Test-statistic correlation and the correlation of underlying observed data

Bin Zhuo[a], Duo Jiang[a], Yanming Di[a,*]

[a] *Department of Statistics, Oregon State University, Corvallis, OR, USA*

Abstract

We investigate the relationship between correlation among test statistics and the correlation of underlying observed data. In false discovery rate (FDR) control procedures and gene-set enrichment analysis, the sample correlation of observed data are often used to approximate the test-statistic correlation. We show that, however, such an approximation is only valid under limited settings. In particular, we derive a formula for the correlation between test statistics when they take a specific form. As a special case, we present the exact expression of test-statistic correlation for equal-variance two-sample t-test statistic under bivariate normal assumption.

Keywords: test statistics correlation, sample correlation

2010 MSC: 00-01, 99-00

1. Introduction

Between-gene correlations are commonly observed in gene expression data [1, 2, 3, 4, 5]. One key task of expression analysis is to detect differentially expressed genes whose expression levels are associated with experimental or environmental variables under study. In such a task, a test statistic is calculated for each gene to quantify the magnitude of differential expression (DE). The test statistics may come from two-sample comparison or regression models for more complex designs. Between a pair of genes, when the observed expression levels

*Corresponding author

Email address: diy@oregonstate.edu (Yanming Di)

are correlated, the test statistics calculated from the observed data will also be correlated [6, 7, 8]. This paper concerns the relationship between test-statistic correlation and the correlation of underlying observed data (e.g., expression levels).

The dependency among test statistics has brought methodological issues to multiple hypothesis testing procedures and gene-set analysis. Multiple hypothesis testing procedures determine a p-value cutoff by controlling *false discovery rate* (FDR) [9] or *q-value* [5]. Many FDR-control procedures are valid only when test statistics are independent [9] or have positive regression dependency [10]. Efron [7] 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. In a gene-set analysis, one tests for over-abundance of differentially expressed genes in a specified gene set (e.g., molecular pathways or gene ontologies) [11]. The correlation among DE test statistics, if not addressed appropriately, will undermine the validity of the gene-set test [2, 8].

A number of attempts have been made to account for test-statistic correlation in FDR control and gene-set analysis. Without replicating the experiment, we cannot directly estimate test-statistic correlation between genes since only a single test statistic value is available for each gene. For this reason, the sample correlation between observed data (after gene treatment effects accounted for) is often used as a surrogate. Efron [7] estimated a quantity called *dispersion variate* from the distribution of correlation among observed data, and then calculated the *false discovery proportion* (FDP) conditioning on this dispersion variate. Wu and Smyth [8] estimated a *variance inflation factor* (VIF) from the sample correlation of observed data and incorporated it into their parametric/rank-based gene-set test procedures. The same VIF was also used by Yaari et al. [12] in their gene-set test that quantifies enrichment status of a gene set by a probability density function.

It is yet unclear when and to what extent the test-statistic correlations can be approximated by sample correlations of observed data. Barry et al. [6] showed by Monte Carlo simulation of gene expression data that a nearly linear

relationship holds between test-statistic correlation and sample correlation of observed data for several forms of test statistics they examined. This Monte Carlo simulation results were cited by Wu and Smyth [8] and Yaari et al. [12] as a justification for estimating their VIF from observed data. Efron [7] also concluded through simulation that the distribution of z-value (the test statistic in that paper) correlation can be nearly represented by the distribution of sample correlation from observed data.

Some gene-set analysis methods based on sample-permutation implicitly assume that the joint distribution of test statistics will not change under sample permutation. Our results will reveal that such assumptions are invalid. The key issue is, as also pointed out by Efron [1], that sample permutation method has an extra assumption (called "subset pivotality"), which states that the test statistics always follow the distribution they have under complete null that no gene is differentially expressed. We will show that that the test-statistic correlation depends on correlation between gene expression levels and also on DE status. Permutation of biological samples will change the DE status and thus alternate the correlations among (and thus the joint distribution of) the test statistics. As a consequence, in the presence of differentially expressed genes, the GSEA [13] procedure will not provide a valid null distribution for gene-set enrichment test. Zhuo and Jiang [14] showed that several gene-set tests will give inflated or overly conservative type I errors in the presence of DE test correlations.

We investigate the relationship between the correlation of test statistics and the correlation of underlying observed data. First, we present a formula for correlation between test statistics when they take a specific form and meet some assumption of independence. Then, under bivariate normal setting, we apply this formula to a special case where two-group comparison experiment is considered. We show that 1) the test-statistic correlation is equal to the correlation of underlying observed data when the test statistics are a linear combination of the observed data, and that 2) the test-statistic correlation is generally weaker than the correlation of underlying observed data when the test

statistics are derived from two-sample t-test under normal assumption. We use simulation results to illustrate our findings.

2. General setup

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 a number of 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. Throughout this paper, we will discuss Pearson's correlation under bivariate settings. We will restrict our interest to two types (following the notation of Lee Rodgers and Nicewander [15]) of Pearson's correlation coefficient. The first type of correlation, which we refer to as *the correlation of underlying observed data* or *the population correlation*, is the standardized covariance

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

In equation (1), μ_X and μ_Y are the expected values of random variables X and Y, and $\sigma_X < \infty$ and $\sigma_Y < \infty$ are the population standard errors. The second type of correlation, which we refer to as *sample correlation*, is a function of raw scores and means

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

where $(\bar{x}, \bar{y})$ is the vector of arithmetic mean of the observations. Fisher [16] proved that sample correlation r is a consistent estimator for the underlying correlation ρ.

