Test-statistic correlation and data-row correlation

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

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

Abstract

When a statistical test is repeatedly applied to rows of a data matrix, correlations among data rows will give rise to correlations among corresponding test statistics. We investigate the relationship between test-statistic correlation and data-row correlation and discuss its implications.

Keywords: test-statistic correlation, bivariate normal, two-sample t-test

2010 MSC: 00-01, 99-00

1. Introduction

Many scientific data sets are organized in matrix forms and statistical inferences—such as hypothesis tests and regression analysis—are often repeatedly applied to individual rows of the data matrix. For example, in gene expression analysis, normalized expression values are often organized in a matrix with rows corresponding to genes and columns corresponding to biological samples (experimental units). In a two-group comparison experiment, a two-sample test will be applied to each row of the data matrix in order to assess differential expression (DE). For more complex experimental designs, regression analysis can be used.

Correlations may exist among the data rows: For example, between-gene correlations are commonly observed in gene expression data [1, 2, 3, 4, 5]. Data-row correlations can give rise to correlations among the test statistic values calculated from the data rows [6, 7, 8]. The dependence among test statistic

*Corresponding author

Email address: rickyb.zh@gmail.com (Bin Zhuo)

values has brought methodological challenges to statistical procedures aiming to summarize the collection of test results. For example, some multiple hypothesis testing procedures determine a p-value cutoff by controlling the *false discovery rate* (FDR) [9] or the *q-value* [5]. Many FDR-control procedures are valid only when the test statistics satisfy certain independence or positive-dependence conditions [9, 10]. Furthermore, Efron [7] showed in a simulation study that for a nominal FDR of 0.1, the actual false discovery proportions (FDP) in individual experiments can easily vary by a factor of 10 when there are correlations among test statistics.

In a gene-set analysis, one tests for over-abundance of DE genes in a specified gene set (e.g., a molecular pathway or a gene ontology category) [11]. The correlations among DE test statistics, if not addressed appropriately, will undermine the validity of many gene-set tests [2, 8, 12]. A better understanding of the test-statistic correlations is thus of fundamental importance and is a first step towards developing statistical methods that correctly account for test-statistic correlations.

Without replicating the experiment, we cannot directly estimate the correlation between a pair of test statistic values, because there is only one observed test statistic value for each data row. For this reason, the correlation between the corresponding data rows (after treatment effects accounted for) is sometimes used as a surrogate—explicitly or implicitly—when one actually needs the test-statistic correlation. It is yet unclear when and to what extent the test-statistic correlation (e.g., as measured by the Pearson correlation coefficient) can be approximated by the corresponding data-row correlation, though some simulation results suggest connections between the two quantities. Efron [7] concluded through simulation that the distribution of z-value (the test statistic considered in that paper) correlation can be nearly represented by the distribution of sample correlation from the data rows. Barry et al. [6] showed by Monte Carlo simulation of gene expression data that a nearly linear relationship holds between test-statistic correlations and data-row correlations for several forms of test statistics they examined. These Monte Carlo simulation results were cited

by Wu and Smyth [8] as a justification for estimating a variance inflation factor from data-row correlations in order to correct for test-statistic correlations.

In this paper, we derive an analytical formula for the test-statistic correlation as a function of the data-row correlation for a general class of test statistics—including the familiar two-sample t-test as a special case. We use simulation results to confirm our analytical findings. We show that 1) the test-statistic correlation is equal to data-row correlation when the test statistic is a linear combination of the observed data, but 2) in general, the test-statistic correlation is weaker than and not well approximated by the corresponding data-row correlation. In particular, our analytical formula reveal that 3) the test-statistic correlation depends on whether the test statistic has an expectation of 0 (which often corresponds to whether the null hypothesis is true). These findings urge us to give more thoughts about correlations when trying to summarize the collection of the test results.

2. Methods and Results

Suppose we have a data matrix and have applied a statistical test to individual rows of the data matrix. We will consider pairwise correlations and focus on two rows of the data matrix: $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with mean vectors $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$. We will assume that the columns of the data matrix are independent so that (X_j, Y_j), $j = 1, \ldots, n$, are independent bivariate random variables: this assumption is usually reasonable in a designed experiment for two-group comparison. The mean of (X_j, Y_j) may vary across experimental units $j = 1, \ldots, n$, but we assume that the population variance-covariance structure remains the same across experimental units, that is,

$$\mathrm{Cov}\begin{pmatrix} X_j \\ Y_j \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{1}$$

for all $j = 1, \ldots, n$. We consider a general class of test statistic of the form

$$T_X = \frac{\boldsymbol{a}^T\boldsymbol{X}}{c_X S_X}, \quad T_Y = \frac{\boldsymbol{a}^T\boldsymbol{Y}}{c_Y S_Y}, \tag{2}$$

where $\boldsymbol{a}$ is a non-zero n-vector, (c_X, c_Y) are non-random constants, and $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ and (S_X, S_Y) are independent. In particular, the familiar two-sample t-test is of this form with S_X and S_Y estimating σ_X and σ_Y respectively. So is the t-test for a regression coefficient in a linear regression model.

We want to investigate the connections between the test-statistic correlation $\rho_T = \mathrm{Cor}(T_X, T_Y)$ and the data-row correlation $\rho = \mathrm{Cor}(X_j, Y_j)$ (common to all units j). First, we present an analytical formula that relates ρ_T to ρ.

Theorem 1. *For the test statistics T_X, T_Y in (2):*

$$\rho_T = \frac{\rho\sigma_X\sigma_Y \,\mathrm{E}(S_X^{-1}S_Y^{-1}) + c^{-1}d_X d_Y \,\mathrm{Cov}(S_X^{-1}, S_Y^{-1})}{\sqrt{\left[\sigma_X^2 \,\mathrm{E}(S_X^{-2}) + c^{-1}d_X^2 \,\mathrm{Var}(S_X^{-1})\right]\left[\sigma_Y^2 \,\mathrm{E}(S_Y^{-2}) + c^{-1}d_Y^2 \,\mathrm{Var}(S_Y^{-1})\right]}} \tag{3}$$

where $d_X = \boldsymbol{a}^T\boldsymbol{\mu}_X$, $d_Y = \boldsymbol{a}^T\boldsymbol{\mu}_Y$ and $c = \boldsymbol{a}^T\boldsymbol{a}$.

