Estimating test statistics correlation from sample correlation

1 Introduction

What's the consequence if the correlation between statistics cannot be represented sample correlation?

1. Methods relating FDR control in terms of type I error seems to be OK?? Because under the null, test statistics correlation are (almost) the same as sample correlation.

2. What about power in terms of FDR control?

3. Competitive gene set test would definitely be affected, in terms of both type I error and power.

4. it seems, according to Efron's 2007 paper, that conditional FDR will also be affected.

What problem do we address in this paper?

Introduction

In gene expression experiments, inter-gene correlations are commonly observed in expression data [1; 2; 3; 4; 5; 6; 7; 8]. The key task of expression analysis is to detect differentially expressed (DE) genes. One common feature of such DE detection is that a summary statistic is calculated for each gene to measure the magnitude of DE. The test statistics are often of familiar form, for example, two-sample comparison or experimental design based regression. However, those test statistics are likely to be correlated, since their corresponding expression levels are correlated. This paper concerns the relation between test statistics correlations and the corresponding expression level correlations.

Why would people care about correlation between genes? The stochastic dependence of test statistics has brought methodological issues, in terms of accessing both individual genes and gene sets. The interest in examining individual genes is to find DE genes among tens of thousands of candidates. Multiple hypothesis testing procedures, such as *false discovery rate* (FDR, **(author?)** 9) and *q-value* [5], are therefore needed. In many cases, such techniques work only when test statistics are independent [9] or have positive regression dependency [10]. The goal of evaluating gene sets is to find molecular pathways or gene networks that are related to the experimental condition or factors of interest. Testing a gene set is usually done by pooling the test statistics of its member genes, and may or may not involve genes not in the test set [11]. In all situations, the correlation between test statistics is a nuisance aspect, which, if not addressed appropriately, will undermine the applicability of the corresponding approaches (REF). For example, **(author?)** [4] showed in a simulation study that for a nominal FDR of 0.1, the actual FDR can easily vary by a factor of 10 when correlation between test statistics exists.

What are existing ways of dealing with inter-gene correlations? A number of attempts have been made to deal with issues of inter-gene correlation when testing either individual genes or gene sets. One approach is to derive certain summary statistic from correlation among test statistics and then use it in the hypothesis testing procedure. (Do I need more examples here) For testing individual genes, **(author?)** [4] calculates the *false discovery proportion* (FDP) conditioning on some dispersion variate which is estimated from correlation among transformed test statistics. For testing gene sets, **(author?)** [7] estimate a *variance inflation factor* (VIF) associated with inter-gene correlation and incorporate it into their parametric/rank-based testing procedures. The same VIF is also used by **(author?)** [12] to account for correlation in their distribution-based gene set testing procedure. Another approach is to permute the labels of biological samples. Sample permutation generates the null distribution of test statistic for each gene. This type of permutation preserves underlying correlation structure between genes, and thus protect the test against such correlations (REF, FDR related and gene set test related). However, sample permutation method has an extra assumption, which states that the test statistics always follow the distribution they have under complete null that no gene is DE [13]. In other words, this assumption expects that the distribution of test statistics under the null is not affected by the presence of non-null cases. The *gene set enrichment analysis* (GSEA) procedure [14] falls into this category.

Key question: Are expression level correlations the same as test statistics correlation?

The first approach requires that the correlations between test statistics are known or at least can be estimated from the data. Without replicating the experiment, however, there's no way to obtain the correlation structure of test statistics because only a single test statistic is available for each gene. In the case of one-sided test (e.g., two sample t-test), one possible choice is to use sample correlations (after gene treatment effects nullified) to represent correlations among test statistics, as is done by **(author?)** [3, 4, 7]. In all of the three works, it is shown by simulation only the equivalence (in terms of either distribution or numerical summarization) of sample correlation coefficient and test statistics correlation coefficient. **(author?)** [4] estimates the distribution of z-value (transformed from corresponding two sample t-test statistics) correlation by sample correlation. **(author?)** [3] show by Monte Carlo simulation of gene expression data that a nearly linear relationship holds between test statistic correlation and sample correlation for several types of test statistic. It has, to the best of our knowledge, not yet been fully explored in the context of two group comparison.

What did we find

In this work, we investigated the effect of testing procedures on inter-gene correlation structure regarding two group comparison. Theoretically, we proved that for two sample z-test, there is a perfect positive correlation between sample correlation coefficient r_{sample} and test statistics correlation $r_{\text{statistic}}$. For two sample t-test, the equivalence does not hold in general for $r_{\text{statistic}}$ and r_{sample}, unless all the test are true null (no DE). We demonstrated by simulation that under the null, such equivalence also holds for two group comparison of Poisson regression.

Relevant but different work

A relevant research was done by **(author?)** [2], in which they studied the effect of different normalization procedures on the inter-gene correlation structure for microarray data. They randomly assigned 330 arrays into 15 pairs, each containing 22 arrays within each array 12558 genes. Then 15 t-statistics were calculated for each gene to mimic 15 two-sample comparisons under null hypothesis of no DE. They compared the histogram of t-statistics correlation for different normalization algorithms, and concluded that the normalization procedures are unable to completely remove the correlation between the test statistics.

2 General setup

2.1 define what do we mean by correlation

Correlation is a statistical quantity used to assess a possible linear relationship between two random variables or two sets of data sets. The degree of correlation is measured by *correlation coefficient*, a scaler taking values on the interval $[-1, 1]$. Correlation coefficient of $+1$ (-1) indicates perfect positive (negative dependence), while correlation coefficient of 0 implies no linear relationship between two random variables. Larger correlation coefficient (in absolute value) corresponds to stronger linear correlation. There are many ways to look at the correlation coefficient , many of which are special cases of Pearson's correlation coefficient [15]. For example, the *Kendall tau rank correlation coefficient* is computed as Pearson's correlation coefficient between the ranked variables.