Let (X_j, Y_j) be a bivariate random variable representing two features (e.g, genes) of sample $j = 1, \ldots, m$, and (x_j, y_j) the corresponding realization. We assume that the population mean of (X_j, Y_j) may differ across samples, but that the population covariance structure remains the same, that is,

$$E \begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \mu_{X,j} \\ \mu_{Y,j} \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\mu}_j, \quad \text{for } j = 1, \ldots, m \tag{3}$$

and

$$\text{Cov} \begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \stackrel{\text{def}}{=} \boldsymbol{\Sigma} \tag{4}$$

where ρ is the population correlation defined by equation (1). In addition, we assume independence across samples (note that independence implies no 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 DE—whether the expression level of a gene is significantly correlated with the treatment or experimental variables. Let $\boldsymbol{a} := (a_1, \ldots, a_m)^T$ be a vector for a contrast of interest, then DE detection for gene X can be statistically formulated as

$$H_0 : \boldsymbol{a}^T\boldsymbol{\mu}_X = 0 \ \textit{Versus}\ H_1 : \boldsymbol{a}^T\boldsymbol{\mu}_X \neq 0, \tag{6}$$

where $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ and $\boldsymbol{\mu}_X = (\mu_{X,1}, \ldots, \mu_{X,m})^T$. DE detection for gene Y can be obtained by applying the same contrast to $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ (simply replacing the subscript X by Y in equation (6)). This hypothesis testing procedure usually results in a "t-test similar" test statistic, in which the numerator is a linear combination of $\boldsymbol{X}$ and the denominator is its standard error. 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 S_X and S_Y are the standard errors for $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{a}^T\boldsymbol{Y}$ respectively. Our main goal is to explore the relationship test-statistic correlation (equation (1))

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

and the underlying correlation of corresponding observed data

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

We will examine under what condition and to what extent $\rho_T(m)$ converges to ρ.

3. Results

In this section we present the exact formula of test-statistics correlation $\rho_T(m)$ by making some assumptions about T_X and T_Y, and show that the test-statistic correlation ρ_T does not always equal or converge to the population correlation ρ. For the case of two-group comparison, we prove that 1) if T_X (or T_Y) is a linear transformation of $\boldsymbol{X}$ (or $\boldsymbol{Y}$), then $\rho_T(m) = \rho$, and that 2) if T_X (or T_Y) is the two sample t-test statistic for $\boldsymbol{X}$ (or $\boldsymbol{Y}$) under normal assumption, then $|\lim_{m\to\infty} \rho_T(m)| \leq |\rho|$. For 2), we show that the relationship between $\lim_{m\to\infty} \rho_T(m)$ and ρ depends on whether the hypothesis tests (equation (6)) are true null or not. We perform simulations for the case of test statistics derived from two-sample t-test to illustrate our findings.

3.1. Theory

Theorem 1. *Let $(X_j, Y_j), j = 1, \ldots, m$ be independent random vectors with mean and covariance structures specified in equation (3). If $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ is independent of (S_X, S_Y), then the correlation of T_X and T_Y in equation (7) can be expressed as*

$$\rho_T(m) = \frac{\rho E(S_X^{-1}S_Y^{-1}) + \frac{\boldsymbol{a}^T\boldsymbol{\mu}_X \cdot \boldsymbol{a}^T\boldsymbol{\mu}_Y}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}} Cov(S_X^{-1}, S_Y^{-1})}{\sqrt{\left[E(S_X^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_X^{-1})\right]\left[E(S_Y^{-2}) + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}} Var(S_Y^{-1})\right]}} \tag{10}$$

Proof: Since samples are independent, we have

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y}) &= \boldsymbol{a}^T\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{a} = \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a},\\
\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{X}) &= \sigma_X^2\boldsymbol{a}^T\boldsymbol{a},\\
E(\boldsymbol{a}^T\boldsymbol{X})^2 &= (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2 + \sigma_X^2\boldsymbol{a}^T\boldsymbol{a},\\
E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})] &= E(\boldsymbol{a}^T\boldsymbol{X})E(\boldsymbol{a}^T\boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})\\
&= (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}
\end{aligned} \tag{11}$$

Note that since $\boldsymbol{a}^T\boldsymbol{X}$ is independent of S_X, we have

$$\begin{aligned}
\mathrm{Var}(T_X) &= E\left[\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)^2\right] - \left[E\left(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}\right)\right]^2\\
&= E[\boldsymbol{a}^T\boldsymbol{X}]^2E[S_X^{-2}] - \left[E(\boldsymbol{a}^T\boldsymbol{X})\right]^2\left[E(S_X^{-1})\right]^2\\
&= \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}E(S_X^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_X)^2\mathrm{Var}(S_X^{-1})
\end{aligned} \tag{12}$$