Proof. (c_X, c_Y) do not affect correction and can be ignored. For any (U_X, U_Y) that are independent of (S_X, S_Y), direct calculation shows that

$$\mathrm{Cov}\left(\frac{U_X}{S_X}, \frac{U_Y}{S_Y}\right) = \mathrm{Cov}(U_X, U_Y)\,\mathrm{E}\left(\frac{1}{S_X S_Y}\right) + \mathrm{E}(U_X)\,\mathrm{E}(U_Y)\,\mathrm{Cov}\left(\frac{1}{S_X}, \frac{1}{S_Y}\right),$$

and

$$\mathrm{Var}\left(\frac{U_i}{S_i}\right) = \mathrm{Var}(U_i)\,\mathrm{E}\left(\frac{1}{S_i^2}\right) + \left(E(U_i)\right)^2 \mathrm{Var}\left(\frac{1}{S_i}\right), \text{ for } i = X, Y.$$

For this theorem, we let $U_X = \boldsymbol{a}^T\boldsymbol{X}$, $U_Y = \boldsymbol{a}^T\boldsymbol{Y}$, then $\mathrm{E}(U_X) = \boldsymbol{a}^T\boldsymbol{\mu}_X$, $\mathrm{E}(U_Y) = \boldsymbol{a}^T\boldsymbol{\mu}_Y$, $\mathrm{Var}(U_X) = \sigma_X^2\boldsymbol{a}^T\boldsymbol{a}$, $\mathrm{Var}(U_Y) = \sigma_Y^2\boldsymbol{a}^T\boldsymbol{a}$, and $\mathrm{Cov}(U_X, U_Y) = \rho\sigma_X\sigma_Y\boldsymbol{a}^T\boldsymbol{a}$ since the columns of the data matrix are assumed independent.

□

To apply equation (3) in practice, we need to compute the involved moments of (S_X^{-1}, S_Y^{-1}), but equation (3) offers some insights without explicit calculation of those quantities.

Corollary 1. *$\rho_T = \rho$ if S_X and S_Y are constants (i.e., not random).*

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 $\mathrm{E}(S_X^{-1}S_Y^{-1}) = S_X^{-1}S_Y^{-1} = \sqrt{E(S_X^{-2})E(S_Y^{-2})}$. □

This corollary says that for z-tests, the test-statistic correlation is the same as the corresponding data-row correlation ([6] also pointed out this). This confirms the simulation results in [7]. Another intuition offered by equation (3) is that the relation between ρ_T and ρ depends on whether one or both of $\boldsymbol{a}^T\boldsymbol{\mu}_X$ and $\boldsymbol{a}^T\boldsymbol{\mu}_Y$ are 0—which often corresponds to whether the corresponding null hypotheses are true. When both $\boldsymbol{a}^T\boldsymbol{\mu}_X$ and $a^T\boldsymbol{\mu}_Y$ are 0, equation (3) will have a simpler form

$$\rho_T = \frac{\rho\,\mathrm{E}(S_X^{-1}S_Y^{-1})}{\sqrt{\mathrm{E}(S_X^{-2})\,\mathrm{E}(S_Y^{-2})}}.$$

Intuitively, in such cases, we can expect $\rho_T \approx \rho$ in large samples if S_X and S_Y are "good" estimators of σ_X and σ_Y: $\mathrm{E}(S_X^{-1}S_Y^{-1})$, $\sqrt{E(S_X^{-2})E(S_Y^{-2})}$ will then both tend to $\sigma_X^{-1}\sigma_Y^{-1}$.

More generally, though, the test-statistic correlation ρ_T is not the same as the data-row correlation ρ. Next, using the important special case of two-sample t-test, we will further demonstrate that, in general, ρ_T is not well approximated by ρ, even in large samples.

For equal-variance two-sample t-test, we let $\boldsymbol{a} = (\underbrace{-\frac{1}{n_1},\ldots,-\frac{1}{n_1}}_{n_1},\underbrace{\frac{1}{n_2},\ldots,\frac{1}{n_2}}_{n_2})^T$, $c_X = c_Y = \sqrt{1/n_1 + 1/n_2}$, and S_X^2, S_Y^2 be the pooled sample variances,

$$S_i^2 = \frac{(n_1-1)S_{i,1}^2 + (n_2-1)S_{i,2}^2}{n_1+n_2-2}, \quad \text{for } i = X, Y,$$

in (2), where $S_{i,1}^2$ and $S_{i,2}^2$ are the sample variances for sample 1 and sample 2 respectively in data row i. From Basu's lemma, $(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{a}^T\boldsymbol{Y})$ are independent of (S_X, S_Y). Typically, the null hypotheses to test are $d_X = \boldsymbol{a}\boldsymbol{\mu}_X = 0$ and $d_Y = \boldsymbol{a}\boldsymbol{\mu}_Y = 0$.

Theorem 2. *For the equal-variance two-sample t-test, when $n = n_1 + n_2 \to \infty$ and $n_1/n \to r$ for some r, $0 < r < 1$,*

$$\rho_T \to \frac{\rho(1+\beta\delta_X\delta_Y\rho)}{\sqrt{(1+\beta\delta_X^2)(1+\beta\delta_Y^2)}}, \tag{4}$$

where $\delta_X = d_X/\sigma_X$, $\delta_Y = d_Y/\sigma_Y$, and $\beta = r(1-r)/2$.