Let (X, Y) be a random vector, and (x_j, y_j) its jth observation. The most familiar measure of dependence between two quantities is the *Pearson's correlation coefficient*. Following the notation of [15], We will restrict our interest to two types of Pearson's correlation coefficient: 1) standardized covariance, which we refer to as *population correlation*

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{1}$$

where μ_X and μ_Y are the expected values and $\sigma_X < \infty$ and $\sigma_Y < \infty$ are the population standard errors, and 2) a function of raw scores and means, which we refer to as *sample correlation*

$$r = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_i (y_j - \bar{y})^2}}, \tag{2}$$

where $(\bar{x}, \bar{y})$ is the vector of arithmetic mean of the observations. Throughout this paper, we will discuss the correlation between X and Y under bivariate settings.

In a two group comparison experiment (e.g., gene expression of treatment and control), let (X_j, Y_j) be the expression level for a pair of genes in sample j where $j = 1, \ldots, n_1$ if it's from the treatment group and $j = n_1+1, \ldots, n_1+n_2$ if it's from the control group. We assume that the covariance for different samples are the same, but the mean level might differ between the two groups, that is,

$$\mathrm{Cov}\begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \stackrel{\mathrm{def}}{=} \mathbf{\Sigma} \tag{3}$$

and

$$\begin{aligned} E\begin{pmatrix} X_j \\ Y_j \end{pmatrix} &= \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\mu}, \quad \text{for } j = 1, \ldots, n_1 \\ E\begin{pmatrix} X_j \\ Y_j \end{pmatrix} &= \begin{pmatrix} \mu_X + \Delta_X \\ \mu_Y + \Delta_Y \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\mu} + \boldsymbol{\Delta}, \quad \text{for } j = n_1 + 1, \ldots, n_1 + n_2. \end{aligned} \tag{4}$$

Here ρ is the population correlation defined by equation (1), and $\boldsymbol{\Delta} = (\Delta_X, \Delta_Y)$ is vector of treatment effect for the pair of genes. In addition, We assume independence across samples (Note that independence implies 0 correlation, but not vise versa),

$$\text{Cov}(X_{j_1}, X_{j_2}) = \text{Cov}(Y_{j_1}, Y_{j_2}) = 0 \quad \text{for } j_1 \neq j_2 \tag{5}$$

In the context of gene expression study, the goal is to detect differential expression (DE)—the mean difference between treatment and control group, which can be statistically formulated as

$$H_{0i} : \Delta_i = 0 \ \textit{Verses}\ H_{1i} : \Delta_i \neq 0, \quad i = X, Y. \tag{6}$$

This hypothesis testing procedure usually results in a "t-test similar" test statistic for each gene. Without a loss of generality, we express the test statistics as follows

$$T_X = \frac{\boldsymbol{a}^T \boldsymbol{X}}{S_X}, \quad T_Y = \frac{\boldsymbol{a}^T \boldsymbol{Y}}{S_Y} \tag{7}$$

where $\boldsymbol{a}$ is a vector of length $n_1 + n_2$ denoting coefficient for comparison(e.g., 1 for treatment and -1 for control), $\boldsymbol{X} = (X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2})$, and S_X and S_Y are the standard error for $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{a}^T\boldsymbol{Y}$ respectively. Depending on the type of test, the standard error S may take different forms. In the case of two sample t-test with unequal variance, $\boldsymbol{a}^T\boldsymbol{X} = \bar{X}_1 - \bar{X}_2$ and $S_X = \sqrt{S_{X,1}^2/n_1 + S_{X,2}^2/n_2}$, where $S_{X,1}^2$ and $S_{X,2}^2$ are the sample variances for the treatment and the control groups.

Our main goal is to explore the relationship between population correlation (equation (1)) for the test statistics

$$\rho_T = \text{Corr}(T_X, T_Y), \tag{8}$$

and that for their corresponding expression level

$$\rho = \text{Corr}(X, Y). \tag{9}$$

We will examine ???HOW MANY??? different test statistics having the form of equation (7).

3 Results

In this section we present the exact expression of statistics correlation coefficient for two sample t-test. In the first part, we conclude theoretically that test statistics correlation and sample correlation are perfect positive dependent for two sample z-test, but that is not always true for two sample t-test. In the second part, we simulate four different cases where test statistics correlation $r_{\text{statistics}}$ may be very different from true correlation ρ or sample correlation r_{sample}.

3.1 Theory

Theorem 1 *For any given sample size (n_1, n_2) and non zero $\boldsymbol{a}$, $\rho_T = \rho$ if S_X and S_Y are constant with respect to $\boldsymbol{X}, \boldsymbol{Y}$.*

Proof: Since samples are independent, we have

$$\begin{aligned} \text{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) &= \boldsymbol{a}^T\text{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{a} = \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}, \\ \text{Var}(\boldsymbol{a}^T\boldsymbol{X}) &= \boldsymbol{a}^T\text{Var}(\boldsymbol{X})\boldsymbol{a} = \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}, \\ \text{Var}(\boldsymbol{a}^T\boldsymbol{Y}) &= \boldsymbol{a}^T\text{Var}(\boldsymbol{Y})\boldsymbol{a} = \sigma_Y^2\boldsymbol{a}^T\boldsymbol{a} \end{aligned} \tag{10}$$

It follows by equation (1) that

$$\rho_T = \frac{\text{Cov}(T_X, T_Y)}{\sqrt{\text{Var}(T_X)\text{Var}(T_Y)}} = \frac{\text{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})/S_XS_Y}{\sqrt{\text{Var}(\boldsymbol{a}^T\boldsymbol{X})\text{Var}(\boldsymbol{a}^T\boldsymbol{Y})/S_XS_Y}} = \rho \tag{11}$$

Note: Theorem 1 states that test statistics correlation and expression level correlation are equal under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Let $\boldsymbol{a} = (\underbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}_{n_1}, \underbrace{-\frac{1}{n_2}, \ldots, -\frac{1}{n_2}}_{n_2})$, and if we set $S_X = 1$, then T_X corresponds to mean difference between treatment and control group; instead, if $S_X = \sigma_X\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, then T_X corresponds to the statistic for two sample z-test. Therefore, $\rho_T = \rho$ if we use mean difference or z-statistic as test statistic.

