootstrap method based on pivotal quantities provides a viable method in microarray gene expression analysis. Extension of bootstrap method to RNA sequencing gene expression analysis is hence appealing. In this article, we propose the simultaneous inferential method based on pivotal quantities to detect differentially expressed isoforms using parametric bootstrap. We investigate the performance of the proposed method in controlling the overall error rates through a simulation study.

2. PROBLEM FORMULATION AND PIVOTAL QUANTITIES

2.1. Experimental design and generalized linear model

To account for different sources of variations in observations from treatment, batch, flow cell, and lane, we consider factorial designs for the Next Generation Sequencing. In brief, bar-coded mRNA samples are pooled and assigned to different lanes of a sequencing device in such a way that there are n biological replicates randomly assigned at each combination

of treatment, lane, and flow cell. For details, see Auer and Doerge [1]. Since we can assign an ID to each isoform sequence in RNA sequencing data file, we may use the term "gene" instead of "isoform" in the following.

For gene l, $l = 1, ..., g$ we let Y_{lijkm} be the the count of readings from the i-th treatment, the j-th flow-cell, the k-th lane, and the m-th biological replicate, $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ..., c$, and $m = 1, ..., n$. We assume Y_{lijkm}'s are independent random observations and the expected value $E(Y_{lijkm}) = \mu_{lijk}$, for $m = 1, ..., n$ follow a per gene Poisson model with log-link (Auer and Doerge, [1]) that

$$\log(\mu_{lijk}) - \log(c_{jk}) = \alpha_l + \tau_{li} + \nu_{lj} + \omega_{lk} \tag{2.1}$$

where α_l is the overall gene l effect; τ_{li} is the i-th treatment effect on gene l with $\sum_i \tau_{li} = 0$; ν_{lj} is the j-th flow cell effect on gene l with $\sum_j \nu_{lj} = 0$; ω_{lk} is the k-th lane effect on gene l with $\sum_k \omega_{lk} = 0$; c_{jk} is a known constant, namely library size, $j = 1, ..., b$, $k = 1, ..., c$ to normalize the readings from j-th flow-cell and k-th lane, see Section 6 and Chen *et al.* [2]. We assume that $\alpha_l, \tau_{li}, \nu_{lj}, \omega_{lk}$, for $l = 1, ..., g$, $i = 1, ..., a$, $j = 1, ..., b$, and $k = 1, ..., c$ in (2.1) are fixed effects. Let $N = abcn$ be the total number of readings from each gene.

We let vector

$$\boldsymbol{Y}_l = \left[Y_{l1111}, ..., Y_{l111n}, ..., Y_{lijk1}, ..., Y_{lijkn}, ..., Y_{labc1}, ..., Y_{labcn}\right]'$$

be a collection of all readings from gene l and let $\boldsymbol{\mu}_l = E(\boldsymbol{Y}_l)$, $l = 1, ..., g$. It is useful to write the model in (2.1) in the form of matrix representation that

$$\log(\boldsymbol{\mu}_l / c_{jk}) = X\boldsymbol{\beta}_l \tag{2.2}$$

where $\boldsymbol{\beta}_l = \left[\alpha_l, \tau_{l1}, ..., \tau_{l(a-1)}, \nu_{l1}, ..., \nu_{l(b-1)}, \omega_{l1}, ..., \omega_{l(c-1)}\right]'$ and X is the corresponding $N \times (a + b + c - 2)$ design matrix.

Since we use per gene generalized linear model, the model for all genes can be written as

$$\mathbf{1}_g \otimes \log(\boldsymbol{\mu}_l / c_{jk}) = \mathbf{1}_g \otimes X\boldsymbol{\beta}_l \,. \tag{2.3}$$

2.2. Pivotal quantities

For gene l, $l = 1, ..., g$ we assume

$$Y_{lijkm} \sim \text{Poisson}(\mu_{lijk}), \qquad \text{for} \quad m = 1, ..., n \,, \tag{2.4}$$

where

$$\mu_{lijk} = \exp\left[(\alpha_l + \tau_{li} + \nu_{lj} + \omega_{lk}) + \log(c_{jk})\right] \tag{2.5}$$

with $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ..., c$, and $m = 1, ..., n$.