Similarly,

$$\mathrm{Var}(T_Y) = \sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}E(S_Y^{-2}) + (\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2\mathrm{Var}(S_Y^{-1}) \tag{13}$$

and

$$\begin{aligned}
&\mathrm{Cov}(T_X, T_Y) = E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}}\cdot\frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right] - E\left[\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X^{-1}}\right]E\left[\frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y^{-1}}\right]\\
&= E[(\boldsymbol{a}^T\boldsymbol{X})(\boldsymbol{a}^T\boldsymbol{Y})]\cdot E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)E[S_X^{-1}]E[S_Y^{-1}]\\
&= [(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y) + \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}]E[S_X^{-1}S_Y^{-1}] - (\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)E[S_X^{-1}]E[S_Y^{-1}]
\end{aligned} \tag{14}$$

The result follows by plugging equations (11)—(14) into equation (1).

corollary 1. *For any non zero* $\boldsymbol{a}$, $\rho_T(m) = \rho$ *if* S_X *is constant with respect to* $\boldsymbol{X}$ *and* S_Y *to* $\boldsymbol{Y}$.

Proof: When S_X and S_Y are constants, $\mathrm{Cov}(S_X^{-1}, S_Y^{-1})$, $\mathrm{Var}(S_X^{-1})$ and $\mathrm{Var}(S_Y^{-1})$ are all 0, and equation (10) reduces to

$$\rho_T(m) = \frac{\rho E(S_X^{-1}S_Y^{-1})}{\sqrt{E(S_X^{-2})E(S_Y^{-2})}} = \rho. \tag{15}$$

Corollary 1 states that test-statistic correlation and the correlation of underlying observed data are equal under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

However, if we assume that (S_X, S_Y) is a non-constant function of $(\boldsymbol{X}, \boldsymbol{Y})$, then the test-statistic correlation in equation (10) can be expressed as

$$\rho_T(m) = \frac{\frac{E(S_X^{-1}S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}\rho + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)(\boldsymbol{a}^T\boldsymbol{\mu}_Y)}{\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}}\rho_s}{\sqrt{\left[\frac{E(S_X^{-2})}{\mathrm{Var}(S_X^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_X)^2}{\sigma_X^2\boldsymbol{a}^T\boldsymbol{a}}\right]\left[\frac{E(S_Y^{-2})}{\mathrm{Var}(S_Y^{-1})} + \frac{(\boldsymbol{a}^T\boldsymbol{\mu}_Y)^2}{\sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}}\right]}}, \tag{16}$$

where

$$\rho_s = \frac{\mathrm{Cov}(S_X^{-1}, S_Y^{-1})}{\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})}}. \tag{17}$$

The correlation between test statistics $\rho_T(m)$ depends on the form of test statistics, and in general, may not converge to the population correlation ρ.

3.2. Application of Theorem 1 under normal distribution

Many gene expression experiments are conducted to compare expression levels under two-treatment conditions. For the rest of this section, we discuss the relationship between ρ_T and ρ under such setting. Let $n = n_1 + n_2$ be the total number of samples, where n_1 of them are from group 1 and n_2 from group 2, and let

$$\boldsymbol{a} = (\underbrace{\frac{1}{n_1}, \ldots, \frac{1}{n_1}}_{n_1}, \underbrace{-\frac{1}{n_2}, \ldots, -\frac{1}{n_2}}_{n_2})^T \tag{18}$$

be the contrast of interest. The mean expression levels are specified as

$$\begin{aligned} &\boldsymbol{\mu}_j = (\mu_X, \mu_Y)^T, \quad j = 1, \ldots, n_1, \\ &\boldsymbol{\mu}_j = (\mu_X, \mu_Y)^T + (\Delta_X, \Delta_Y)^T, \quad j = n_1 + 1, \ldots, n_1 + n_2. \end{aligned} \tag{19}$$

If we set $S_X = 1$, then T_X corresponds to mean difference between groups 1 and 2; instead, if $S_X = \sigma_X\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where σ_X is known, then T_X corresponds to the statistic for two-sample z-test with pooled variance. Therefore, according to Corollary 1, $\rho_T(n_1, n_2) = \rho$ if we use mean difference or z-value as test statistics.

The two-sample t-statistic is also a commonly used statistic in DE analysis. In the case of two sample t-test with equal variance, with the contrast $\boldsymbol{a}$ defined in equation (18), the test statistic for X is

$$T_X = \frac{\bar{X}_1 - \bar{X}_2}{S_{p,X}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{20}$$

where $S_{p,X}$ is the pooled variance

$$S_{p,X}^2 = \frac{(n_1-1)S_{X,1}^2 + (n_2-1)S_{X,2}^2}{n_1+n_2-2}. \tag{21}$$

Similarly, we obtain T_Y by replacing the subscript "X" with "Y" in equations (20) and (21). Under normal distribution assumption, we have the following theorem for two-sample t-test with equal variance:

