Statistica Sinica Preprint No: SS-2018-0467	
Title	On simultaneous calibration of two-sample t-tests for high-dimension low-sample-size data
Manuscript ID	SS-2018-0467
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0467
Complete List of Authors	Chunming Zhang Shengji Jia and Yongfeng Wu
Corresponding Author	Chunming Zhang
E-mail	cmzhang@stat.wisc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

On simultaneous calibration of two-sample t-tests for high-dimension low-sample-size data

Chunming Zhang, Shengji Jia, Yongfeng Wu

Department of Statistics, 1300 University Avenue,

University of Wisconsin, Madison, 53706, USA,

cmzhang@stat.wisc.edu, *shengji@stat.wisc.edu*, *wu364@wisc.edu*

April 21, 2020

Abstract

The exact distribution is typically unavailable for a two-sample t-statistic in a single test for equal population means if we have nonGaussian samples, unequal population variances, or unequal sample sizes n_1 and n_2. In this case, a calibration method using a reference distribution offers a practically feasible substitute. This study simultaneously calibrates a diverging number m of two-sample t-statistics for inferences of significance in high-dimensional data from a small sample. For the Gaussian calibration method, we demonstrate the following. First, the simultaneous "**general**" two-sample t-statistics achieve the overall significance level, as long as $\log(m)$ increases at a strictly slower rate than $(n_1+n_2)^{1/3}$ as n_1+n_2 diverges. Second, directly applying the same calibration method to simultaneous "pooled" two-sample t-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

statistics may substantially lose the overall level accuracy. The proposed "**adaptively pooled**" two-sample t-statistics overcome such incoherence, while operating as simply and performing as well as the "general" two-sample t-statistics. Third, we propose a "**two-stage**" t-test procedure to effectively alleviate the skewness commonly encountered in various two-sample t-statistics in practice, thus increasing the calibration accuracy. Lastly, we discuss the implications of these results using simulation studies and real-data applications.

Key words and phrases: familywise error rate; multiple hypothesis testing; overall significance level; simultaneous inference; skewness.

Short title: On simultaneous calibration of two-sample t-tests

1 Introduction

With the advancement of high-throughput technology, large-scale simultaneous inference procedures [5, 12, 22, 20, 28, 29] arise naturally from high-dimensional data from small samples, with wide applications in biology, genetics, astronomy, economics, and neuroscience research among others. This problem is characterized by simultaneously carrying out a large number of hypothesis tests, where each test involves a relatively short data vector. For example, in microarray gene expression studies, the number of genes could be in the order of thousands or higher, but sample sizes could be in the order of tens or hundreds. Such procedures implicitly assume that some marginal quantities, such as the significance levels (or type-I error rates) and p-values, can be calculated exactly for each of the simultaneous tests. In practice, such an assumption may not be realistic when the exact distributions of the test statistics in finite-sample cases are not directly available. This motivates the need to estimate the distributions from which the marginal quantities are computed. However, it is unclear how good the approximation must be for the simultaneous inference to be

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

feasible.

This study investigates the performance of simultaneously conducting a diverging number m of two-sample t-tests for the equality of the mean effects of two groups, where m frequently exceeds the sample sizes n_1 and n_2 in the two groups, although the combined sample size $n = n_1 + n_2$ is still moderately large. Three issues arise naturally from analyzing such matrix-type data. First, it is well known that the exact distribution of an individual two-sample t-statistic for comparing population means is typically unavailable if we have nonGaussian samples, unequal population variances, or unequal sample sizes. Indeed, this issue remains one of the unsolved problems in the statistical literature, the so-called Behrens–Fisher problem [26, 27]. In practice, a calibration method using a reference distribution, such as the standard Gaussian distribution $\mathbb{N}(0, 1)$, serves as a feasible substitute, provided that the approximation accuracy suffices for finite sample sizes. Second, the two-sample problem is more important, in a certain sense, but more complex and challenging than the one-sample problem. Moreover, unlike the one-sample t-statistic, there is no unique method for choosing a two-sample t-statistic. The two most common choices are the "**general**" two-sample t-statistic and the "pooled" two-sample t-statistic. Nonetheless, no studies have examined whether the calibration methods for the two choices are equally applicable. Third, in practice, asymmetric populations are common, but reduce the accuracy of a single two-sample t-statistic. Here, no studies have examined simultaneous inferences based on a diverging number of two-sample t-statistics.

Owing to the popularity of two-sample t-tests, it is highly desirable to investigate how many and which two-sample t-statistics can be calibrated simultaneously before the overall level accuracy becomes poor. This study addresses three new issues for two-sample t-statistics involving independent and dependent data.

