Overview of Simultaneous Inference of Relative Gene Isoform Expressions for RNA Sequencing Factorial Designs

Bo Li *

Abstract

In this article, we overview a simultaneous inference method in Li (2020) to detect differentially expressed genes based on Poisson generalized linear models for RNA sequencing factorial designs. We use a real example for illustration.

Key Words: RNA Sequencing Gene Expression Analysis; Simultaneous Confidence Intervals; Simultaneous Tests; Asymptotic Distribution of Pivotal Quantities

1. Introduction

In RNA sequencing data analysis, researchers often conduct large scale multiple hypothesis testing for relative gene isoform expressions simultaneously. A major consideration is to control the familywise error rate (FWER), otherwise the probability of claiming at least one false positives can easily go up to 1, see Li (2020). We may call gene isoforms, genes in the following for simplicity.

Chen et al (2014) apply LRT test statistics for relative gene expressions and adjust $p - value$ using popular multiple comparisons, such as Bonferroni method, Holm's step-down multiple hypothesis testing method, or Benjamini and Hochberg step-up multiple hypothesis testing method. Li et al (2012) attempt to control the empirical false discovery rate in detecting differentially expressed genes based on permutation datasets. Note that under the complete null hypothesis, the false discovery rate equals the familywise error rate; under the partial null hypothesis, the false discovery rate is bounded below by the familywise error rate, see Dudoit et al (2003). For an overview of these methods, see Li (2019). For RNA sequencing factorial designs, Li (2020) proposed simultaneous inference method to detect differentially expressed genes based on gene wise Poisson generalized linear models. In this article, we overview a simultaneous inference method in Li (2020). We use a real example to illustrate the application.

2. Simultaneous Inference for Gene Expressions

Let Y_{lijk} be the independent count from $k-th$ replicate under $i-th$ treatment group and $j-th$ block of gene l, $i = 1, 2,\ j = 1, \cdots, a,\ k = 1, \cdots, n_{ij}$ and $l = 1, \cdots, g$ such that

$$log(E(Y_{lijk})) - log(c_{ijk}) = \gamma_l + \tau_{li} + \beta_{lj} \tag{2.1}$$

*Department of Mathematical Sciences, The Citadel, The Military College of South Carolina, Charleston, SC, 29409.

where γ_l is the mean effect of gene $l,\ l = 1, \cdots, g$; τ_{li} is the $i-th$ treatment effect on gene l with constraint $\tau_{l1} + \tau_{l2} = 0$ for each l; β_{lj} is the $j-th$ block effect on gene l with constraint $\sum_j \beta_{lj} = 0$ for each l; c_{ijk} is a scaler for normalization; for each triplet $i,\ j$ and k, we take the $0.75-th$ quantile of $Y'_{lijk}s,\ l = 1, \cdots, g$ as c_{ijk}, see Bullard et al (2010). The total number of observations for each gene is given by $N = \sum_{i,j} n_{ij}$.

We estimate the parameters in (2.1) using maximum likelihood estimation method. We apply the iterative weighted least squares method of Wedderburn (1974) for estimation. Denote $\widehat{\gamma}_l,\ \widehat{\tau}_{li}$, and $\widehat{\beta}_{lj}$ the maximum likelihood estimation of $\gamma_l,\ \tau_{li}$, and $\beta_{lj},\ i = 1, 2,\ j = 1, \cdots, a$ for each $l,\ l = 1, \cdots, g$.

For gene $l,\ l = 1, \cdots, g$ test the hypotheses that

$$H_{0_l} : \tau_{l1} - \tau_{l2} = 0 \ \ vs. \ \ H_{1_l} : \tau_{l1} - \tau_{l2} \neq 0. \tag{2.2}$$

To proceed, we define a sequence of pivotal quantities in association to $\tau_{l1} - \tau_{l2}$ given by

$$T_l(\tau_{l1}, \tau_{l2}) = \widehat{\sigma}_l^{-1}[(\widehat{\tau}_1 - \widehat{\tau}_2) - (\tau_1 - \tau_2)], \tag{2.3}$$

$l = 1, \cdots, g$. Let $\boldsymbol{T}(\boldsymbol{\tau}) = [T_1(\tau_{11}, \tau_{12}), \cdots, T_g((\tau_{g1}, \tau_{g2}))]'$. It follows the proof of Theorem 2.1 of Li (2020) that the joint limiting distribution is given in the following Corollary.