Theorem 2. *Let $(X_i, Y_i), i = 1, \ldots, n$ follow a bivariate normal distribution with mean specified by equations (19) and covariance $\mathbf{\Sigma}$ (see equation (3)). If T_X and T_Y are statistics for equal-variance two-sample t-test, then*

$$\rho_T(n_1, n_2) = \frac{\frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} C\rho_s + \rho B + \rho_s \rho (A - B)}{\sqrt{\left[\frac{\Delta_X^2}{\sigma_X^2} C + A\right]\left[\frac{\Delta_Y^2}{\sigma_X^2} C + A\right]}} \tag{22}$$

where

$$\begin{aligned} A &= \frac{n_1+n_2-2}{n_1+n_2-4}, \quad B = \frac{(\frac{n_1+n_2-2}{2})\Gamma^2(\frac{n_1+n_2-4}{2} + \frac{1}{2})}{\Gamma^2(\frac{n_1+n_2-2}{2})}, \\ \rho_s &= Corr(S_X^{-1}, S_Y^{-1}), \quad C = \frac{(n_1+n_2)(A-B)}{(2 + n_1 n_2^{-1} + n_1 n_2^{-1})}. \end{aligned} \tag{23}$$

The proof of Theorem 2 is presented in Appendix A. Next we present the limit of $\rho_T(n_1, n_2)$.

Theorem 3. *If there exists positive constants M_1 and M_2, such that $M_1 \leq n_1 n_2^{-1} \leq M_2$, then*

$$\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) = \frac{\rho(1 + \beta \frac{\Delta_X \Delta_Y}{\sigma_X \sigma_Y} \rho)}{\sqrt{\left[1 + \beta \frac{\Delta_X^2}{\sigma_X^2}\right]\left[1 + \beta \frac{\Delta_Y^2}{\sigma_Y^2}\right]}} \tag{24}$$

where $\beta = \lim_{n_1+n_2\to\infty} C = (4 + 2n_1^{-1}n_2 + 2n_1 n_2^{-1})^{-1}$.

The proof is provided in Appendix B. Theorem 3 states that as long as n_1 and n_2 grow proportionally to infinity, the quantity $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)$ is a function of population correlation ρ, the signal-to-noise ratios $(\delta_X, \delta_Y) = (\Delta_X/\sigma_X, \Delta_Y/\sigma_Y)$ and the sample ratio n_1/n_2. We have the following observations:

1. If both tests are true null (i.e., $\mathbf{\Delta} = \mathbf{0}$), then $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) = \rho$.

2. If only one test is true null, then $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)$ is proportional to and smaller in absolute value than ρ (i.e., $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) = \gamma_0\rho,\ 0 < \gamma_0 < 1$).

3. If both tests are true alternative (i.e., $\mathbf{\Delta} \neq \mathbf{0}$), then $\rho_T \neq \rho$ in general. Specifically,

   i) when $\Delta_X\Delta_Y > 0$ (i.e., both genes are DE towards the same direction), we have $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) > \rho$ for $\rho < 0$ and $0 \leq \lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) \leq \rho$ for $\rho \geq 0$.

   ii) when $\Delta_X\Delta_Y < 0$ (i.e., genes are DE towards different directions), we have $\rho < \lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) < 0$ for $\rho < 0$ and $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) < \rho$ for $\rho > 0$.

   Therefore in either case, we have $|\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)| \leq |\rho|$.

We note that $|\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)| \leq |\rho|$ when test statistics are derived from two sample t test with equal variance. In other words, the correlation between T_X and T_Y are always weaker than that between X and Y. It's also interesting to note that when both genes are differentially expressed, $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) = 0$ at $\rho = -\frac{\sigma_X\sigma_Y}{\beta\Delta_X\Delta_Y}$ provided that $\frac{\sigma_X\sigma_Y}{\beta\Delta_X\Delta_Y} \in (-1, 1)$. Figure 1 shows the contour plots of $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)$ versus the signal-to-noise ratios δ_X $(= \Delta_X/\sigma_X)$ and δ_Y $(= \Delta_Y/\sigma_Y)$ for different ρ's. The largest value of $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2)$ (in absolute value) is always at the center, where both δ_X and δ_Y are 0 (i.e., $\Delta_X = \Delta_Y = 0$).

In addition, if $n_1/n_2 \to 0$ or ∞, then $\beta = 0$ and we have $\lim_{n_1+n_2\to\infty} \rho_T(n_1, n_2) = \rho$. That is, when sample size of one group is not proportional to that of the other, $\rho_T(n_1, n_2)$ will converge to ρ regardless of whether the tests are under the null or not.

3.3. Simulation

We perform simulations to evaluate the correlations between test statistics and those between expression levels under two sample t-test. We simulate the expression data from normal distributions. Specifically, we let (X, Y) be the