Theorem 2 *Let $(X_i, Y_i), i = 1, \ldots, n_1 + n_2$ follow a bivariate normal distribution with mean specified in equation (4) and covariance in (3). If T_X and T_Y are statistics for equal-variance two-sample t-test, that is,*

$$\begin{aligned} &\boldsymbol{a}^T\boldsymbol{X} = \bar{X}_1 - \bar{X}_2 \overset{def}{=} U_X, \quad \boldsymbol{a}^T\boldsymbol{Y} = \bar{Y}_1 - \bar{Y}_2 \overset{def}{=} U_Y \\ &S_i^2 = \frac{(n_1 - 1)S_{i,1}^2 + (n_2 - 1)S_{i,2}^2}{n_1 + n_2 - 2}, \quad i = X, Y. \end{aligned} \tag{12}$$

then

$$Corr(T_X, T_Y) = \frac{\frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} C\rho_s + \rho B + \rho_s \rho (A - B)}{\sqrt{\left[\frac{\Delta_X^2}{\sigma_X^2} C + A\right]\left[\frac{\Delta_Y^2}{\sigma_X^2} C + A\right]}} \tag{13}$$

where

$$\begin{aligned} A &= \frac{n_1 + n_2 - 2}{n_1 + n_2 - 4}, \quad B = \frac{(\frac{n_1+n_2-2}{2})\Gamma^2(\frac{n_1+n_2-4}{2} + \frac{1}{2})}{\Gamma^2(\frac{n_1+n_2-2}{2})}, \\ \rho_s &= Corr(\frac{1}{S_X}, \frac{1}{S_Y}), \quad C = \frac{(n_1 + n_2)(A - B)}{(2 + n_1 n_2^{-1} + n_1 n_2^{-1})}. \end{aligned} \tag{14}$$

The proof of Theorem 2 is presented in Section 4. Next we give the limit of $\mathrm{Corr}(T_X, T_Y)$.

Theorem 3 *If there exists a positive number M, such that $n_1 n_2^{-1} \leq M$ and $n_1 n_2^{-1} \leq M$, then*

$$\rho_T = \lim_{n_1+n_2 \to \infty} Corr(T_X, T_Y) = \frac{\rho + \beta \frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} \rho_s}{\sqrt{\left[1 + \beta \frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta \frac{\Delta_Y^2}{\sigma_Y^2}\right]}} \tag{15}$$

where ρ_s is defined in equation (14) and $\beta = \lim_{n_1+n_2 \to \infty} C$.

When $\mathbf{\Delta} = \mathbf{0}$ then $\rho_T = \rho$; but when $\mathbf{\Delta} \neq \mathbf{0}$, then $\rho_T \neq \rho$ in general.

corollary 1 *If $\mathbf{\Delta} = (\Delta_X, \Delta_Y) = \mathbf{0}$, and there exists a positive number M, such that $n_1 n_2^{-1} \leq M$ and $n_1 n_2^{-1} \leq M$, then $Corr(T_X, T_Y) \to \rho$ as $n_1 + n_2 \to \infty$.*

Proof: If null is true for both test or $\mathbf{\Delta} = 0$, then equation (13) reduces to

$$\mathrm{Corr}(T_X, T_Y) = \left[\rho_s \cdot 1 + (1 - \rho_s)\frac{B}{A}\right] \rho \tag{16}$$

The term in the square bracket is a weighted average of 1 and $\frac{B}{A}$, with the latter converging to 1 as n_1+n_2 grows to infinity. Therefore $\lim_{n_1+n_2 \to \infty} \mathrm{Corr}(T_X, T_Y) = \rho$.

corollary 2 *If $\mathbf{\Delta} = (\Delta_X, \Delta_Y) \neq \mathbf{0}$, then $Corr(T_X, T_Y)$ does not converge to ρ in general.*

The result immediately follows from lemma (4) in appendix.

Depending on the true value of $\mathbf{\Delta}$ (DE or not DE, up-regulated or down-regulated if DE), $\text{Corr}(T_X, T_Y)$ might be far from ρ in different ways, discussed below.

We show via simulation [figure (1)] that for ρ growing from -1 to 1, ρ_s in equation (14) has a "U" shape whose minimum is located near $\rho = 0$, and

$$0 \leq r_s \leq |\rho| \quad \text{ONLY BASED ON SIMULATION} \tag{17}$$

(17) is useful in comparing $\rho(T_1, T_2)$ and ρ. For $\rho < 0$

1. if $\Delta_1\Delta_2 > 0$, then gene 1 and gene 2 are DE in the same direction (both up-regulated or both down-regulated), then

$$\rho(T_1, T_2) = \frac{\rho + \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2} r_{\text{s}}}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \frac{\rho}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho$$

2. if $\Delta_1\Delta_2 < 0$, then gene 1 and gene 2 are DE in different directions (one up-regulated and the other down-regulated), then by $r_s < -\rho$,

$$\rho(T_1, T_2) = \frac{\rho + \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2} r_{\text{s}}}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho\frac{1 - \frac{\Delta_1\Delta_2}{8\sigma_1\sigma_2}}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]\left[1 + \frac{\Delta_2^2}{8\sigma_2^2}\right]}} > \rho$$

3. if $\Delta_1\Delta_2 = 0$, then one is DE but the other is not. Suppose gene 1 is not DE, then

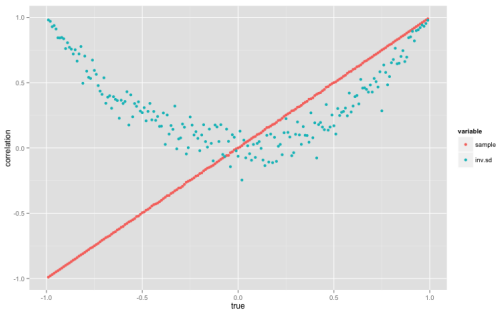
$$\rho(T_1, T_2) = \frac{\rho}{\sqrt{\left[1 + \frac{\Delta_1^2}{8\sigma_1^2}\right]}} > \rho$$

Therefore in any case, $\rho(T_1, T_2) \geq \rho$ when $\rho < 0$. Similarly it can be shown that for $\rho > 0$, $\rho(T_1, T_2) \leq \rho$. In simple words, T_1 and T_2 are "less" correlated than the samples are.