Proof. As $n = n_1 + n_2 \to \infty$ and $n_1/n \to r$, in equation (3),

$$nc = n\boldsymbol{a}^T\boldsymbol{a} = n(\frac{1}{n_1} + \frac{1}{n_2}) \to \frac{1}{r} + \frac{1}{1-r} \text{ and } (nc)^{-1} \to r(1-r) = 2\beta.$$

The key of the proof is to determine the limits of the moments $\mathrm{E}(S_X^{-1}S_Y^{-1})$, $\mathrm{E}(S_X^{-2})$, $\mathrm{E}(S_Y^{-2})$, $\mathrm{Cov}(S_X^{-1}, S_Y^{-1})$, $\mathrm{Var}(S_X^{-1})$, and $\mathrm{Var}(S_Y^{-1})$. By the consistency of (S_X^2, S_Y^2) and the continuous mapping theorem,

$$S_i^{-1} \xrightarrow{p} \sigma_i^{-1}, \quad S_i^{-2} \xrightarrow{p} \sigma_i^{-2}, \text{ for } i = X, Y, \text{ and } S_X^{-1}S_Y^{-1} \xrightarrow{p} \sigma_X^{-1}\sigma_Y^{-1}.$$

For large v $(= n-2)$, $\mathrm{E}(S_i^{-4}) = \sigma_i^{-4}v^2/(v-2)(v-4) < 2\sigma_i^{-4}$, for $i = X, Y$. This implies the S_X^{-1}, S_Y^{-1}, S_X^{-2}, S_Y^{-2} and $S_X^{-1}S_Y^{-1}$ are all uniformly integrable (note that $\mathrm{E}(S_X^{-2}S_Y^{-2}) \le \sqrt{\mathrm{E}(S_X^{-4})\,\mathrm{E}(S_Y^{-4})}$), and thus

$$\mathrm{E}(S_i^{-1}) \to \sigma_i^{-1}, \quad \mathrm{E}(S_i^{-2}) \to \sigma_i^{-2}, \text{ for } i = X, Y, \quad \mathrm{E}(S_X^{-1}S_Y^{-1}) \to \sigma_X^{-1}\sigma_Y^{-1}.$$

In Lemma 1 in the Appendix, we show that

$$\sqrt{v}\left[\begin{pmatrix} S_X^{-1} \\ S_Y^{-1} \end{pmatrix} - \begin{pmatrix} \sigma_X^{-1} \\ \sigma_Y^{-1} \end{pmatrix}\right] \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma_X^{-2} & \rho^2\sigma_X^{-1}\sigma_Y^{-1} \\ \rho^2\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\right]. \quad (5)$$

This together with the continuous mapping theorem suggests that

$$\mathrm{Var}(\sqrt{v}S_i^{-1}) = \mathrm{E}\left[v(S_i^{-1} - \sigma_X^{-1})^2\right] \to \frac{1}{2}\sigma_i^{-2}, \text{ for } i = X, Y,$$

$$\mathrm{Cov}(\sqrt{v}S_X^{-1}, \sqrt{v}S_Y^{-1}) = \mathrm{E}\left[v(S_X^{-1} - \sigma_X^{-1})(S_Y^{-1} - \sigma_Y^{-1})\right] \to \frac{1}{2}\rho^2\sigma_X^{-1}\sigma_Y^{-1}.$$

For these moments limits to hold, we need to show that the involved moments are uniformly integrable (see, e.g., Theorem 6.2 of [13]). It is sufficient to show that $\mathrm{E}\left[(\sqrt{v}\cdot(S_X^{-1} - \sigma_X^{-1}))^4\right]$ is bounded for large v: in Lemma 2 in the Appendix, we show that

$$\mathrm{E}\left[(\sqrt{v}\cdot(S_X^{-1} - \sigma_X^{-1}))^4\right] = \frac{3}{4}\sigma_X^{-4} + O(v^{-1}).$$

Plugging the limiting values of $\mathrm{E}(S_X^{-1}S_Y^{-1})$, $\mathrm{E}(S_X^{-2})$, $\mathrm{E}(S_Y^{-2})$, $\mathrm{Cov}(S_X^{-1}, S_Y^{-1})$, $\mathrm{Var}(S_X^{-1})$, and $\mathrm{Var}(S_Y^{-1})$ into equation (3) gives equation (4). □

In the Appendix, we will also explain how to compute ρ_T in finite samples for two-sample t-test. It is mainly $\mathrm{E}(S_X^{-1}S_Y^{-1})$ that is difficult to compute.

Theorem 2 (and equation (10) in the Appendix) reaffirms that the relation between ρ_T and ρ depends on $(\delta_X, \delta_Y) = (d_X/\sigma_X, d_Y/\sigma_Y)$. Figure 1 shows the contour plot of the limiting value of ρ_T when $n_1 = n_2 \to \infty$ ($r = 1/2$, $\beta = 1/8$) as a function of $(\delta_X,\ \delta_Y)$, for $\rho = -0.7, -0.1, 0.1, 0.7$. Note that $\rho_T \to \rho$ if $d_X = d_Y = 0$: typically, this means both null hypotheses are true. One can show that $|\lim_{n\to\infty} \rho_T| \leq |\rho|$. That is to say, in general, the test-statistic correlation is weaker than the corresponding data-row correlation.