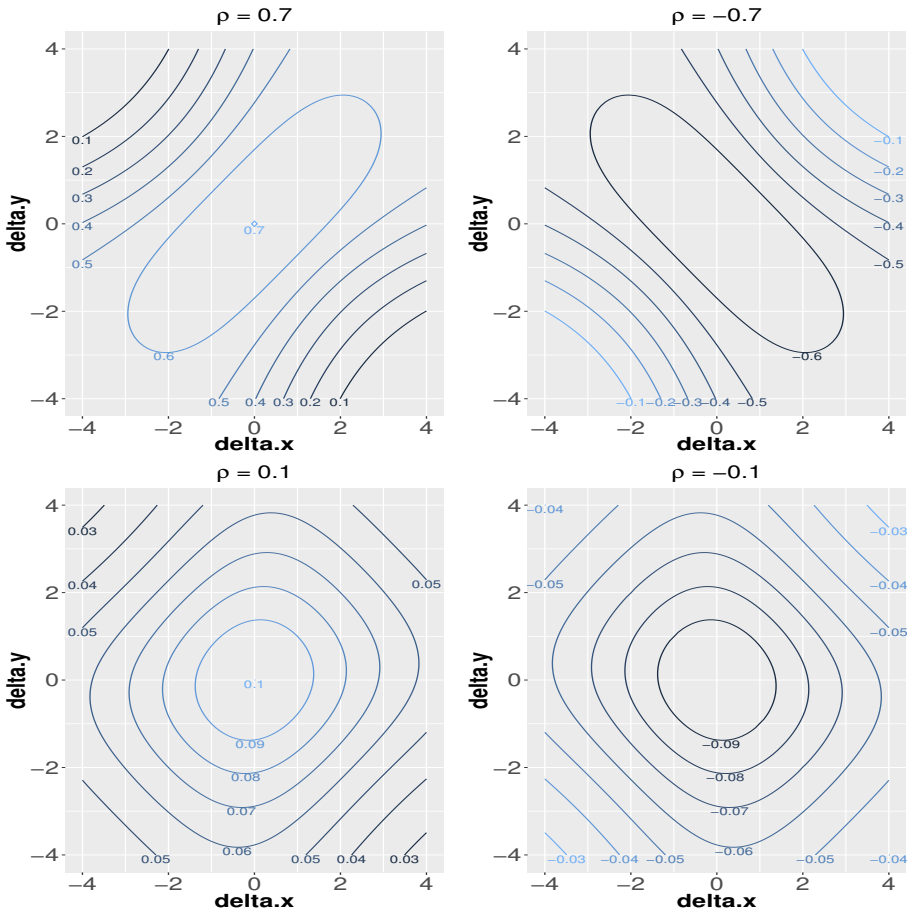

Figure 1: Contour plot of theoretical correlation between test statistics. For each fixed ρ and each pair of δ_X ($= \Delta_X/\sigma_X$) and δ_Y ($= \Delta_Y/\sigma_Y$), the theoretical correlation ρ_T is calculated according to equation (24).

expression levels of genes X and Y, and

$$\begin{aligned}
\begin{pmatrix} X_j \\ Y_j \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], j = 1, \ldots, n_1 \\
\begin{pmatrix} X_j \\ Y_j \end{pmatrix} &\sim N\left[\begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], j = n_1 + 1, \ldots, n_1 + n_2
\end{aligned} \tag{25}$$

For each given ρ, we consider these $n = n_1 + n_2$ pairs of (X, Y) as observations from one *simulated* experiment. Out of this experiment, we calculate

$q = (T_X, T_Y)$ where T_X and T_Y are the test statistics for gene X and gene Y respectively using two-sample t-test for equal variance procedure. We replicate the simulated experiment for $B = 1000$ times, resulting in a matrix $\boldsymbol{Q}_{1000\times 2}$. We take the correlation between the first and the second columns of $\boldsymbol{Q}$ as an estimate for test-statistics correlation $r_{\text{statistics}}$.

We increase ρ from -0.99 to 0.99 by fixed step size 0.01, and examine the relationship between $r_{\text{statistics}}$ and ρ under the following different cases:

a) $\delta_X = \delta_Y = 0$;

b) $\delta_X = 0, \delta_Y = 2$;

c) $\delta_X = 0.5, \delta_Y = 2$;

d) $\delta_X = 1, \delta_Y = 2$;

e) $\delta_X = 3, \delta_Y = 2$;

f) $\delta_X = -3, \delta_Y = 2$.

We conduct simulations for two different sample sizes: we set $n_1 = n_2 = 1000$ to assess asymptotic performance of $\rho_T(n_1, n_2)$ in equation (10), and set $n_1 = n_2 = 3$ to mimic small sample size scenarios which are typical in gene expression study.

In Figure 2, we plot $r_{\text{statistics}}$ against the underlying true population correlation ρ under both large and small sample size scenarios. In case a) where both tests are true null, $r_{\text{statistics}}$ is close to the true correlation ρ when sample size is large ($n_1 = n_2 = 1000$), but smaller (in absolute value) than ρ when sample size is small ($n_1 = n_2 = 3$). In cases b)—f) where there is at least one true alternative, the estimate $r_{\text{statistics}}$ can be very different from ρ. In case b) where only one gene is differentially expressed, the magnitude of $r_{\text{statistics}}$ is proportional to, and smaller in absolute value than ρ. It is more interesting to note that $r_{\text{statistics}}$ is not monotone with respect to ρ when both genes are differentially expressed. If genes are differentially expressed towards the same direction as in the case of e), $r_{\text{statistics}}$ first decreases until it reaches the minimum (a negative value), and then gradually increases to 1, as ρ grows from -1 to 1. When genes are differentially expressed towards opposite directions like the case f), however,

the trend is reversed from that of e): $r_{\text{statistics}}$ increases from -1 to its maximum (a positive value), and then decreases. This set of simulation results is reflected in the test statistics correlation formula of equation (24). We also demonstrate the process of how ρ_T changes from being a linear function of ρ to a quadratic function, by fixing $\delta_Y = 2$ while increasing δ_X from 0 (case b) to 3 (case e).