Issue 1: We demonstrate that for the Gaussian calibration method, the overall significance

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

level of the simultaneous "**general**" two-sample t-statistics can be achieved, provided that $\log(m)$ increases at a strictly slower rate than $(n_1+n_2)^{1/3}$ as n_1+n_2 diverges. Furthermore, we show that the choice of (m, n_1, n_2) controls the false discovery rates (FDRs) of some multiple testing procedures based on calibrated p-values.

Issue 2: In contrast, the "pooled" two-sample t-statistics may behave substantially differently to the "general" two-sample t-statistics, particularly when a "composite variance quantity" (CVQ; defined in (2.7)) exceeds one. The proposed "**adaptively pooled**" two-sample t-statistics in Section 3.2 operate as simply, but perform as well as the "general" two-sample t-statistics.

Issue 3: Moreover, we propose a "**two-stage**" t-test procedure in Section 3.3 to effectively alleviate the skewness effects commonly encountered from various types of two-sample t-statistics in practice, thus increasing the calibration accuracy.

In the case of simultaneous one-sample t-statistics under independence and positive regression dependence on subsets [2], calibration using a Gaussian or Student's t distribution and the bootstrap method was studied in [13], assuming that the number m_0 of true null hypotheses is identical to m; that is, $m_0 = m$, which is restrictive in applications. Here, we examine the validity of the Gaussian calibration method applied to different choices of two-sample t-statistics under independence and general dependency, where $m_0 \leq m$ is allowed and m_0 is a nonrandom quantity. To control the FDR asymptotically, we apply the factor model to deal with several practically motivated dependence models, including the jointly Gaussian distributed test statistics.

The rest of the paper is organized as follows. Section 2 formulates the overall significance level of simultaneous two-sample t-statistics that compare the means of two populations. Section 3 addresses Issues 1–3 in detail. Section 4 discusses the effect on the calibration method of dependence between observations. Sections 5 and 6 present our simulation

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

studies and real-data examples, respectively. Section 7 concludes the paper. All technical details, figures and tables are relegated to the online Supplementary Material.

2 Model structure and significance testing

Many applications test data from two groups, such as a normal control group and a cancer patient group. More formally, we consider observations $\{X_{i,j}\}$ of the X-group and $\{Y_{i,j}\}$ of the Y-group described by the signal plus noise model

$$\begin{aligned} X_{i,j} &= \mu_{X;i} + \varepsilon_{i,j}, \quad 1 \le i \le m, \ 1 \le j \le n_1, \\ Y_{i,j} &= \mu_{Y;i} + e_{i,j}, \quad 1 \le i \le m, \ 1 \le j \le n_2, \end{aligned} \tag{2.1}$$

where the index i refers to the ith test (for example, gene or brain voxel), j indicates the jth sample (for example, array or subject), constants $\mu_{X;i}$ and $\mu_{Y;i}$ stand for the mean effects from the X-group and Y-group, respectively, in the ith test, and $\varepsilon_{i,j}$ and $e_{i,j}$ are the respective random errors. Some basic assumptions are collected in conditions A1–A3 for our statistical analysis. We test the following hypotheses:

$$H_{0,i} : \mu_{X;i} = \mu_{Y;i} \quad \text{against} \quad H_{1,i} : \mu_{X;i} \neq \mu_{Y;i}, \tag{2.2}$$

simultaneously for $1 \le i \le m$. One-sided alternatives can be formulated similarly.

2.1 Single two-sample t-statistic

For testing a single null hypothesis $H_{0,i}$ in (2.2), two-sample t-statistics denoted by $T_{i;n_1,n_2}$, along with their variants, are widely used. One version is formed by the "**general**" two-sample t-statistic ([26], equation (2)),

$$T^{\text{general}}_{i;n_1,n_2} = \frac{\overline{X}_i - \overline{Y}_i}{\sqrt{s^2_{X;i}/n_1 + s^2_{Y;i}/n_2}}, \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

where $\overline{X}_i = \sum_{j=1}^{n_1} X_{i,j}/n_1$ and $\overline{Y}_i = \sum_{j=1}^{n_2} Y_{i,j}/n_2$ are the sample means within the ith test, and $s^2_{X;i} = \sum_{j=1}^{n_1}(X_{i,j} - \overline{X}_i)^2/(n_1 - 1)$ and $s^2_{Y;i} = \sum_{j=1}^{n_2}(Y_{i,j} - \overline{Y}_i)^2/(n_2 - 1)$ are the sample variances within the ith test. Under conditions A1–A3, the distribution of $T^{\text{general}}_{i;n_1,n_2}$ is given as follows.

(a1) In the special case of Gaussian errors $\varepsilon_{i,j} \sim \mathbb{N}(0, \sigma^2_{\varepsilon;i})$ and $e_{i,j} \sim \mathbb{N}(0, \sigma^2_{e;i})$, with equal variances $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i}$ and equal sample sizes $n_1 = n_2$, $T^{\text{general}}_{i;n_1,n_2}$ under $H_{0,i}$ of (2.2) follows the t_{2n_1-2}-distribution.