Let $\widehat{\boldsymbol{\beta}}_{l,N}$ be the maximum likelihood estimation of $\boldsymbol{\beta}_l$, $l = 1, ..., g$. We apply Newton–Raphson method using Fisher Scoring to compute the estimation. We may suppress the notation of the dependence on N and denote the estimation by $\widehat{\boldsymbol{\beta}}_l$.

Now, we define a $q \times (a+b+c-2)$ comparison matrix C to detect differential gene expression among treatments. In gene expression studies, researchers often interest in i) all-pairwise comparisons of gene expression over treatments, or ii) comparing gene expression for several treatments versus a control, Hsu *et al.* [6]. We focus on all-pairwise comparisons in this article and analogous results should hold for multiple comparisons to a control. As an example of comparison matrix C for all-pairwise comparisons, see (4.1) in Section 4.

Let W_l be $N \times N$ diagonal weight matrix whose diagonal elements are given by μ_{l1111}, ..., μ_{l111n}, ..., μ_{lijk1}, ..., μ_{lijkn}, ..., μ_{labc1}, ..., μ_{labcn} in order. The vector containing pivotal quantities is given by

$$\boldsymbol{T}(\boldsymbol{\beta}_l) = \Sigma_l^{-1/2}\big[C(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)\big] \tag{2.6}$$

where Σ_l is a diagonal matrix whose diagonal elements equal to the diagonal elements in $C(X'W_lX)^{-1}C'$, $l = 1, ..., g$.

In relation to the Poisson generalized linear model in (2.3), (2.4), and (2.5), consider gene expression by letting

$$\boldsymbol{T}(\boldsymbol{\beta}) = \big[\boldsymbol{T}(\boldsymbol{\beta}_1)', ..., \boldsymbol{T}(\boldsymbol{\beta}_l)', ..., \boldsymbol{T}(\boldsymbol{\beta}_g)'\big]'.$$

The joint limiting distribution of $\boldsymbol{T}(\boldsymbol{\beta})$ is given by the following Theorem.

Theorem 2.1. *Suppose $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_g$ are independent vectors, for $\frac{1}{N}(X'W_lX) \overset{N\to\infty}{\longrightarrow} \mathcal{W}_l$, which is positive definite, for $l = 1, ..., g$, then*

$$\sqrt{N}\boldsymbol{T}(\boldsymbol{\beta}) \xrightarrow{D} \text{MVN}(\mathbf{1}_g \otimes \mathbf{0}_q, \Lambda), \qquad \text{as} \quad N \to \infty, \tag{2.7}$$

where Λ is a $gq \times gq$ block diagonal matrix such that the l-th $(q \times q)$ diagonal block matrix $\Lambda_l = \lim_{N\to\infty} N\Sigma_l^{-1/2}C(X'W_lX)^{-1}C'\Sigma_l^{-1/2}$, $l = 1, ..., g$.

Proof of Theorem 2.1 immediately follows equation (5.25) and (S.17) of McCulloch *et al.* [10]. Note: since Λ_l is unknown in practice, we use a consistent estimator $\widehat{\Lambda}_l = N\widehat{\Sigma}_l^{-1/2}C(X'\widehat{W}_lX)^{-1}C'\widehat{\Sigma}_l^{-1/2}$ where $\widehat{\Sigma}_l$ is a diagonal matrix whose elements equal to the diagonal elements in $C(X'\widehat{W}_lX)^{-1}C'$, and $\widehat{W}_l$ has diagonal elements given by $\exp\{(\widehat{\alpha}_l + \widehat{\tau}_{l1} + \widehat{\nu}_{l1} + \widehat{\omega}_{l1}) + \log(c_{11})\}$, ..., $\exp\{(\widehat{\alpha}_l + \widehat{\tau}_{li} + \widehat{\nu}_{lj} + \widehat{\omega}_{lk}) + \log(c_{jk})\}$, ..., $\exp\{(\widehat{\alpha}_l + \widehat{\tau}_{la} + \widehat{\nu}_{lb} + \widehat{\omega}_{lc}) + \log(c_{bc})\}$ in order, $l = 1, ..., g$. In the expression, $\widehat{\alpha}_l$, $\widehat{\tau}_{li}$, $\widehat{\nu}_{lj}$, and $\widehat{\omega}_{lk}$ are maximum likelihood estimation of the parameters, $i = 1, ..., a$; $j = 1, ..., b$; $k = 1, ..., c$. Application of the large-sample approximation method is not trivial since the multivariate normal distribution in Theorem 2.1 has mean and variance with dimension $(gq) \times 1$ and $(gq) \times (gq)$ respectively and the total number of genes g, in RNA-sequencing experiments, is typically very large. We propose an Algorithm in Section 4 to reduce the computational burden in RNA-sequencing gene expression analysis.