3.2 Simulation

The simulations are performed under two different testing procedures. The first is two sample t-test, where we evaluate the true correlation and statistics correlation. In the second setting, we simulate correlated Poisson data to mimic RNA-Seq counts, and evaluate the relationship between the two for

Figure 1: $\text{Corr}(S_1^{-1}, S_2^{-1})$ against ρ



score test of Poisson regression.

For the normal case, we let

$$\begin{aligned} \boldsymbol{X}_i &= \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ -10 \end{pmatrix}, \begin{pmatrix} 0.1 & \rho\sqrt{0.1 \cdot 0.3} \\ \rho\sqrt{0.1 \cdot 0.3} & 0.3 \end{pmatrix} \right] \\ \boldsymbol{Y}_j &= \begin{pmatrix} Y_{1j} \\ Y_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 10 + \Delta_1 \\ -10 + \Delta_2 \end{pmatrix}, \begin{pmatrix} 0.1 & \rho\sqrt{0.1 \cdot 0.3} \\ \rho\sqrt{0.1 \cdot 0.3} & 0.3 \end{pmatrix} \right] \end{aligned} \tag{18}$$

with ρ growing continuously from -0.99 to 0.99 by 0.01. The sample size n is set to be 1000 ($j = 1, \ldots, 500$ for each group). For each given ρ, we generate $50,000$ samples for control group and another $50,000$ samples for the treatment group. The $50,000$ samples within each group are then randomly split into 100 blocks of size 500. Next, a pair is formed by taking one block (500 samples) from treatment and one block from control, mimicking one experiment for two group comparison. Therefore, 100 pairs are obtained to represent 100 replicates of the same experiment, from which 100 test statistics are computed for each gene.

The sample correlation r_{sample} is calculated by (**??**). The correlations between t-test statistics are calculated by the sample correlation of $(\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_{100})$. Specifically, we compared $r_{\text{statistics}}$, r_{sample} and ρ for the following four cases:

a) no DE genes;

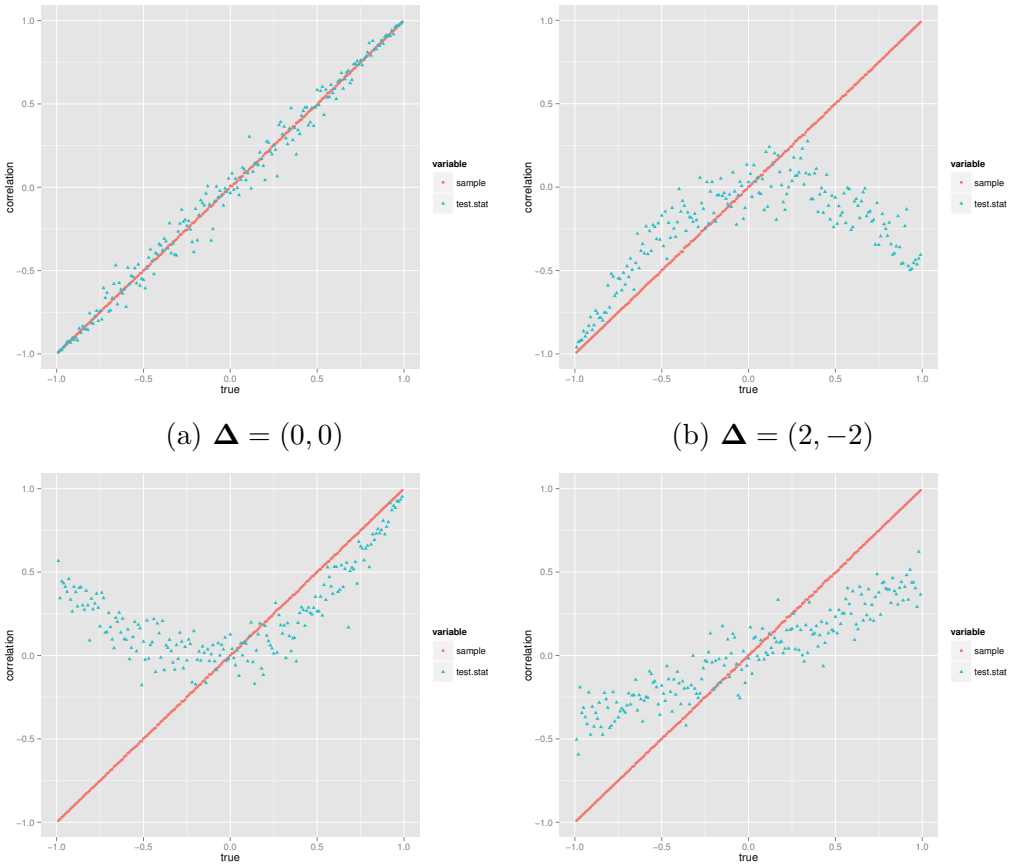
b) DE in opposite directions;

c) DE in the same direction;

d) gene 1 DE and gene 2 null.

Figure (2) plots $r_{\text{statistics}}$ and r_{sample} against ρ. While the equivalence between those three holds when neither gene is DE [case a)], it fails as long as DE

exists. $r_{\text{statistics}}$ is almost always negative, if genes are DE in different direction [case b)], and almost always positive if genes are DE in the same direction [case c)]. When only one gene is DE, $r_{\text{statistics}}$ is positively proportional to ρ [case d)]. Note that in all cases, $|r_{\text{statistics}}| \leq |\rho|$, in other words, the test statistics tend to be "less correlated" than the samples are.



(a) $\boldsymbol{\Delta} = (0, 0)$ (b) $\boldsymbol{\Delta} = (2, -2)$

(c) $\boldsymbol{\Delta} = (2, 2)$ (d) $\boldsymbol{\Delta} = (-2, 0)$

Figure 2: Under t-test, the relationship between r_{sample} (solid dots), $r_{\text{statistics}}$ (triangles) and r_{true} (horizontal axis), for the case (a): gene 1 and gene 2 are not DE; (b): both genes are DE, but in different direction; (c): both genes are DE, in the same direction; (d): gene 1 is DE, but gene 2 is not. $\boldsymbol{\Delta}$ is the magnitude of DE.