(a2) In other cases, the exact distribution of $T^{\text{general}}_{i;n_1,n_2}$ under $H_{0,i}$ is typically unavailable, but the central limit theorem (CLT) and Slutsky's theorem [10] give

$$T^{\text{general}}_{i;n_1,n_2} \xrightarrow{\mathcal{D}} \mathbb{N}(0,1), \quad \text{under } H_{0,i}, \tag{2.4}$$

as $n_1 \to \infty$ and $n_2 \to \infty$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Another commonly used form is the "pooled" two-sample t-statistic ([26], equation (1); [4], Section 4.9.3; [12], Section 2.1; [5]),

$$T^{\text{pool}}_{i;n_1,n_2} = \frac{\overline{X}_i - \overline{Y}_i}{s_{\text{pool}_{X;Y};i}\sqrt{1/n_1 + 1/n_2}}, \tag{2.5}$$

where $s^2_{\text{pool}_{X;Y};i} = \{(n_1-1)s^2_{X;i} + (n_2-1)s^2_{Y;i}\}/(n_1 + n_2 - 2)$ acts as a pooled sample variance within the ith test. Under conditions A1–A3, the distribution of $T^{\text{pool}}_{i;n_1,n_2}$ is given as follows.

(b1) In the special case of Gaussian errors $\varepsilon_{i,j} \sim \mathbb{N}(0, \sigma^2_{\varepsilon;i})$ and $e_{i,j} \sim \mathbb{N}(0, \sigma^2_{e;i})$, with equal variances $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i}$, $T^{\text{pool}}_{i;n_1,n_2}$ under $H_{0,i}$ of (2.2) follows the $t_{n_1+n_2-2}$-distribution.

(b2) In other cases, the exact distribution of $T^{\text{pool}}_{i;n_1,n_2}$ under $H_{0,i}$ is typically unavailable. In a large sample analysis, if $n_1 \to \infty$ and $n_2 \to \infty$ such that $n_1/(n_1 + n_2) \to \rho \in (0,1)$, then it can be shown that

$$T^{\text{pool}}_{i;n_1,n_2} \xrightarrow{\mathcal{D}} \mathbb{N}(0, \sigma^2_{\rho;\theta_{(\varepsilon,e);i}}), \quad \text{under } H_{0,i}, \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

where

$$\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} = \frac{(1-\rho)+\rho\,\theta_{(\varepsilon,e);i}}{\rho+(1-\rho)\,\theta_{(\varepsilon,e);i}}, \quad \text{with } \theta_{(\varepsilon,e);i} = \sigma^2_{e;i}/\sigma^2_{\varepsilon;i}. \tag{2.7}$$

The derivation of (2.6) is relegated to Appendix A. We call $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}}$ the CVQ, which aggregates the ratio of sample sizes and the ratio of population variances. Clearly, $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} = 1$ holds only in Case I or Case II below:

$$\text{Case I}: \quad \rho = 1/2, \text{ that is, equal sample sizes with } n_1 = n_2; \tag{2.8}$$

$$\text{Case II}: \quad \theta_{(\varepsilon,e);i} = 1, \text{ that is, equal population variances with } \sigma^2_{\varepsilon;i} = \sigma^2_{e;i}. \tag{2.9}$$

Note too that $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} > 1$ holds only if $n_1 < n_2$ and $\sigma^2_{\varepsilon;i} > \sigma^2_{e;i}$, or if $n_1 > n_2$ and $\sigma^2_{\varepsilon;i} < \sigma^2_{e;i}$. In general, the limiting distribution in (2.6) cannot be used directly, because the population variances $\sigma^2_{\varepsilon;i}$ and $\sigma^2_{e;i}$ in $\theta_{(\varepsilon,e);i}$ are typically unknown in practical settings.

2.2 Simultaneous two-sample t-statistics

When calibrating multiple two-sample t-tests $\{T_{i;n_1,n_2}\}_{i=1}^m$ simultaneously, the accuracy of the overall significance level is used to control some aspects of the overall error rate. We first use the "**general**" two-sample t-statistics $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$ to introduce some necessary notation. We discuss extensions to alternative choices $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$ in Section 3.2. For a critical value t, the significance level of the ith test is

$$\alpha_{i;n_1,n_2}(\mathrm{t}) = \mathrm{P}_{H_{0,i}}(|T^{\text{general}}_{i;n_1,n_2}| > \mathrm{t}), \tag{2.10}$$