We illustrate in Figure 2 the variation in $r_{\text{statistics}}$ with respect to change in sample size n. For each fixed ρ under cases a)—f), the absolute value of $r_{\text{statistics}}$ increases when we change the sample size n from 6 to 2000. The change in $r_{\text{statistics}}$ induced by sample size could be substantial, especially when the population correlation is large (e.g., $\rho > 0.2$). This simulation shows that test-statistics correlation ρ_T can be over-estimated by sample correlation, especially when sample size is small.

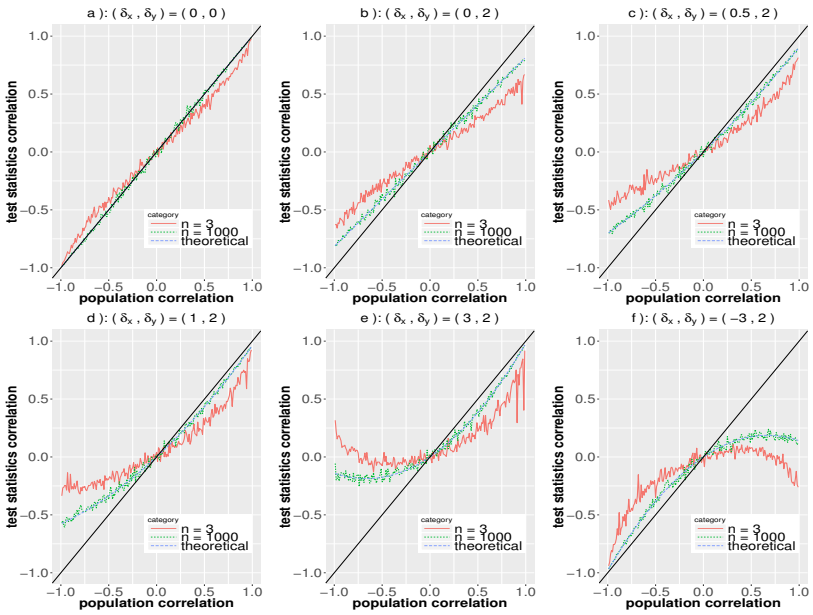

Figure 2: Plots of test-statistic correlation against true population correlation. The test statistics are calculated using two sample t-test with equal variance, and the theoretical correlation is calculated by equation (24).

4. Conclusion and discussion

This article discusses the relationship between population correlation ρ and the corresponding test-statistic correlation $\rho_T(m)$. We investigate $\rho_T(m)$ for test statistics of the form $(\frac{\boldsymbol{a}^T\boldsymbol{X}}{S_X}, \frac{\boldsymbol{a}^T\boldsymbol{Y}}{S_Y})$ (see equation (7)), where the denominator is the standard error of the numerator. Assuming independence between $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ and (S_X, S_Y), we derive the formula for test-statistic correlation ρ_T, and show that ρ_T may not equal population correlation ρ.

In two group comparison setting, we conclude that $\rho_T(m) = \rho$ when S_X (or S_Y) is constant with respect to $\boldsymbol{X}$ (or $\boldsymbol{Y}$). That is, $\rho_T(m) = \rho$ under linear transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is the case for two-sample z-test. However, when S_X (or S_Y) is a function of $\boldsymbol{X}$ (or $\boldsymbol{Y}$), as is the case of two-sample t-test, this equality may not hold. For two-sample t-test, we prove that $\lim_{m\to\infty} \rho_T(m) = \rho$ only if the null in equation (6) is true for all the tests considered, and that $|\lim_{m\to\infty} \rho_T(m)| \leq |\rho|$ otherwise. In the case where one test is true null and the other true alternative, $\lim_{m\to\infty} \rho_T(m)$ is directly proportional to ρ, while when both tests are true alternatives, $\lim_{m\to\infty} \rho_T(m)$ is a quadratic function of ρ.

We note that cares need to be taken when estimating correlations between test statistics. In gene expression analysis, the two-sample t-test [6, 7, 18] or moderated t-test [8] are used to calculate test statistics for DE detection, and the sample correlation (after treatment effects nullified) are used to account for correlation between those test statistics. Our study shows that, however, for differentially expressed genes, $\rho_T(m)$ may be overestimated when two genes are positively correlated, and underestimated when they are negatively correlated. If there are true differentially expressed genes whose expression levels are correlated in either way, the VIF may not be accurately estimated in Wu and Smyth [8], resulting in biased test for their enrichment analysis as shown in from Zhuo and Jiang [14]. Our results also indicate that the variance of $\rho_T(m)$ may also be overestimated in Efron [7], which leads to inflated variation in estimating their conditional FDP.

Theorem (1) and the subsequent results hold when the following two assumptions are met: 1) the test statistic has the of the form $\boldsymbol{a}^T\boldsymbol{X}/S_X$, and 2) $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are independent. In practice, both assumptions are vulnerable. The test statistic may take different forms, depending on many factors such as the nature of the data (RNA-Seq or microarray), the experimental design structure, and the statistical hypothesis to be tested. The independence assumption between $\boldsymbol{a}^T\boldsymbol{X}$ and S_X are unlikely to hold unless the statistic is derived from two-sample t-test for normally distributed random variables. Therefore, the application of Theorem 1 is somewhat limited. Yet one goal of this study is to raise awareness that the approximation of $\rho_T(m)$ by sample correlation (after treatment effect removed) should not be taken for granted. In the future, we will explore the relationship between $\rho_T(m)$ and ρ for more general cases and for other types of statistics.