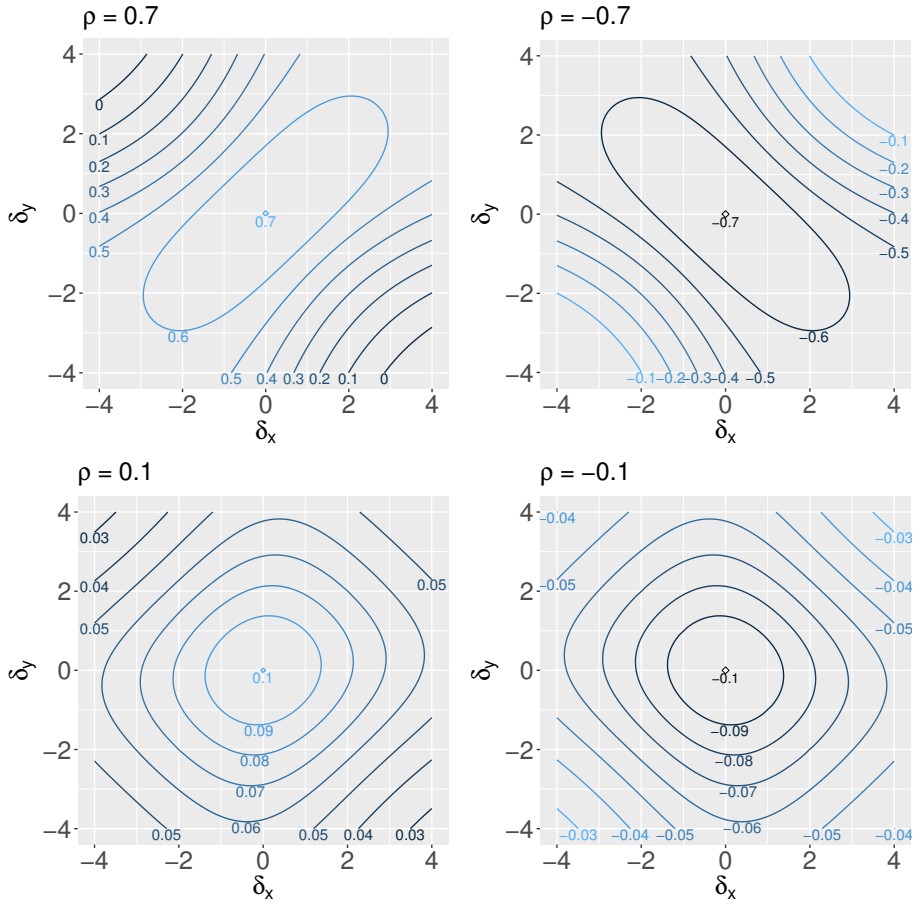

Figure 1: Contour plot of the limiting values of ρ_T as $n_1 = n_2 \to \infty$. For each ρ, the asymptotic value of ρ_T is plotted as a function of (δ_X, δ_Y).

In Figure 2, we plotted ρ_T as a function of ρ when $n_1 = n_2 = 3, 10$ or ∞ for a few selected values of (δ_X, δ_Y). We also added simulated values of ρ_T (for $n_1 = n_2 = 3, 10$) to confirm our analytical findings: For each (δ_X, δ_Y) value, we let ρ vary form -1 to 1 by a step size of 0.01. For each ρ, we simulated a pair of data rows $\boldsymbol{X}$, $\boldsymbol{Y}$ according to independent bivariate normal distributions:

$$\begin{pmatrix} X_j \\ Y_j \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], j = 1, \ldots, n_1,$$

$$\begin{pmatrix} X_j \\ Y_j \end{pmatrix} \sim N\left[\begin{pmatrix} \delta_X \\ \delta_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], j = n_1 + 1, \ldots, n_1 + n_2$$

and computed the two-sample t-test statistics T_X and T_Y. $H = 5000$ pairs of (T_X, T_Y) were simulated and their sample correlation $\hat{\rho}_T$ were shown in Figure 2.

Let $\rho_T^\infty = \lim \rho_T$ as $n_1 = n_2 \to \infty$. We see from Figure 2 that when $\delta_X = \delta_Y = 0$, $\rho_T^\infty = \rho$; when $\delta_X = 0$, ρ_T^∞ is a linear function of ρ; and when δ_X and δ_Y are both non-zero, ρ_T^∞ is a quadratic function of ρ. These features are predictable from the analytical formula (4) in Theorem 2 and they hold approximately in finite samples if n is large. In fact, we see that when $n_1 = n_2 = 10$, the ρ_T values are already remarkably close to ρ_T^∞.

In small samples (e.g., $n_1 = n_2 = 3$), there is more difference between ρ_T and ρ_T^∞: ρ_T is often weaker than ρ_T^∞ (i.e., $|\rho_T| < |\rho_T^\infty|$) with a couple of exceptions (e.g., when $\delta_X = \pm 5$, $\delta_Y = 5$), which is reasonable since (S_X^2, S_Y^2) are "noisier" in small samples and noise in general reduces correlation. When both δ_X and δ_Y are non-zero (this typically means both null hypotheses are false), ρ does not approximate ρ_T well no matter what the sample size is: $|\rho|$ can significantly overestimate $|\rho_T|$. In extreme cases when $\delta_X \delta_Y$ is big, ρ_T and ρ can have opposite signs.

3. Conclusion and discussion

This article discusses the relation between test-statistic correlation ρ_T and the corresponding data-row correlation ρ. Our results indicate that only in