Correlated Poisson data was simulated according to [16]. Briefly, a 2-vector standard normal $\boldsymbol{Z}$ is first generated with correlation matrix $\boldsymbol{\Sigma}_Z$, and Z_i's are converted to $U_i = \Phi(Z_i)$ where Φ is standard normal CDF. U_i's,

uniform on (0, 1), are then transformed to $Y_i \equiv F_i^{-1}(U_i)$ with

$$F_i^{-1}(u) = \inf\{y : F_i(y) \geq u\} \tag{19}$$

The element in $\mathbf{\Sigma}_Z$ are chosen such that the desired Pearson correlation can be achieved. Technical details are available in [16] and thus not discussed here.

The two group comparison under Poisson regression are simulated as follows: for control group $X_{1i} \sim \text{Pois}(20)$ and $X_{2i} \sim \text{Pois}(50)$ with $\text{Cov}(X_{1i}, X_{2i}) = \rho$; for treatment group, a shift $\mathbf{\Delta}$ is added to the mean vectors, in other words, $Y_{1i} \sim \text{Pois}(20+\Delta_1)$ and $Y_{2i} \sim \text{Pois}(50+\Delta_2)$ with $\text{Cov}(Y_{1j}, Y_{2j}) = \rho$. The test statistics are calculated from score test (derivation is available in appendix),

$$U = \frac{\sqrt{\frac{n}{2}}(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y}_1 + \bar{y}_2}}. \tag{20}$$

Unlike the normal distribution whose shape is determined by both the mean and variance parameters, the shape of a Poisson distribution is totally determined by its mean parameter. For a score test statistic such as (20), the denominator and the numerator are no longer independent. Subsequently, the derivation of test statistics correlation for t-test is invalid for Poisson regression. We will only demonstrate via simulation the relationship between $r_{\text{statistics}}$, r_{sample} and ρ.

Figure (3) presents the simulation under scenarios a)-d). The equivalence of $r_{\text{statistics}}$, r_{sample} and ρ still holds in general when neither gene is DE.

4 Method

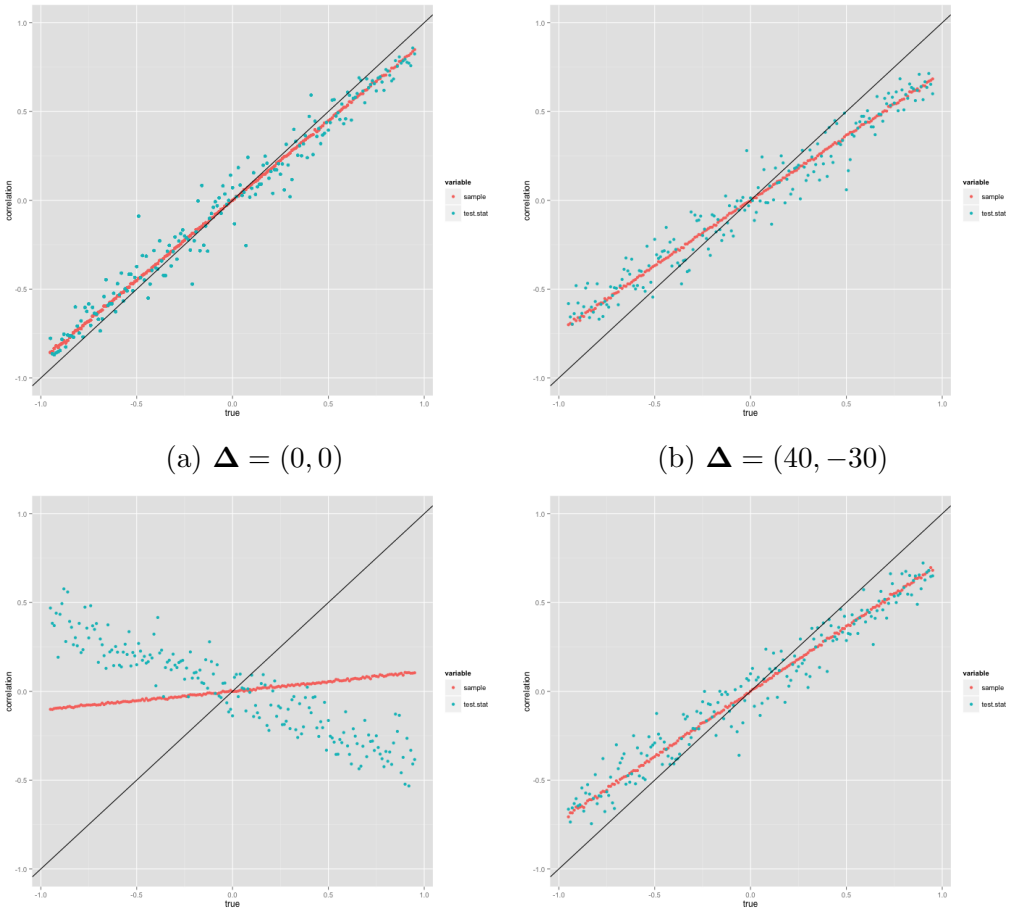
Lemma 1 *Sample correlation coefficient is a consistent estimator for ρ,*

$$\sqrt{n}(r_{sample} - \rho) \xrightarrow{D} N\left(0, (1-\rho^2)^2\right).$$

The proof of lemma 1 can be found in [17].

To prove Theorem 2, it is useful to note that $\boldsymbol{U}$ is independent of $\boldsymbol{S}$, following from Lemmas 2 and 3.