A challenge besetting RNA-sequencing gene expression analysis may be the overdispersion among counting data, Auer and Doerge [1] and Wang *et al.* [11]. To proceed, we let ϕ_l be the dispersion parameter and overdispersion occurs when $\phi_l > 1$, $l = 1, ..., g$.

It is suggested in Auer and Doerge [1] that statistics for detecting differential gene expression should be scaled by the dispersion parameter. Hence, a sequence of pivotal quantities, considering overdispersion, are given by

$$\boldsymbol{T}(\boldsymbol{\beta}_l, \phi_l) = (\phi_l\Sigma_l)^{-1/2}\big[C(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)\big]. \tag{2.8}$$

The pivotal quantities in (2.6) can be considered as a special case of (2.8) when $\phi_l = 1$. We focus on gene expression analysis for RNA-sequencing data, which presents mild overdispersion such that ϕ_l is in a neighborhood of 1, and examine the validity of robustness of the large-sample approximation method through a simulation study in Section 4 in this article.

3. SIMULTANEOUS INFERENCE USING BOOTSTRAP

3.1. Simultaneous inference

In relation to the generalized linear model in (2.3), let the relative gene expression be $\tau_{li} - \tau_{li'},\ i \neq i' = 1, ..., a,\ l = 1, ..., g$. Detecting all-pairwise differential gene expression can be formulated as testing a sequence of hypotheses that:

$$H_{0_l,ii'}:\ \tau_{li} - \tau_{li'} = 0 \qquad \text{vs.} \qquad H_{1_l,ii'}:\ \tau_{li} - \tau_{li'} \neq 0 \tag{3.1}$$

for $i \neq i' = 1, ..., a,\ l = 1, ..., g$. Hence we conduct $q \times g$ tests simultaneously, where q is the number of rows in comparison matrix C such that $C\boldsymbol{\beta}_l = [\tau_{l1} - \tau_{l2}, ..., \tau_{l(a-1)} - \tau_{la}]'$, see (4.1) for example.

The resulting test statistics are given by

$$\boldsymbol{T}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l) = (\widehat{\phi}_l \widehat{\Sigma}_l)^{-1/2} C\widehat{\boldsymbol{\beta}}_l \tag{3.2}$$

for $l = 1, ..., g$ where the plug-in estimation of ϕ_l in Auer and Doerge [1] is given by

$$\widehat{\phi}_l = \left(\sum_{i,j,k,m} \frac{\left(Y_{lijkm} - \exp\{(\widehat{\alpha}_l + \widehat{\tau}_{li} + \widehat{\nu}_{lj} + \widehat{\omega}_{lk}) + \log(c_{jk})\}\right)^2}{\exp\{(\widehat{\alpha}_l + \widehat{\tau}_{li} + \widehat{\nu}_{lj} + \widehat{\omega}_{lk}) + \log(c_{jk})\}} \right) \Big/ \big(N - (a + b + c - 2)\big). \tag{3.3}$$

For gene l, write

$$\boldsymbol{T}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l) = \left[T_{12}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l), ..., T_{ii'}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l), ..., T_{(a-1)a}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)\right]'$$

in association to the hypotheses in (3.1) and the test statistics in (3.2). For all-pairwise comparisons, the total number of comparisons (the total number of elements in $\boldsymbol{T}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)$) $q = \binom{a}{2}$.

Simultaneous level-α tests reject hypothesis $H_{0_l,ii'},\ i \neq i' = 1, ..., a,\ l = 1, ..., g$ if:

$$\left|T_{ii'}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)\right| > q_\alpha \tag{3.4}$$

where q_α is the upper α-th quantile of the distribution of maximum modulus statistics $\max\limits_{\substack{i \neq i' = 1, ..., a \\ l = 1, ..., g}} \{|T_{ii'}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)|\}$.