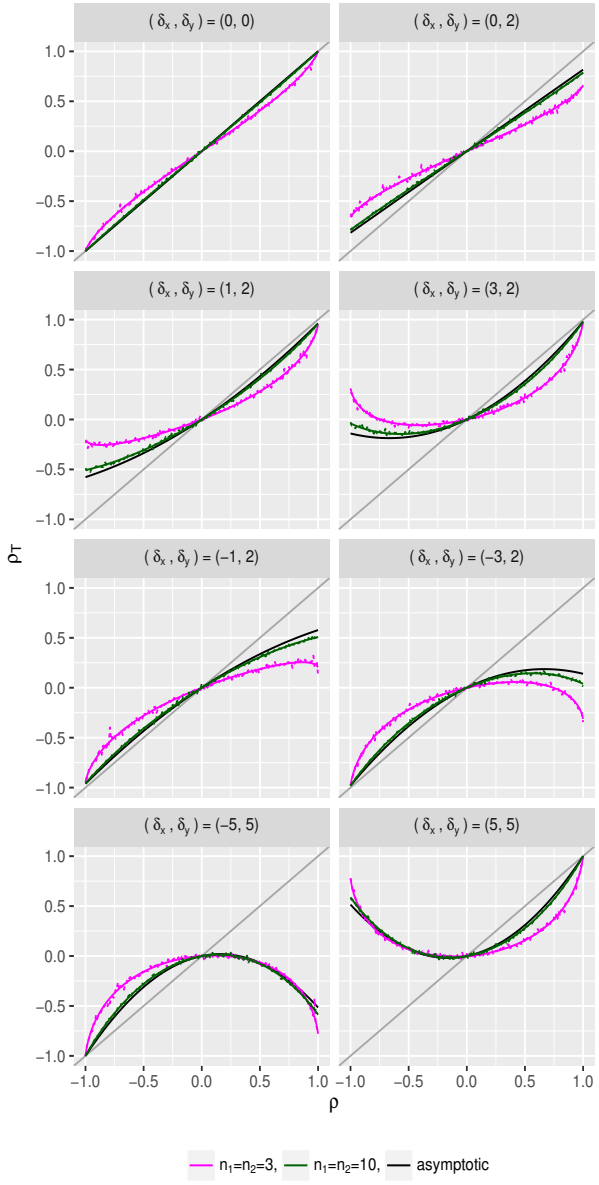

Figure 2: Test-statistic correlation ρ_T versus data-row correlation ρ at different (δ_X, δ_Y) values, when $n_1 = n_2 = 3$, 10, or ∞. The simulated values of ρ_T are also shown for $n_1 = n_2 = 3$ and 10. The solid (smooth) lines represent theoretical value, and dashed (jagged) lines represent simulated values.

limited settings, ρ_T can be well approximated by ρ: for example, $\rho_T = \rho$ for z-test and $\rho_T \approx \rho$ in large samples if both null hypotheses are true. For two-sample t-test, the relation between ρ_T and ρ will depend on (δ_X, δ_Y), the expected mean differences divided by the respective standard deviations of the data rows. When δ_X and δ_Y are both non-zero, ρ_T is a quadratic function of ρ, ρ_T can be much weaker than ρ ($|\rho_T| < |\rho|$), and ρ_T and ρ can sometimes have opposite signs.

Our findings have practical implications in statistical inferences aiming to summarize the collection of test results. For example, our results indicate that it is not reliable to approximate the distribution of test-statistic correlations by the distribution of data-row correlations if we expect the null hypotheses to be false for a significant proportion of the rows—which is often the case in gene expression analysis. If one wants to assess the null distribution of the test-statistic p-values by permuting the columns of the data matrix, then one has to realize the permutation will also change the correlations among the test-statistic values (since (δ_X, δ_Y) values will change after each permutation). In separate ongoing work, we are delving into these and related issues to better understand the impact of test-statistic correlation on gene set enrichment analysis, where one wants to test for overabundance of DE genes in a pre-specified set ([14] is one such attempt).

Wu and Smyth [8] discussed a variance inflation factor (VIF) which is useful when estimating the variance of the sum or average of m test statistics $t_1, t_2, \ldots, t_m$ when the corresponding genes (data rows) are correlated. In that paper, VIF is defined as $1 + (m-1)\bar{\rho}_T$, where $\bar{\rho}_T$ is the average of test-statistic correlations (i.e., ρ_T's) over all pairs of data rows in the set. (If all t_i's have the same variance τ^2, then $\mathrm{Var}(\bar{t}) = \mathrm{VIF} \cdot \tau^2/m$.) It was mentioned that $\bar{\rho}_T$ can be estimated by the average of data-row correlations. Our results indicate that replacing test-statistic correlations by data-row correlations will not be accurate if there are mean differences between the two groups among the data rows. For example, if we consider the two-sample t-test performed on $m = 21$ data rows in a matrix with correlated data rows ($\rho = 0.1$ for all pairs, variance $\sigma^2 = 1$ for all rows) and mean differences ranging from -3 to 3 (uniformly spaced, i.e.,

$\delta = -3, -2.4, -1.8, ..., 3$ for the 21 rows) between two groups ($n_1 = n_2 = 30$), then the true VIF value computed using test-statistic correlations (which we can compute using asymptotic formula (4) in Theorem 2) should be 2.48; the VIF computed using the data-row correlations is 3.00, which overestimates the true VIF by 21%. In practice, we can estimate ρ_T for each pair of data rows by plugging the corresponding estimated values of ρ, d, σ values into equation (4). In the Appendix, we use a simulation to show that estimating VIF using estimated ρ_T values will outperform approximating ρ_T by the sample data-row correlations.

One reviewer asked whether our results apply to the moderated t-test where the variance estimation is based on a shrinkage method. The short answer is "no". It is difficult to derive an analytical formula for the correlation between a pair of moderated t-test statistic values where a shrinkage method is used for estimating the variances of the test-statistic values, since information from all data rows are used for estimating the variances. Through a simple simulation where we applied moderated t-test in the `limma` package ([15]), we observed that the test-statistic correlations among moderated t-test statistic values still depend on δ_X and δ_Y, but the relationship between test-statistic correlations and data-row correlations do not follow the analytical formula that we derived in Theorem 2. In particular, we observed that for moderated t-test, the test-statistic correlations tend to be greater than the data-row correlations in some cases where $\delta_X = \delta_Y \neq 0$. For the usual two-sample t-test statistic, we have shown earlier that the magnitude of test-statistic correlation tends to be less than that of the data-row correlation. We included the details on the simulation settings and results of this simulation on moderated t-test in the Appendix.

In this paper, we assumed that the columns of the data matrix are independent and the explicit formula mainly focused on the two-sample t-test. We believe these are good starting points for discussing this complex issue of test-statistic correlations resulting from data-row correlations. In the future, we plan to extend our investigation into more general settings: for example, the test for regression coefficients in a generalized linear model.

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

Acknowledgement

Research reported in this article was partially 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 research was part of doctor dissertation of BZ under the supervision of YD. We thank the reviewers for their helpful comments.

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) (2007) 93–103.