The R codes for reproducing results in this paper are available at Github: `https://github.com/zhuob/CorrelatedTest`.

Acknowledgement

Research reported in this article was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977 (to YD). We thank Sarah Emerson for valuable comments and suggestions in method development and manuscript preparation. This article is part of doctor dissertation of BZ under the supervision of YD.

References

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

[2] D. M. Gatti, W. T. Barry, A. B. Nobel, I. Rusyn, F. A. Wright, Heading down the wrong pathway: on the influence of correlation within gene sets, BMC genomics 11 (1) (2010) 574.

[3] Y.-T. Huang, X. Lin, Gene set analysis using variance component tests, BMC Bioinformatics 14 (1) (2013) 210.

[4] X. Qiu, A. I. Brooks, L. Klebanov, A. Yakovlev, The effects of normalization on the correlation structure of microarray data, BMC bioinformatics 6 (1) (2005) 120.

[5] J. D. Storey, The positive false discovery rate: a bayesian interpretation and the q-value, Annals of statistics (2003) 2013–2035.

[6] W. T. Barry, A. B. Nobel, F. A. Wright, A statistical framework for testing functional categories in microarray data, The Annals of Applied Statistics (2008) 286–315.

[7] B. Efron, Correlation and large-scale simultaneous significance testing, Journal of the American Statistical Association 102 (477).

[8] D. Wu, G. K. Smyth, Camera: a competitive gene set test accounting for inter-gene correlation, Nucleic acids research 40 (17) (2012) e133–e133.

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

[10] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Annals of statistics (2001) 1165–1188.

[11] J. J. Goeman, P. Bühlmann, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics 23 (8) (2007) 980–987.

[12] G. Yaari, C. R. Bolen, J. Thakar, S. H. Kleinstein, Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations, Nucleic Acids Research (2013) gkt660.

[13] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for

interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America 102 (43) (2005) 15545–15550.

[14] B. Zhuo, D. Jiang, Meaca: efficient gene-set interpretation of expression data using mixed models, bioRxiv (2017) 106781.

[15] J. Lee Rodgers, W. A. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician 42 (1) (1988) 59–66.

[16] R. A. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika (1915) 507–521.

[17] A. H. Joarder, Moments of the product and ratio of two correlated chi-square variables, Statistical Papers 50 (3) (2009) 581–592.

[18] X. Qiu, L. Klebanov, A. Yakovlev, Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes, Statistical Applications in Genetics and Molecular Biology 4 (1).

Appendix A

Proof of Theorem 2

It is useful to note that $\boldsymbol{U} = (\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ is independent of $\boldsymbol{S} = (S_X, S_Y)$, following from Lemmas 1 and 2.

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

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

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

For proof of Lemma 1, interested readers are referred to Joarder [17]. It immediately follows from Lemma 1 that $\boldsymbol{W}_1 = (\frac{(n_1-1)S_{X,1}^2}{\sigma_X^2}, \frac{(n_1-1)S_{Y,1}^2}{\sigma_Y^2})$ follows bivariate chi-square distribution with degree of freedom $n_1 - 1$. Similarly, $\boldsymbol{W}_2 = (\frac{(n_2-1)S_{X,2}^2}{\sigma_X^2}, \frac{(n_2-1)S_{Y,2}^2}{\sigma_Y^2})$ follows a bivariate chi-square distribution with degree of freedom $n_2 - 1$. Note that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independent since the samples are independent.

Lemma 2. *$\boldsymbol{U} = (U_X, U_Y)$ is independent of $\boldsymbol{S} = (S_X, S_Y)$.*

Proof: By Lemma 1, the density function of $\boldsymbol{W}_1 + \boldsymbol{W}_2$ only involves $\sigma_X^2, \sigma_Y^2, \rho$ and sample size n_1, n_2, therefore we can denote its density by some function $g(\sigma_X^2, \sigma_Y^2, \rho, n_1 + n_2)$. Note that $\boldsymbol{S}^2 = \frac{(\sigma_X^2,\ \sigma_Y^2)}{n_1+n_2-2}(\boldsymbol{W}_1 + \boldsymbol{W}_2)^T$ is a linear transformation of $\boldsymbol{W}_1 + \boldsymbol{W}_2$, so its density also can be expressed in terms of $\sigma_1^2, \sigma_2^2, \rho, n_1, n_2$. 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 2 implies that $U_X U_Y$ is also independent of $S_X^{-1}S_Y^{-1}$, and therefore $E(\frac{U_X}{S_X} \cdot \frac{U_Y}{S_Y})$ can be expressed as $E(U_X U_Y)E(S_X^{-1}S_Y^{-1})$. We can apply Theorem

1 to calculate the correlation between T_X and T_Y under two sample t-test for equal variance.