When the magnitude of differential gene expression is of interest, a $(1-\alpha)$ 100% simultaneous confidence interval of $\tau_{li} - \tau_{li'},\ i \neq i' = 1, ..., a,\ l = 1, ..., g$ is given by

$$\boldsymbol{c}'_{ii'}\widehat{\boldsymbol{\beta}}_l \ \pm\ q_\alpha \big\{\widehat{\phi}_l\, \boldsymbol{c}'_{ii'}(X'\widehat{W}_l X)^{-1}\boldsymbol{c}_{ii'}\big\}^{1/2} \tag{3.5}$$

where $\boldsymbol{c}_{ii'}$ is the row vector of C in association to $\tau_{li} - \tau_{li'},\ i \neq i' = 1, ..., a$ for all $l = 1, ..., g$.

3.2. Bootstrap based on pivotal quantities

It can be shown that the upper α-th quantile of the multivariate normal distribution defined in (2.7) is a consistent estimator of q_α. RNA sequencing data analysis is often complicated by a large number of unknown parameters but a limited number of observations. Using the large-sample approximation method indicated by Theorem 2.1 can be problematic in the estimation of q_α as it will be shown in Section 4. We propose the parametric bootstrap method based on pivotal quantities to approximate quantiles q_α in detecting differentially expressed genes for RNA sequencing data.

For $r = 1, ..., B$, we define the $q \times 1$ vector of pivotal quantities based on the r-th bootstrap sample by

$$\boldsymbol{T}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l) = \big(\widehat{\phi}_l^{(r)} \widehat{\Sigma}_l^{(r)}\big)^{-1/2} C\big[\widehat{\boldsymbol{\beta}}_l^{(r)} - \widehat{\boldsymbol{\beta}}_l\big], \qquad l = 1, ..., g\,, \tag{3.6}$$

where $\widehat{\phi}_l^{(r)}$, $\widehat{\Sigma}_l^{(r)}$, and $\widehat{\boldsymbol{\beta}}_l^{(r)}$ are estimated based on the r-th bootstrap data set. Analogously, we write

$$\boldsymbol{T}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l) = \Big[T_{12}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l), ..., T_{ii'}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l), ..., T_{(a-1)a}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)\Big]'.$$

We use the following Algorithm to approximate quantiles q_α. For each r, $r = 1, ..., B$,

(i) for each l, $l = 1, ..., g$ generate random variables $\{Y_{lijkm}\}$ from Poisson$\big(\exp\big\{(\widehat{\alpha}_l + \widehat{\tau}_{li} + \widehat{\nu}_{lj} + \widehat{\omega}_{lk}) + \log(c_{jk})\big\}\big)$, $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ..., c$, and $m = 1, ..., n$;

(ii) obtain maximum modulus statistics

$$T_M^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l) = \max_{i \neq i' = 1, ..., a} \big\{\big|T_{ii'}^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)\big|\big\}\,, \qquad l = 1, ..., g\,,$$

and

$$T_M^{(r)}(\widehat{\boldsymbol{\beta}}, \widehat{\phi}) = \max_{l = 1, ..., g} \big\{T_M^{(r)}(\widehat{\boldsymbol{\beta}}_l, \widehat{\phi}_l)\big\}.$$

Repeat (i) and (ii) B times, and the upper α-th quantile of the sampling distribution of $T_M^{(r)}(\widehat{\boldsymbol{\beta}}, \widehat{\phi})$ is an approximation of q_α.

As it will be shown in Section 4, the bootstrap method provides a viable alternative of the large-sample approximation method when the overdispersion parameter is in a neighborhood of $\phi_l = 1$, $l = 1, ..., g$.

4. SIMULATION STUDY

In this section, we investigate the performance of the proposed method in terms of controlling the family-wise error rate (FWER) using Monte Carlo simulation.

We assign the following values to the parameters of the model in (2.1). Let

$$\tau_{li} = 0, \quad \text{for } l = 1, ..., 20, \quad i = 1, 2, 3, 4 \quad (\textit{Complete Null})\,,$$
$$\tau_{li} = 0, \quad \text{for } l = 1, ..., 15, \quad i = 1, 2, 3, 4 \quad (\textit{Partial Null})\,.$$

To study the power rates under partial null hypotheses, we let $\tau_{l1} = -0.02$, $\tau_{l2} = 0.01$, $\tau_{l3} = 0.01$, and $\tau_{l4} = 0$, for $l = 16, ..., 20$.