Corollary 2.1. *For independent observations* $Y_{1111}, \cdots, Y_{l2an_{2a}}, \cdots, Y_{g111}, \cdots, Y_{g2an_{2a}}$, *assume that* $\frac{1}{N}(X'W_lX) \longrightarrow V_l$, *a positive definite matrix, as* $N \longrightarrow \infty$, *for all* $l,\ l = 1, \cdots, g$. *We have*

$$\boldsymbol{T}(\boldsymbol{\tau}) \xrightarrow{D} MVN(\mathbf{0}_g, I_g),\ as\ N \longrightarrow \infty \tag{2.4}$$

where $W_l = diag\{\mu_{l111}, \cdots, \mu_{l2an_{2a}}\},\ l = 1, \cdots, g$, $\mathbf{0}_g$ *is* $g \times 1$ *vector of* $0's$ *and* I_g *is* $g \times g$ *identity matrix.*

Let X be the design matrix corresponding to the right hand side of (2.1). Consider the overdispersion ϕ_l among observations of gene $l,\ l = 1, \cdots, g$. We use the plug-in estimation of ϕ_l in Auer and Doerger (2010) given by

$$\widehat{\phi}_l = (\sum_{i,j,k} \frac{(Y_{lijk} - exp\{(\widehat{\gamma}_l + \widehat{\tau}_{li} + \widehat{\beta}_{lj}) + log(c_{ijk})\})^2}{exp\{(\widehat{\gamma}_l + \widehat{\tau}_{li} + \widehat{\beta}_{lj}) + log(c_{ijk})\}})/(N - (1 + a)). \tag{2.5}$$

Under the complete null hypothesis $\cap_{l=1}^{g} H_{0_l}$, the test statistic in Li (2020) is given by

$$T_l = \widehat{\sigma}_l^{-1}(\widehat{\tau}_1 - \widehat{\tau}_2), \tag{2.6}$$

for each $l,\ l = 1, \cdots, g$ where $\widehat{\sigma}_l^2$ is given by the second diagonal element of $\widehat{\phi}_l(X'\widehat{W}_lX)^{-1}$ quadruple, $\widehat{W}_l = diag\{\widehat{\mu}_{l111}, \cdots, \widehat{\mu}_{l2an_{2a}}\}$ and $\widehat{\mu}_{lijk} = exp\{(\widehat{\gamma}_l + \widehat{\tau}_{li} + \widehat{\beta}_{lj}) + log(c_{ijk})\},\ i = 1, 2,\ j = 1, \cdots, a$ and $k = 1, \cdots, n_{ij}$.

Let q_α be the upper $\alpha/2 - th$ quantile of the multivariate normal distribution in (2.4). We have that a two-sided level-α simultaneous test for H_{0_l} in (2.2) is given by rejecting H_{0_l} if $|T_l| > q_\alpha,\ l = 1, \cdots, g$.

Analogously, when the magnitude of gene expressions is of interest, a $(1-\alpha)100\%$ simultaneous confidence interval of $\tau_{l1}-\tau_{l2}$ is given by

$$[(\widehat{\tau}_{l1}-\widehat{\tau}_{l2})-q_\alpha\widehat{\sigma}_l,(\widehat{\tau}_{l1}-\widehat{\tau}_{l2})+q_\alpha\widehat{\sigma}_l], \tag{2.7}$$

$l=1,\cdots,g$.

Note that the number of genes in RNA sequencing datasets is often large. For instance, the number of genes in section 3 is 12839; $i.e.$, the dimension of the multivariate normal random vector in (2.4) is 12839. Hence, it is challenging to approximate the quantiles q_α. In section 4 of Li (2020), the algorithm based on simulation is used to approximate q_α, alternatively.