[8] D. Wu, G. K. Smyth, Camera: a competitive gene set test accounting for inter-gene correlation, Nucleic acids research 40 (17) (2012) 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. DasGupta, Asymptotic theory of statistics and probability, Springer Science & Business Media, 2008.

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

[15] G. K. Smyth, Limma: linear models for microarray data, in: Bioinformatics and computational biology solutions using R and Bioconductor, Springer, 2005, pp. 397–420.

[16] F. G. Tricomi, A. Erdélyi, The asymptotic expansion of a ratio of gamma functions, Pacific Journal of Mathematics 1 (1) (1951) 133–142.

[17] F. Olver, Asymptotics and special functions, AK Peters/CRC Press, 1997.

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

Appendix

Lemmas for Theorem 2

We prove two lemmas for Theorem 2 under the theorem's conditions and specifications:

Lemma 1. *The asymptotic distribution of* $(S_X^{-1}, S_Y^{-1})^T$ *is given by (5).*

Proof. For the pooled variances in the two data rows $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$(n-2)\begin{pmatrix} S_X^2 \\ S_Y^2 \end{pmatrix} = (n_1-1)\begin{pmatrix} S_{X,1}^2 \\ S_{Y,1}^2 \end{pmatrix} + (n_2-1)\begin{pmatrix} S_{X,2}^2 \\ S_{Y,2}^2 \end{pmatrix}$$

Let

$$Z_{n,s} = \sqrt{n_s - 1}\left(\begin{pmatrix} S_{X,s}^2 \\ S_{Y,s}^2 \end{pmatrix} - \begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \end{pmatrix}\right), \quad \text{for } s = 1, 2.$$

Then

$$Z_n = \sqrt{n-2}\left(\begin{pmatrix} S_X^2 \\ S_Y^2 \end{pmatrix} - \begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \end{pmatrix}\right) = \alpha_{n,1} Z_{n,1} + \alpha_{n,2} Z_{n,2},$$

where $\alpha_{n,1} = \sqrt{n_1 - 1}/\sqrt{n-2} \to \sqrt{r}$, $\alpha_{n,2} = \sqrt{n_2 - 1}/\sqrt{n-2} \to \sqrt{1-r}$.

From the central limit theorem (or the property of MLE), as $n_1, n_2 \to \infty$, $Z_{n,1}$ and $Z_{n,2}$ are both asymptotically normally distributed as

$$N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 2\begin{pmatrix} \sigma_X^4 & \rho^2\sigma_X^2\sigma_Y^2 \\ \rho^2\sigma_X^2\sigma_Y^2 & \sigma_Y^4 \end{pmatrix}\right] \tag{6}$$

$Z_{n,1}$'s and $Z_{n,2}$'s are independent. So $Z_n = \alpha_{n,1} Z_{n,1} + \alpha_{n,2} Z_{n,2}$ converges to the same distribution in (6) by Slutsky's theorem and the continuous mapping theorem (note that $\alpha_1^2 + \alpha_2^2 = 1$).

Letting $g(x) = x^{-\frac{1}{2}}$ and applying delta method to the asymptotic distribution (6) of Z_n, we obtain (let $v = n - 2$)

$$\sqrt{v}\left[\begin{pmatrix} S_X^{-1} \\ S_Y^{-1} \end{pmatrix} - \begin{pmatrix} \sigma_X^{-1} \\ \sigma_Y^{-1} \end{pmatrix}\right] \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \sigma_X^{-2} & \rho^2\sigma_X^{-1}\sigma_Y^{-1} \\ \rho^2\sigma_X^{-1}\sigma_Y^{-1} & \sigma_Y^{-2} \end{pmatrix}\right]. \tag{7}$$

□

The next lemma shows that $\mathrm{E}\left[(\sqrt{v}\cdot(S_X^{-1}-\sigma_X^{-1}))^4\right]$ is bounded for large v.

Lemma 2.

$$\mathrm{E}\left[(\sqrt{v}\cdot(S_X^{-1}-\sigma_X^{-1}))^4\right]=\frac{3}{4}\sigma_X^{-4}+O(v^{-1})$$

Proof. For the pooled sample variance, S_X^2, in the equal-variance two-sample t-test, $Q_X=vS_X^2/\sigma_X^2\sim\mathrm{ChiSq}(v)$ with $v=n-2$ degrees of freedom, and

$$S_X^2=\frac{\sigma_X^2\cdot Q_X}{v},$$

$$S_X^{-1}=\left(\frac{\sigma_X^2\cdot Q_X}{v}\right)^{-1/2}=\sqrt{v}\cdot\sigma_X^{-1}\cdot Q_X^{-1/2},$$
$$\sqrt{v}\cdot(S_X^{-1}-\sigma_X^{-1})=\sqrt{v}\cdot\sigma_X^{-1}\cdot(\sqrt{v}\cdot Q_X^{-1/2}-1).$$

For a chi-square random variable with v degrees of freedom, $Q\sim\mathrm{ChiSq}(v)$,

$$E(Q^k)=\frac{2^k\Gamma(\frac{v}{2}+k)}{\Gamma(\frac{v}{2})},\qquad\text{for } v/2+k>0. \tag{8}$$

For large v, using expansion (12) on page 138 of [16] with $z=v/2$, $\alpha=k$ and $\beta=0$ (see also the equivalent expansion (5.02) on page 119 of [17]), we see that

$$\frac{\Gamma(\frac{v}{2}+k)}{\Gamma(\frac{v}{2})}=\left(\frac{v}{2}\right)^k\cdot\left(1+\frac{k(k-1)}{v}+\frac{k(k-1)(3(k-1)^2-k-1)}{6v^2}+O(v^{-3})\right). \tag{9}$$