Lemma 2 *Let $(X_j, Y_j), j = 1 \ldots, n$ be independent random variables satisfying equation (5), then $\boldsymbol{W} = (W_X, W_Y) = (\frac{(n-1)S_X^2}{\sigma_X^2}, \frac{(n-1)S_Y^2}{\sigma_Y^2})$ follows a*



(a) $\boldsymbol{\Delta} = (0, 0)$ (b) $\boldsymbol{\Delta} = (40, -30)$

(c) $\boldsymbol{\Delta} = (40, 70)$ (d) $\boldsymbol{\Delta} = (40, 0)$

Figure 3: Under score test of Poisson regression, the relationship between r_{sample} (solid dots), $r_{\text{statistics}}$ (triangles) and r_{true} (horizontal axis), for the case (a): gene 1 and gene 2 are not DE; (b): both genes are DE, but in different direction; (c): both genes are DE, in the same direction; (d): gene 1 is DE, but gene 2 is not.

bivariate chi square distribution *with density*

$$f(w_x, w_y) = \frac{2^{-n}(w_x w_y)^{(n-3)/2} e^{-\frac{w_x + w_y}{2(1-\rho^2)}}}{\sqrt{\pi}\Gamma(\frac{n}{2})(1-\rho^2)^{(n-1)/2}} \times \sum_{k=0}^{\infty} [1 + (-1)^k] \left(\frac{\rho\sqrt{w_x w_y}}{1-\rho^2}\right)^k \frac{\Gamma(\frac{k+1}{2})}{k!\Gamma(\frac{k+n}{2})} \tag{21}$$

for $n > 3$ *and* $-1 < \rho < 1$.

For proof of Lemma 2, interested readers are referred to [18]. It immediately follows from Lemma 2 that $\boldsymbol{W}_1 = (\frac{(n_1-1)S^2_{X,1}}{\sigma^2_X}, \frac{(n_1-1)S^2_{Y,1}}{\sigma^2_Y})$ and $\boldsymbol{W}_2 = (\frac{(n_2-1)S^2_{X,2}}{\sigma^2_X}, \frac{(n_2-1)S^2_{Y,2}}{\sigma^2_Y})$ both follow bivariate chi-square distributions, with degree of freedom $n_1 - 1$ and $n_2 - 1$ respectively. Note that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independent since the samples are independent.

Lemma 3 *$\boldsymbol{U} = (U_X, U_Y)$ is independent of $\boldsymbol{S} = (S_X, S_Y)$, where $\boldsymbol{U}$ and $\boldsymbol{S}$ are defined in equation (12).*

Proof: By Lemma 2, the density function of $\boldsymbol{W}_1+\boldsymbol{W}_2$ only involves $\sigma^2_X, \sigma^2_Y, \rho$ and sample size n_1, n_2, therefore we can denote its density by some function $g(\sigma^2_X, \sigma^2_Y, \rho, n_1 + n_2)$. Note that $\boldsymbol{S}^2 = \frac{(\sigma^2_X,\ \sigma^2_Y)}{n_1+n_2-2}(\boldsymbol{W}_1 + \boldsymbol{W}_2)^T$ is a linear transformation of $\boldsymbol{W}_1 + \boldsymbol{W}_2$, so its density also can be expressed in terms of $\sigma^2_1, \sigma^2_2, \rho, n_1, n_2$. Therefore $\boldsymbol{S} = (S_X, S_Y)$ is an ancillary statistic for $\boldsymbol{\Delta}$. On the other hand, it can be shown that $\boldsymbol{U} = (U_X, U_Y)$ is a complete sufficient statistic for $\boldsymbol{\Delta}$. It follows by Basu's theorem that $\boldsymbol{U}$ and $\boldsymbol{S}$ are independent.

Lemma 3 implies that $U_X U_Y$ is also independent of $\frac{1}{S_X S_Y}$, and therefore $E(\frac{U_X}{S_X} \cdot \frac{U_Y}{S_Y})$ can be expressed as $E(U_X U_Y)E(\frac{1}{S_X S_Y})$. Additionally, if we know $\mathrm{Corr}(\frac{1}{S_X}, \frac{1}{S_Y})$, then the t-test statistics correlation can be accurately represented.

Proof of theorem (2)

First note that

$$\begin{aligned}\mathrm{Cov}(T_X, T_Y) &= E(T_X T_Y) - E(T_X)E(T_Y) \\ &= \frac{1}{c_0^2}\left[E(U_X U_Y)E(\frac{1}{S_X S_Y}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})\right] \qquad \text{(by lemma 3)}\end{aligned}$$

where $c_0 = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and $\mathrm{Var}(T_X) = \mathrm{Var}(\frac{U_X}{c_0 S_X}) = \frac{1}{c_0^2}\mathrm{Var}(\frac{U_X}{S_X})$. Note that

$$\begin{aligned}\mathrm{Corr}(T_X, T_Y) &= \frac{\mathrm{Cov}(T_X, T_Y)}{\sqrt{\mathrm{Var}(T_X)\mathrm{Var}(T_Y)}} \\ &= \frac{E(U_X U_Y)E(\frac{1}{S_X S_Y}) - E(\frac{U_X}{S_X})E(\frac{U_Y}{S_Y})}{\sqrt{\mathrm{Var}(\frac{U_X}{S_X})\mathrm{Var}(\frac{U_Y}{S_Y})}}\end{aligned} \tag{22}$$

We need to calculate $E(U_X U_Y)$, $E(\frac{1}{S_X S_Y})$, $E(\frac{U_i}{S_i})$ and $\mathrm{Var}(\frac{U_i}{S_i})$ for $i = X, Y$.