3. Example

To study the effect of strains on gene expressions, Bottomly et al (2011) compared the RNA sequencing gene expressions in C57BL/6J and DBA/2J mouse striatum, namely T1 and T2, respectively. Flow-cells FC1, FC2 and FC3 are administrated in the sequencing devices in such a way that 3 replicates are assigned for the combination $T1$ and $FC1$, 4 for $T1$ and $FC2$, 3 for $T1$ and $FC3$, 4 for $T2$ and $FC1$, 3 for $T2$ and $FC2$, and 4 for $T2$ and $FC3$. The data is from ReCount RNA Sequencing Database "http://bowtie-bio.sourceforge.net/recount/". The observations are the number of copies of the gene sequences under each configuration above for 12839 genes. Note that it follows Bottomly et al (2011) that we exclude genes, whose observations are all $0's$ in at least one of $T1$ and $T2$ groups.

Let the strains $T1$ and $T2$ be the treatment effects on gene expressions. Let the flow-cell assignments $FC1,\ FC2$ and $FC3$ be the block effects on gene expressions. We fit the observations to the Poisson model in (2.1) and test the hypotheses in (2.2) based on the test statistics in (2.6). Using the result in Corollary 2.1, the quantile $q_{0.05}=4.61$. We have that 596 genes are differentially expressed with nominal significance level 0.05. The 95% simultaneous confidence intervals for the top 10 (ranked by the absolute value of T_l in (2.6)) differentially expressed genes are listed in Table 1.

As a side note for the interpretation of the result, notice that the simulation study in Li (2020) shows that the large-sample approximation method in section 2.1 can be anti-conservative in overdispersed data. In the future study, we will develop robust simultaneous inference method when overdispersion besets RNA sequencing gene expression analysis.

Table 1: Simultaneous Confidence Intervals for the Top 10 Differentially Expressed Genes - Nominal Confidence Level $1 - \alpha = 0.95$.

Gene ID	$\lvert T_l \rvert$	Interval Estimation
$ENSMUSG00000030532$	26.193	$(-1.231, -0.863)$
$ENSMUSG00000023236$	24.289	$(-1.148, -0.782)$
$ENSMUSG00000015484$	23.125	$(1.130, 1.693)$
$ENSMUSG00000037461$	21.493	$(0.382, 0.591)$
$ENSMUSG00000062822$	18.600	$(0.889, 1.475)$
$ENSMUSG00000028393$	17.955	$(-1.514, -0.895)$
$ENSMUSG00000024248$	16.978	$(1.665, 2.906)$
$ENSMUSG00000054579$	16.812	$(-1.865, -1.062)$
$ENSMUSG00000056592$	16.704	$(-0.998, -0.566)$
$ENSMUSG00000024206$	16.589	$(-0.858, -0.485)$

The "Interval Estimation" denotes the 95% *simultaneous confidence interval given in (2.7) based on normal theory Corollary 2.1.*

REFERENCES

Auer P. L., Doerge R.W. (2010), "Statistical Design and Analysis of RNA Sequencing Data," *Genetics*, 185(2), 405 – 416.

Bottomly D., Walter N.A.R., Hunter J.E., Darakjian P., Kawane S., Buck K.J., Searles R.P., Mooney M., McWeeney S.K., Hitzemann R. (2011), "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays," *PLoS ONE*, 6(3), e17820.

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, 94.

Chen Y.S., Lun A.T.L., Smyth G.K. (2014), "Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR," in *Statistical Analysis of Next Generation Sequencing Data*, eds. Somnath Datta and Daniel S Nettleton, New York: Springer.

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

Li B. (2019), "Multiple Hypothesis Testing in RNA Sequencing Gene Isoform Expression Analysis," *JSM Proceedings*, 2404 -2408.

Li B. (2020), "Simultaneous Inference of Differentially Expressed Isoforms for RNA Sequencing Data," *REVSTAT – Statistical Journal*, 18(2), 153 – 163.

Li J., Witten D.M., Johnstone I.M., Tibshirani R. (2012), "Normalization, testing, and false discovery rate estimation for RNA-sequencing data," *JBiostatistics*, 13(3), 523–538.

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