Letting $k=-1/2,-1,-3/2$, and -2, it then follows from (8) and (9) that

$$\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}\right)\right]=1+\frac{3}{4v}+\frac{25}{32v^2}+O(v^{-3}),$$
$$\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}\right)^2\right]=1+\frac{2}{v}+\frac{4}{v^2}+O(v^{-3}),$$
$$\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}\right)^3\right]=1+\frac{15}{4v}+\frac{385}{32v^2}+O(v^{-3}),$$
$$\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}\right)^4\right]=1+\frac{6}{v}+\frac{28}{v^2}+O(v^{-3}),$$

and thus

$$\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}-1\right)^4\right]=\frac{3}{4v^2}+O(v^{-3}).$$

$$\begin{aligned}\mathrm{E}\left[(\sqrt{v}\cdot(S_X^{-1}-\sigma_X^{-1}))^4\right] &= v^2\cdot\sigma_X^{-4}\cdot\mathrm{E}\left[\left(\sqrt{v}\cdot Q_X^{-1/2}-1\right)^4\right]\\ &= v^2\cdot\sigma_X^{-4}\cdot\left(\frac{3}{4v^2}+O(v^{-3})\right)\\ &= \frac{3}{4}\sigma_X^{-4}+O(v^{-1}).\end{aligned}$$

□

Compute ρ_T in finite samples

It follows from (8) that for $v>2$,

$$\mathrm{E}(S_X^{-2})=A_v\sigma_X^{-2},\quad \mathrm{E}(S_X^{-1})=B_v\sigma_X^{-1},$$

where

$$A_v=\frac{v}{v-2},\quad B_v=\sqrt{\frac{v}{2}}\frac{\Gamma(\frac{v}{2}-\frac{1}{2})}{\Gamma(\frac{v}{2})},$$

and thus

$$\mathrm{Var}(S_X^{-1})=(A_v-B_v^2)\sigma_X^{-2}.$$

For bivariate normal data, Joarder [18] derived a formula (Theorem 3.1 on page 586 of that paper) for computing product moments of the form $\mathrm{E}(Q_X^aQ_Y^b)$ as infinite sums, for $(Q_X,Q_Y)=(vS_X^2/\sigma_X^2,vS_Y^2/\sigma_Y^2)$:

$$\begin{aligned}\mathrm{E}(Q_X^aQ_Y^b) &= 2^{a+b}(1-\rho^2)^{a+b+v/2}\\ &\quad\times\sum_{k=0,2,4,\ldots}(2\rho)^k\frac{\Gamma\left(\frac{k+v}{2}+a\right)\Gamma\left(\frac{k+v}{2}+b\right)\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi}k!\Gamma\left(\frac{k+v}{2}\right)\Gamma\left(\frac{v}{2}\right)}.\end{aligned}$$

So we can numerically compute $C_v(\rho)=v\,\mathrm{E}(Q_X^{-1/2}Q_Y^{-1/2})$ (for $v>2$)—which depends on ρ. Then

$$\mathrm{E}(S_X^{-1}S_Y^{-1})=C_v(\rho)\sigma_X^{-1}\sigma_Y^{-1},$$

$$\mathrm{Cov}(S_X^{-1},S_Y^{-1})=(C_v(\rho)-B_v^2)\sigma_X^{-1}\sigma_Y^{-1},$$

and equation (3) becomes

$$\rho_T = \frac{\rho C_v(\rho) + \frac{1}{vc}\delta_X\delta_Y v(C_v(\rho) - B_v^2)}{\sqrt{\left(A_v + \frac{1}{vc}v(A_v - B_v^2)\delta_X^2\right)\left(A_v + \frac{1}{vc}v(A_v - B_v^2)\delta_Y^2\right)}}. \tag{10}$$

One can show that, under the conditions of Theorem 2, as $v = n - 2 \to \infty$, $vc \to \frac{1}{r(1-r)}$, $A_v \to 1$, $B_v \to 1$, and $v(A_v - B_v^2) \to 1/2$, and thus $\mathrm{E}(S_X^{-1}) \to \sigma_X^{-1}$ and $\mathrm{Var}(\sqrt{v}S_X^{-1}) \to \frac{1}{2}\sigma_X^{-2}$. The asymptotic result in Theorem 2 suggests that as $v \to \infty$,

$$C_v(\rho) \to 1, \quad v(C_v(\rho) - B_v^2) \to \frac{1}{2}\rho^2,$$

but we did not find a direct analytical proof for these limits.

Estimating the variance inflation factor

The variance inflation factor (VIF) for a set of m test statistics, $t_1, \ldots, t_m$, is defined as

$$1 + (m-1)\bar{\rho}_T,$$

where $\bar{\rho}_T$ is the average of all pairwise test-statistic correlations (ρ_T's). In the case of two-sample t-test, given the data-row correlations and mean differences between the two groups in all data rows, we can use equation (4) to compute ρ_T for all row pairs, and in turn the VIF. If we consider the two-sample t-test performed on $m = 21$ data rows in a matrix with correlated data rows ($\rho = 0.1$ for all pairs, variance $\sigma^2 = 1$ for all rows) and mean differences ranging from -3 to 3 (uniformly spaced, i.e., $\delta = -3, -2.4, -1.8, ..., 3$ for the 21 rows) between two groups ($n_1 = n_2 = 30$), the true VIF value computed using test-statistic correlations should be 2.48; the VIF computed using the data-row correlations is 3.00, which overestimates the true VIF. In practice, for each data row $i = 1, ..., m$, the mean difference d_i can be estimated by sample mean difference. Let $r_i^T = (r_{i1}, ..., r_{in})$ be the vector of residuals after fitting the two group means, then σ_i can be estimated by $s_i = \sqrt{\frac{r_i^T \cdot r_i}{(n-2)}}$. Between a pair of data rows $i = X, Y$, the data-row correlation ρ can be estimated by the sample correlation

coefficient between the residual vectors r_X and r_Y, and the test-statistic correlation ρ_T is estimated by plugging in estimated values of ρ, d_X, d_Y, σ_X, σ_Y into equation (4). We simulated $H = 5000$ data matrices with the above specified m, n_1, n_2, ρ, and δ values, and estimated VIF by either using the estimated test-statistic correlations or directly replacing the test-statistic correlations by the estimated data-row correlations. Figure 3 summarizes the histograms of the estimated VIF values. We can see using the estimated ρ_T values to compute the VIF gives less biased results; using the data-row correlations in place of the test-statistic correlations tends to overestimate the VIF.