For nuisance parameters, we let $\alpha_l = -3$ and

$$\nu_{lj} = \begin{cases} 0.5, & \text{if } j = 1, \\ -1, & \text{if } j = 2, \\ 0.5, & \text{if } j = 3, \end{cases}$$

for $l = 1, ..., 20$. Let

$$\omega_{lk} = \begin{cases} 0.25, & \text{if } k = 1, \\ -0.5, & \text{if } k = 2, \\ 0.75, & \text{if } k = 3, \\ -1.25, & \text{if } k = 4, \\ 1.5, & \text{if } k = 5, \\ -0.75, & \text{if } k = 6, \end{cases}$$

for $l = 1, ..., 20$.

Assume the library size for each lane and flow cell $c_{jk} = 1,000,000$ for all $j = 1, 2, 3$ and $k = 1, ..., 6$.

We may rewrite the model in (2.1) as $\log(\lambda_{lijk}) = \alpha_l + \tau_{li} + \nu_{lj} + \omega_{lk}$, where the sampling rate $\lambda_{lijk} = E(Y_{lijk}/c_{jk})$ and c_{jk} is a given constant. The observations Y'_{lijkm} are generated from Poisson(μ_{lijk}) where $\mu_{lijk} = c_{jk}\lambda_{lijk}$, for $m = 1, 2$. To exam the performance of the proposed method under mild overdispersion, we add Gaussian noise $\epsilon_{lijkm} \sim N(0, (\phi_l - 1)\mu_{lijk})$ ($\phi_l > 1$) to the observations that $Y_{lijkm} = Y'_{lijkm} + [\epsilon_{lijkm}]$, $i = 1, ..., 4$, $j = 1, 2, 3$, $k = 1, ..., 6$, $m = 1, 2$ for gene l, $l = 1, ..., 20$ as it is treated in Auer and Doerge [1]. Note that $E(Y'_{lijkm} + \epsilon_{lijkm}) = \mu_{lijk}$ and $\text{Var}(Y'_{lijkm} + \epsilon_{lijkm}) = \phi_l\mu_{lijk}$. We choose $\phi_l = 1.1$, 1.05, 1.01, and 1.001 respectively and let $Y_{lijkm} = Y'_{lijkm}$ for $\phi_l = 1$. In addition, we let the observations equal to zero if it generates "negative" counts, though the chance of generating "negative" counts is rare when the value of $(\phi_l - 1)$ is small.

Hence, the vector of parameters $\boldsymbol{\beta}_l = [\alpha_l, \tau_{l1}, \tau_{l2}, \tau_{l3}, \nu_{l1}, \nu_{l2}, \omega_{l1}, ..., \omega_{l5}]'$, $l = 1, ..., g$. Let X be the corresponding design matrix for all genes. Consider all-pairwise comparisons among treatments. Let C be the 6×11 comparison matrix given by

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & 0 & \cdots & 0 \\ 0 & 2 & 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 2 & 0 & \cdots & 0 \end{bmatrix}. \tag{4.1}$$

We run simultaneous tests in (3.4) 1,000 times and compute the empirical overall error rates. Widely used measures of the overall error rates in gene expression analysis are the family-wise error rate (FWER) and the false discovery rate (FDR). Let FWER_0 be the probability that at least one true null hypotheses rejected under complete null hypotheses. Let FWER_1 be the probability that at least one true null hypotheses rejected under partial null hypotheses. The false discovery rate (FDR) is computed as the average proportion of

wrongly rejected null hypotheses among all rejected hypotheses. FDR is defined as 0 if no rejection were made. To investigate the power of the simultaneous tests, we compute the proportional power rate by obtaining the average proportion of genes found differentially expressed among all misexpressed genes, Dudoit *et al.* [3].

To evaluate the performance of the large-sample approximation method, we use the following Algorithm to generate quantiles based on the multivariate normal distribution defined in Theorem 2.1. In specific, for each r, $r = 1, ..., B$,

(**i***) generate random variables $\boldsymbol{T}_l^{(r)}$ from MVN$(\mathbf{0}, \widehat{\Lambda}_l)$, for all $l = 1, ..., g$;

(**ii***) obtain maximum modulus statistics $T_{M_l}^{(r)} = \max\{|\boldsymbol{T}_l^{(r)}|\}$, $l = 1, ..., g$ and $T_M^{(r)} = \max\{T_{M_l}^{(r)}\}$.