1. Note that $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1} + \frac{1}{n_2})\right), i = X, Y$.

$$\begin{aligned}E(U_X U_X) &= \mathrm{Cov}(U_X, U_Y) + E(U_X)E(U_Y) \\ &= \rho\sigma_X\sigma_Y\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \Delta_X\Delta_Y\end{aligned} \tag{23}$$

2. Since $\frac{(n_1-1)S_X^2}{\sigma_X^2}$ and $\frac{(n_2-1)S_Y^2}{\sigma_Y^2}$ are independent and follow $\chi^2(n_1-1)$ and $\chi^2(n_2-1)$ respectively, , we have $W_X = \frac{(n_1+n_2-2)S_X^2}{\sigma_X^2} \sim \chi^2(n_1+n_2-2)$. It can be shown that

$$E(W_X^k) = \frac{2^k\Gamma(\frac{n_1+n_2-2}{2}+k)}{\Gamma(\frac{n_1+n_2-2}{2})}$$

Therefore

$$E\left(\frac{1}{S_X}\right) = \frac{\sqrt{B}}{\sigma_X}, \quad \mathrm{Var}\left(\frac{1}{S_X}\right) = \frac{A-B}{\sigma_X^2} \tag{24}$$

Note that $\rho_s = \mathrm{Corr}(\frac{1}{S_X}, \frac{1}{S_Y})$, we have

$$\begin{aligned} E(\frac{1}{S_X S_Y}) &= E(\frac{1}{S_X})E(\frac{1}{S_Y}) + \rho_s\sqrt{\mathrm{Var}(\frac{1}{S_X})\mathrm{Var}(\frac{1}{S_Y})} \\ &= \frac{B}{\sigma_X\sigma_Y} + \rho_s\frac{A-B}{\sigma_X\sigma_Y} \end{aligned} \tag{25}$$

3. $U_i \sim N\left(\Delta_i, \sigma_i^2(\frac{1}{n_1}+\frac{1}{n_2})\right)$ and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2} \sim \chi^2(n_1+n_2-2)$ and by Lemma 3 U_i and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}$ are independent for $i = X, Y$, we have

$$\frac{\frac{U_i-\Delta_i}{\sigma_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}/(n_1+n_2-2)} = \frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{26}$$

It follows from

$$E\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = 0, \quad \mathrm{Var}\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{n_1+n_2-2}{n_1+n_2-4} \tag{27}$$

that

$$\begin{aligned} E\left(\frac{U_i}{S_i}\right) &= \frac{\Delta_i}{\sigma_i}\sqrt{B} \\ \mathrm{Var}\left(\frac{U_i}{S_i}\right) &= A\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{\Delta_i^2}{\sigma_i^2}(A-B) \end{aligned} \tag{28}$$

Finally, the test statistics correlation (13) is obtained by plugging equations (23–28) into equation (22).

Up to now we have obtained an exact expression for $\mathrm{Corr}(T_X, T_Y)$, which depends not only on the sample size n_1 and n_2, but also on Δ/σ, the relative magnitude of DE. The rest of this section discusses asymptotic property of $\mathrm{Corr}(T_X, T_Y)$ for large sample size.

Lemma 4 *If there exists a positive number M, such that $n_1 n_2^{-1} \leq M$ and $n_1 n_2^{-1} \leq M$, then the following results hold:*

1. $\lim_{n_1+n_2\to\infty} A = 1.$
2. $\lim_{n_1+n_2\to\infty} B = 1.$
3. $\lim_{n_1+n_2\to\infty} C = \beta.$

where A, B and C are defined in equation (14), and $\beta = (4 + n_1 n_2^{-1} + n_1^{-1} n_2)^{-1}$.

Proof: Note that

$$B = \begin{cases} \frac{(k-1)\Gamma^2(k-\frac{3}{2})}{\Gamma^2(k-1)}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})\Gamma^2(k-1)}{\Gamma^2(k-\frac{1}{2})}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{29}$$

We will use second order Stirling's formula,

$$k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k (1 + \frac{1}{12k}) \tag{30}$$

Using Stirling's formula (30) and $\Gamma(k+\frac{1}{2}) = \frac{(2k)!}{4^k k!}\sqrt{\pi}$, it can be shown that

$$B \approx \begin{cases} \frac{(k-1)(k-2)(k-2+\frac{1}{24})^2}{(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{(k-\frac{1}{2})(k-1+\frac{1}{12})^4}{(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{31}$$

It can also be shown using equation (31) that

$$A - B \approx \begin{cases} \frac{\frac{1}{4}(k-1)(k-2)^3+o((k-2)^4)}{(k-2)(k-2+\frac{1}{12})^4}, & \text{if } n_1+n_2 = 2k, k \geq 2 \\ \\ \frac{\frac{1}{4}(k-1)^3(k-\frac{1}{2})(k-3)+o((k-1)^4)}{(k-\frac{3}{2})(k-1+\frac{1}{24})^2(k-1)^3}, & \text{if } n_1+n_2 = 2k+1, k \geq 2 \end{cases} \tag{32}$$

And the results immediately follow.

5 Conclusion

State the major findings

This article discusses the relationship between sample correlation coefficients r_{sample} (after treatment effects removed) and test statistics correlation $r_{\text{statistics}}$ in a two group comparison setting. We proved that under normal distribution assumption, $r_{\text{statistics}}$ and r_{sample} have a perfect positive correlation for two sample z test. However, for two sample t-test this correspondence does not hold in general, unless the null in (**??**) is true for all the tests considered. The results for two sample t-test can be applied to the case of two group mean comparison under Poisson regression, as shown by simulation. Consequently, that estimating $r_{\text{statistics}}$ by r_{sample} after nullifying treatment effects can not be taken for granted.

State the practical meaningness of the findings

In gene expression analysis, cares need to be taken when estimating test statistics correlation from sample correlation. For microarray data, two sample t test ([4], [3]) or its moderated version [7] are used in detecting DE, with $r_{\text{statistics}}$ estimated from sample correlation to adjust for inter-gene correlation. Our study shows, however, that for DE genes, $r_{\text{statistics}}$ may be either overestimated if two genes are positively correlated, or underestimated if two genes are negatively correlated. If we believe that most genes are positively correlated (if any) and that there are true DE genes, then the VIF factor may be overestimated in [7], which may result in conservative test for enrichment analysis; the variance of $r_{\text{statistics}}$ may also be overestimated in [4], which leads to larger variation in estimating conditional FDP. The situation may be more complicated for RNA-Seq data, which are counts in nature and therefore need to be modeled by more sophisticated regression tools (e.g. logistic regression, negative binomial regression, etc.).