Next we prove Theorem 2. First note that by Lemma 2 we have

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

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

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

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

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

$$\begin{aligned}E(U_X U_X) &= \mathrm{Cov}(U_X, U_Y) + E(U_X)E(U_Y) \\ &= \rho\sigma_X\sigma_Y\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \Delta_X\Delta_Y\end{aligned} \tag{A.2}$$

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

$$E\left(S_X^{-1}\right) = \frac{\sqrt{B}}{\sigma_X}, \quad \mathrm{Var}\left(S_X^{-1}\right) = \frac{A-B}{\sigma_X^2} \tag{A.3}$$

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

$$\begin{aligned}E(S_X^{-1}S_Y^{-1}) &= E(S_X^{-1})E(S_Y^{-1}) + \rho_s\sqrt{\mathrm{Var}(S_X^{-1})\mathrm{Var}(S_Y^{-1})} \\ &= \frac{B}{\sigma_X\sigma_Y} + \rho_s\frac{A-B}{\sigma_X\sigma_Y}\end{aligned} \tag{A.4}$$

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 2 U_i and $\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}$ are independent for $i = X, Y$, we have

$$\frac{\frac{U_i-\Delta_i}{\sigma_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{(n_1+n_2-2)S_i^2}{\sigma_i^2}/(n_1+n_2-2)} = \frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{A.5}$$

It follows from

$$E\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = 0, \quad \text{Var}\left(\frac{U_i-\Delta_i}{S_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = \frac{n_1+n_2-2}{n_1+n_2-4} \tag{A.6}$$

that

$$\begin{aligned} E\left(\frac{U_i}{S_i}\right) &= \frac{\Delta_i}{\sigma_i}\sqrt{B} \\ \text{Var}\left(\frac{U_i}{S_i}\right) &= A\left(\frac{1}{n_1}+\frac{1}{n_2}\right) + \frac{\Delta_i^2}{\sigma_i^2}(A-B) \end{aligned} \tag{A.7}$$

Finally, the test-statistic correlation (22) is obtained by plugging equations (A.2–A.7) into equation (A.1).

Appendix B

Proof of Theorem (3)

Lemma 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 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 (23), and $\beta = (4+n_1n_2^{-1}+n_1^{-1}n_2)^{-1}$.

Proof: Note that

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

We will use second order Stirling's formula,

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

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

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

It can also be shown following equation (B.3) that

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

And the results immediately follow.

Lemma 4. *Let $(X_j, Y_j), j = 1, \ldots, n$ be i.i.d. random variables under the two sample t-test for equal variance setting, with mean specified in equation (19) covariance structure in equation (3). Then we have $\lim_{n\to\infty} \rho_s = \rho^2$.*

Proof: Let's first look at samples $j = 1, \ldots, n_1$. Note that

$$S^2_{X,1} = \frac{1}{n_1}\sum_{j=1}^{n_1}(X_j - \bar{X}_1)^2 \tag{B.5}$$

is the *maximum likelihood estimator* (MLE) for σ^2_X. By invariance property of MLE,

the pooled variance estimator

$$\begin{pmatrix} S^2_X \\ S^2_Y \end{pmatrix} = a_1 \begin{pmatrix} S^2_{X,1} \\ S^2_{Y,1} \end{pmatrix} + a_2 \begin{pmatrix} S^2_{X,2} \\ S^2_{Y,2} \end{pmatrix} \tag{B.6}$$

where

$$n = n_1 + n_2, \quad a_1 = \frac{n_1 - 1}{n-2}, \quad a_2 = \frac{n_2 - 1}{n-2}$$

is also MLE for $(\sigma_X^2, \sigma_Y^2)^T$ respectively. It can be shown that

$$\begin{gathered} E[S_X^2] = \sigma_X^2,\ E[S_Y^2] = \sigma_Y^2, \\ \mathrm{Var}[S_X^2] \to \frac{2\sigma_X^4}{n},\ \mathrm{Var}[S_Y^2] \to \frac{2\sigma_Y^4}{n},\ \mathrm{Cov}(S_X^2, S_Y^2) \to \frac{2\rho^2\sigma_X^2\sigma_Y^2}{n} \end{gathered} \tag{B.7}$$

We have

$$\sqrt{n}\left[\begin{pmatrix} S_{X,1}^2 \\ S_{Y,1}^2 \end{pmatrix} - \begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\begin{pmatrix} \sigma_X^4 & \rho^2\sigma_X^2\sigma_Y^2 \\ \rho^2\sigma_X^2\sigma_Y^2 & \sigma_Y^4 \end{pmatrix}\right] \tag{B.8}$$

If we let $g(x) = x^{-\frac{1}{2}}$, and apply δ-method to equation (B.8), we obtain

$$\sqrt{n}\left[\begin{pmatrix} S_X^{-1} \\ S_Y^{-1} \end{pmatrix} - \begin{pmatrix} \sigma_X^{-1} \\ \sigma_Y^{-1} \end{pmatrix}\right] \xrightarrow{d.} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma_X^{-2} & \rho^2\sigma_X^{-1}\sigma_Y^{-1} \\ \rho^2\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\right] \tag{B.9}$$

Therefore we have $\mathrm{Corr}(S_X^{-1}, S_Y^{-1}) \to \rho^2$.

Theorem (3) follows from Lemma (3) and (4).