Repeat (i*) and (ii*) B times, and the upper α-th quantile of the empirical distribution of $T_M^{(r)}$ is an approximation of q_α based on Theorem 2.1.

The performance of the large-sample approximation method and the bootstrap method in the simulation study are summarized in Table 1.

Table 1: Error rates of detecting differentially expressed genes/isoforms — nominal type-1 error rate $\alpha = 0.05$.

Method	ϕ_l	FWER_0	FWER_1	FDR	Prop. Power
No Adjustment	1.000	0.993	0.970	0.146	—
MVN	1.000†	0.072	0.059	0.003	0.889
	1.050 (1.1)‡	0.084	0.061	0.003	0.861
	1.010	0.073	0.045	0.002	0.887
	1.001	0.065	0.065	0.003	0.886
Bootstrap Method	1.000	0.052	0.037	0.002	0.878
	1.050 (1.1)	0.051	0.035	0.002	0.849
	1.010	0.049	0.034	0.002	0.874
	1.001	0.050	0.045	0.002	0.872

Notes:

i) Simulation size = 1,000. Bootstrap size $B = 200$.
ii) FWER_0 denotes the family-wise error rate under complete null hypotheses.
iii) FWER_1 denotes the family-wise error rate under partial null hypotheses.
iv) MVN denotes the method of large-sample approximation in Section 3.1.
v) "Bootstrap Method" means the parametric bootstrap method in Section 3.2.
vi) † The same value of ϕ_l is assigned to all genes.
vii) ‡ The first 15 genes have $\phi_l = 1.05$ and the last 5 genes have $\phi_l = 1.1$.
viii) The total computation user time was about 16 hours on a desktop with processor with the following specifications: Intel(R) Core(TM) i5-7600 CPU @ 3.50GHz, 3504 Mhz and Installed physical memory (RAM): 16.0 GB.

It shows that the bootstrap method based on pivotal quantities controls FWER under both complete and partial null hypotheses. This implies the proposed method controls FWER strongly, see Dudoit *et al.* [3]. Without adjustment of multiplicity, it is well known that the overall error rates often exceed the nominal level, particularly in large-scale tests.

Simultaneous tests based on large-sample approximation fail to control FWER in the strong sense in RNA sequencing data analysis. While the overall error rates are controlled at nominal level $\alpha = 0.05$, in average more than 85% of "real" misexpressed genes are detected as differentially expressed genes using the bootstrap method in Section 3.2. Note that it is not useful to address the power rates when the method does not control FWER.

To investigate the performance of the bootstrap method in estimation of quantiles, we generate $1,000$ samples as described above and obtain the $(1-\alpha)$-th quantile of the sampling distribution of pivotal quantities in (2.8). Since the quantiles are generated from a given underlying distribution of maximum modulus distribution empirically, it can be used as a benchmark to evaluate the performance of the proposed method. The results are summarized in Table 2.

Table 2: Quantiles q_α of detecting differentially expressed genes/isoforms — nominal type-1 error rate $\alpha = 0.05$.

ϕ_l	Simulation	MVN	Bootstrap
1.000	3.604	3.519 (0.090)	3.606 (0.094)
1.050 (1.1)	3.617	3.522 (0.086)	3.611 (0.094)
1.010	3.605	3.524 (0.090)	3.609 (0.095)
1.001	3.604	3.516 (0.084)	3.604 (0.097)

Notes:

i) Simulation size = 1,000. Bootstrap size $B = 200$.

ii) MVN denotes the method of large-sample approximation in Section 3.1 and the Algorithm in Section 4. The quantile is generated from $B = 200$ samples. We repeat the process for $1,000$ times. The mean value of these repeats is included outside of the parentheses and standard deviation is tabulated in the parentheses.

iii) "Bootstrap" means the parametric bootstrap method in Section 3.2. The quantile is generated from $B = 200$ bootstrap samples. We repeat the process for 1,000 times. The mean value of these repeats is included outside of the parentheses and standard deviation is tabulated in the parentheses.

iv) "Simulation" means: we generate observations from the model in (2.1) with the parameter value assigned in Section 4 and given underlying distributions for $1,000$ times; the upper α-th quantile of maximum modulus statistics based on pivotal quantiles in (2.8) is tabulated in the table.

v) The total computation user time was about 8 hours on a desktop with processor with the following specifications: Intel(R) Core(TM) i5-7600 CPU @ 3.50GHz, 3504 Mhz and Installed physical memory (RAM): 16.0 GB.