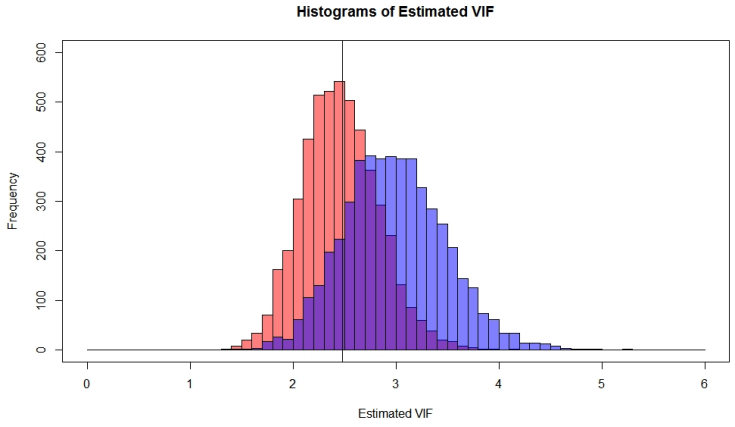

Figure 3: Histograms of the VIF values estimated by estimating ρ_T for all pairs (left) or by replacing ρ_T by corresponding sample data-row correlations (right). The vertical line indicates the true VIF value of 2.48.

Pairwise correlations among moderated t-test statistic values

We use simulation to examine pairwise test-statistic correlations for the moderated t-test as implemented in the R package `limma` ([15]) when the data rows are correlated. When computing the moderated t-statistic, the variance estimations for the data rows are shrunk towards a common value. Effectively, the variance estimation for each data row draws information from all data rows. This makes deriving an analytical formula for the pairwise test-statistic corre-

lation difficult.

To examine the correlation among test statistic values, we need to simulate multiple data sets. To see the effect of shrinkage, it is also necessary that we simulate many data rows in each data set. In each data set, we simulated 1000 rows of normal data from two samples of size 10 each—resulting in a 1000×20 data matrix. In the control group (first 10 columns), the mean was set to 0; in the treatment group (the last 10 columns), the mean was simulated to be -3 for rows 1–50, 2 for rows 51–100, and 0 for rows 101–1000 respectively. The variance was 1 for all data points, the correlation between any pair of data rows was simulated to be $\rho = 0.1$ or $\rho = 0.5$. We simulated $H = 10000$ date sets under each ρ value. To fix ideas, one can think each simulated data set as representing normalized (log-transformed) gene expression data from microarrays. The different mean levels under treatment ($\{-3, 2, 0\}$), represent different degrees of differential expression (DE). The parameter settings were kept simple here—3 DE classes ($\delta_X, \delta_Y \in \{-3, 2, 0\}$) and a constant data-row correlation ($\rho = 0.1$ or 0.5)—just enough to reveal the connection between data-row correlations and test-statistic correlations as (δ_X, δ_Y) vary.

In each data set, the moderated t-statistic value was computed for each row. Pairwise sample test-statistic correlations were then computed over the 10000 independently simulated data sets—giving a 1000×1000 matrix of sample pairwise test-statistic correlations. We grouped these pairwise test-statistic correlations according to (δ_X, δ_Y) values and summarized the distribution in each group in Figure 4 (data-row correlation is fixed at $\rho = 0.1$ or $\rho = 0.5$). The main conclusions are:

1. Test-statistic correlations are generally not same as data-row correlations and their relationship depends on the DE statuses (i.e., depending on (δ_X, δ_Y) values)—this is similar to the standard two-sample t-test case that we explored in the paper.

2. However, the relationship between test-statistic correlation and data-row correlation does not follow the analytical formula derived for the two-sample t-test case. In particular, we see that when both data rows are from the same

DE class ($\delta_X = \delta_Y = -2$ or $\delta_X = \delta_Y = 3$), the test-statistic correlations tend to be greater than the data-row correlation. For the usual two-sample t-test statistic, we have seen earlier that the magnitude of test-statistic correlation tends to be less than that of the corresponding data-row correlation.

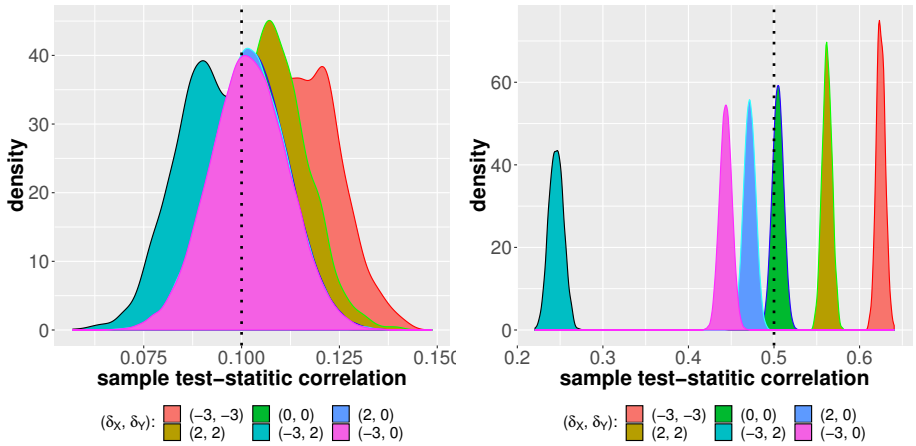

Figure 4: Distribution of pairwise sample test-statistic correlations for the moderated-t test. The pair-wise test-statistic correlations are grouped by the (unordered) values of (δ_X, δ_Y) and the distribution in each group is plotted. The data-row correlation is fixed for all pairs at $\rho = 0.1$ (left panel) or $\rho = 0.5$ (right panel) and is indicated by the vertical line in each plot.