Acknowledge the study's limitations

One assumption yet to be justified

In the context of two sample t-test, the simulation results agree with our theoretical conclusion, assuming that $0 \leq r_s \leq |\rho|$ in (17) is true. Our simulation does suggest

$$r_s = \rho^2, \tag{33}$$

as shown in figure (1). If (33) can be justified theoretically, it is possible to approximate the true value of $\rho(T_1, T_2)$, which will correct the bias of estimating $r_{\text{statistics}}$ by r_{sample}. Another remaining challenge is to assess the relationship of $r_{\text{statistics}}$ and ρ for non-normal distributions, or for other hypothesis testing under different regression models (e.g., negative binomial regression).

6 Appendix

Score test statistics correlation under Poisson regression
For a gene, let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the gene expression level, and $X = (1, \ldots, 1, 0, \ldots, 0)$ be the indicator of whether sample is from treatment or control group. A Poisson regression model

$$\begin{aligned} Y_i &\sim Pois(\mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

From the likelihood function

$$L = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}$$

we obtain the log-likelihood function

$$\begin{aligned} l(\beta_0, \beta_1) = \log L &= \sum_{i=1}^{n} (y_i \log \mu_i - \log y_i! - \mu_i) \\ &= \sum y_i(\beta_0 + \beta_1 x_i) - \sum \log y_i! - \sum \exp(\beta_0 + \beta_1 x_i) \end{aligned} \tag{34}$$

For testing $H_0 : \beta_1 = 0$, the score test statistics is

$$U = [Z(\tilde{\beta})^T I^{-1}(\tilde{\beta}) Z(\tilde{\beta})]^{1/2}$$

where $\tilde{\beta} = (\hat{\beta}_0, 0)$. From (34) we have

$$\frac{\partial l}{\partial \beta_0} = \sum_i y_i - \sum_i \exp(\beta_0) \Rightarrow \quad \hat{\beta}_0 = \log(\bar{y})$$

Therefore

$$Z(\tilde{\beta}) = \begin{bmatrix} \sum_i y_i - \sum_i \exp(\beta_0 + \beta_1 x_i) \\ \sum_i y_i x_i - \sum_i \exp(\beta_0 + \beta_1 x_i) x_i \end{bmatrix} |_{\beta_1 = 0} = \begin{bmatrix} \sum y_i - \exp(\hat{\beta}_0) \\ \sum y_i x_i - \sum \exp(\hat{\beta}_0) x_i \end{bmatrix} = \begin{bmatrix} 0 \\ \sum y_i x_i - \bar{y} \sum x \end{bmatrix}$$

$$I(\tilde{\beta}) = \begin{bmatrix} \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i \\ \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i & \sum \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i^2 \end{bmatrix} = \begin{bmatrix} \sum y_i & \bar{y} \sum x_i \\ \bar{y} \sum x_i & \bar{y} \sum x_i^2 \end{bmatrix}$$

and it follows that

$$U = [Z(\tilde{\beta})^T I^{-1}(\tilde{\beta}) Z(\tilde{\beta})]^{1/2} = \left(\frac{n(\sum y_i x_i - \bar{y} \sum x_i)^2}{\bar{y}[n \sum x_i^2 - (\sum x_i)^2]} \right)^{1/2} \tag{35}$$

To simplify the above expression, let's assume the first $n/2$ elements of $\boldsymbol{X}$ are 1, therefore we have $\sum x_i = \sum x_i^2 = n/2$

$$U = \sqrt{\frac{n(\sum_{i=1}^{n/2} y_i - \bar{y} \cdot n/2)^2}{\bar{y}[n \cdot n/2 - (n/2)^2]}} = \sqrt{\frac{\frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2}{\bar{y}_1 + \bar{y}_2}} = \pm \frac{\sqrt{\frac{n}{2}}(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y}_1 + \bar{y}_2}} \tag{36}$$

where $\bar{y}_1 = \frac{\sum_{i=1}^{n/2} y_i}{n/2}$ and $\bar{y}_2 = \frac{\sum_{i=n/2+1}^{n} y_i}{n/2}$ are just group means. It resembles a t test statistic.

References

[1] Efron, B. (2004) *Journal of the American Statistical Association.*

[2] Qiu, X., Brooks, A. I., Klebanov, L., and Yakovlev, A. (2005) *BMC bioinformatics* **6(1)**, 120.

[3] Barry, W. T., Nobel, A. B., and Wright, F. A. (2008) *The Annals of Applied Statistics* pp. 286–315.

[4] Efron, B. (2007) *Journal of the American Statistical Association* **102(477).**

[5] **Storey, J. D. (2003)** ***Annals of statistics*** **pp. 2013–2035.**

[6] **Huang, Y.-T. and Lin, X. (2013)** ***BMC Bioinformatics*** **14(1), 210.**

[7] **Wu, D. and Smyth, G. K. (2012)** ***Nucleic acids research*** **40(17), e133–e133.**

[8] **Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010)** ***BMC genomics*** **11(1), 574.**

[9] **Benjamini, Y. and Hochberg, Y. (1995)** ***Journal of the Royal Statistical Society. Series B (Methodological)*** **pp. 289–300.**

[10] **Benjamini, Y. and Yekutieli, D. (2001)** ***Annals of statistics*** **pp. 1165–1188.**

[11] **Goeman, J. J. and Bühlmann, P. (2007)** ***Bioinformatics*** **23(8), 980–987.**

[12] **Yaari, G., Bolen, C. R., Thakar, J., and Kleinstein, S. H. (2013)** ***Nucleic Acids Research*** **p. gkt660.**

[13] **Efron, B. (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, volume 1, Cambridge University Press, .**

[14] **Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005)** ***Proceedings of the National Academy of Sciences of the United States of America*** **102(43), 15545–15550.**

[15] **Lee Rodgers, J. and Nicewander, W. A. (1988) *The American Statistician* 42(1), 59–66.**

[16] **Madsen, L. and Birkes, D. (2013) *Journal of Statistical Computation and Simulation* 83(4), 677–691.**

[17] **Fisher, R. A. (1915) *Biometrika* pp. 507–521.**

[18] **Joarder, A. H. (2009) *Statistical Papers* 50(3), 581–592.**