It shows from Table 2 that the bootstrap quantiles in Section 3.2 are closer to the simulated quantiles as compared to that generated from MVN. A closer examination sees the quantiles based on normal theory are generally below the simulated quantiles. Therefore, the large-sample approximation method provides a liberal estimation of FWER, as evidenced in Table 1.

5. CONCLUSION AND FUTURE WORK

In this article, we have proposed the parametric bootstrap method based on pivotal quantities in detecting differentially expressed genes for RNA-sequencing data. We have formulated the problem using the Poisson generalized linear models. We have derived the joint limiting distribution of the vector containing pivotal quantities. We have conducted an empirical study to show that the proposed method controls FWER and FDR strongly in detecting differentially expressed genes. The bootstrap method requires a large computation time, parallel computation is recommended particularly for large-scale inference. When data "apparently" violate Poisson distributional assumption, we will investigate the methods involving a large value of overdispersion. To capture the within genes' variation and between genes' variation, we will study the resampling methods, such as moving block bootstrap method in the future work.

6. SOFTWARE

We use the function `glm()` in `R` to obtain maximum likelihood estimation of the parameters in model (2.1). Note that computation of the estimation using `glm()` in `R` may encounter non-convergence. Alternatively, iterative weighted least squares method of Wedderburn [12] may be used in the estimation. Our experience in the simulation study (results not shown) shows that using 20-step iterative weighted least squares method provides satisfactory approximation of the overall Type-I error rates. We use the function `rmvnorm()` of Genz *et al.* [5] in `R` to generate multivariate normal random variables. We use the function `calcNormFactors()` of Chen *et al.* [2] to obtain the library size. Software in the form of `R` code is available on request from the author (`bli@citadel.edu`).

ACKNOWLEDGMENTS

The author would like to thank the referees for providing valuable comments on this article.

REFERENCES

[1] Auer, P.L. and Doerge, R.W. (2010). Statistical Design and Analysis of RNA Sequencing Data, *Genetics*, **185**, 405–416.

[2] Chen, Y.; Lun, A.T.L. and Smyth, G.K. (2014). *Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR*. In: "Statistical Analysis of Next Generation Sequencing Data" (S. Datta and D. Nettleton, Eds.), New York, Springer.

[3] Dudoit, S.; Shaffer, J.P. and Boldrick, J.C. (2003). Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science*, **18**(1), 71–103.

[4] Dudoit, S.; Yang, Y.H.; Callow, M.J. and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111–139.

[5] Genz, A.; Bretz, F.; Miwa, T.; Mi, X.; Leisch, F.; Scheipl, F. and Hothorn, T. (2017). mvtnorm: Multivariate Normal and t Distributions, R package version 1.0-6.
URL: http://CRAN.R-project.org/package=mvtnorm

[6] Hsu, J.C.; Chang, J.Y. and Wang, T. (2006). Simultaneous confidence intervals for differential gene expressions, *Journal of Statistical Planning and Inference*, **136**(7), 2182–2196.

[7] Kerr, M.K.; Martin, M. and Churchill, G.A. (2000). Analysis of Variance for Gene Expression Microarray Data, *Journal of Computational Biology*, **7**(6), 818–837.

[8] Li, B. and Mansouri, H.G. (2016). Simultaneous Rank Tests for Detecting Differentially Expressed Genes, *Journal of Statistical Computation and Simulation*, **86**(5), 959–972.

[9] Mansouri, H.G. and Li, B. (2019). On simultaneous confidence intervals based on rank-estimates with application to analysis of gene expression data, *Communications in Statistics – Theory and Methods*, **48**(17), 4339–4349.
DOI: 10.1080/03610926.2018.1494287

[10] McCulloch, C.E.; Searle, S.E. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*, New Jersey, Wiley, 2008.

[11] Wang, J.; Huang, M.; Torre, E.; Dueck, H.; Shaffer, S.; Murray, J.; Raj, A.; Li, M. and Zhang, N.R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing, *Proceedings of the National Academy of Sciences*, **115**(28), E6437–E6446.
DOI: 10.1073/pnas.1721085115

[12] Wedderburn, R.W.M. (1974). Quasi-likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method, *Biometrika*, **61**(3), 439–447.

[13] Westfall, P.H. and Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, New York, Wiley, 1993.