where $\mathrm{P}_{H_{0,i}}$ denotes the probability calculated when $H_{0,i}$ is true. When testing m null hypotheses simultaneously, the indices of the true null hypotheses are collected in the set $\mathcal{I}_0 = \{i : H_{0,i} \text{ is true}\}$, with cardinality $m_0 = |\mathcal{I}_0|$. The overall significance level is captured by the family-wise-error-rate (abbreviated as FWER or FWER_1), $\text{FWER}(\mathrm{t}) = \mathrm{P}(V_m(\mathrm{t}) \geq 1)$, where $V_m(\mathrm{t}) = \sum_{i=1}^m \mathrm{I}(H_{0,i} \text{ is true},\ |T^{\text{general}}_{i;n_1,n_2}| > \mathrm{t}) = \sum_{i\in\mathcal{I}_0} \mathrm{I}(|T^{\text{general}}_{i;n_1,n_2}| > \mathrm{t})$ denotes the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

number of false rejections, with an indicator operator $\mathrm{I}(\cdot)$. More generally, for integers $k \geq 1$, $\mathrm{FWER}_k(\mathrm{t}) = \mathrm{P}(V_m(\mathrm{t}) \geq k)$ denotes the k-fold family-wise-error-rate (abbreviated as FWER_k, see [21]).

Recall from Section 2.1 that exact values of $\alpha_{i;n_1,n_2}(\mathrm{t})$ based on the exact null distribution of $T^{\mathrm{general}}_{i;n_1,n_2}$ are unavailable in many practical settings. However, when $n_1 \to \infty$ and $n_2 \to \infty$, the null distribution of $T^{\mathrm{general}}_{i;n_1,n_2}$ can be approximated by $\mathbb{N}(0,1)$, as seen in (2.4). This result motivates the approximation using $\mathbb{N}(0,1)$ random variables $\{T^{\mathrm{a}}_i\}_{i=1}^m$. It is thus natural to use the quantities,

$$\begin{aligned} &\alpha^{\mathrm{a}}_i(\mathrm{t}) = \mathrm{P}(|T^{\mathrm{a}}_i| > \mathrm{t}), && V^{\mathrm{a}}_m(\mathrm{t}) = \textstyle\sum_{i\in\mathcal{I}_0} \mathrm{I}(|T^{\mathrm{a}}_i| > \mathrm{t}), \\ &\mathrm{FWER}^{\mathrm{a}}(\mathrm{t}) = \mathrm{P}(V^{\mathrm{a}}_m(\mathrm{t}) \geq 1), && \mathrm{FWER}^{\mathrm{a}}_k(\mathrm{t}) = \mathrm{P}(V^{\mathrm{a}}_m(\mathrm{t}) \geq k), \end{aligned}$$

which are computationally feasible, as substitutes for $\alpha_{i;n_1,n_2}(\mathrm{t})$, $V_m(\mathrm{t})$, $\mathrm{FWER}(\mathrm{t})$, and $\mathrm{FWER}_k(\mathrm{t})$, respectively, when n_1 and n_2 are large.

In this study, we examine the relation between the number of tests m and the sample sizes n_1 and n_2 within each test. Here, applying appropriate choices of the critical values $\mathrm{t}^{\mathrm{a}}_{\alpha;m}$ and $\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}$ (obtained from the calibrated distributions (through $\{T^{\mathrm{a}}_i\}_{i=1}^m$)) to the two-sample t-statistics $\{T^{\mathrm{general}}_{i;n_1,n_2}\}_{i=1}^m$ and $\{T^{\mathrm{pool}}_{i;n_1,n_2}\}_{i=1}^m$ guarantees that

$$\mathrm{FWER}_1(\mathrm{t}^{\mathrm{a}}_{\alpha;m}) \leq \alpha + o(1), \tag{2.11}$$

$$\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}) \leq \alpha + o(1), \tag{2.12}$$

as $m \to \infty$, $n_1 \to \infty$, and $n_2 \to \infty$, where α is the control level. Similarly, it is ideal to control the FDR based on a certain threshold $\tau_{\alpha;m;n}$ for the true p-values $\{P_i\}$; that is, $\mathrm{FDR}(\tau_{\alpha;m;n}) \leq \alpha + o(1)$, where $\mathrm{FDR}(\tau) = \mathrm{E}[\frac{\sum_{i\in\mathcal{I}_0} \mathrm{I}(P_i \leq \tau)}{\{\sum_{i=1}^m \mathrm{I}(P_i \leq \tau)\} \vee 1}]$, with $a \vee b = \max\{a, b\}$. When the exact $\{P_i\}$ are unavailable, it is more realistic to control the corresponding FDR based on some threshold $\tau^{\mathrm{a}}_{\alpha;m;n}$ for the calibrated p-values $\{P^{\mathrm{a}}_i\}_{i=1}^m$, such that

$$\mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n}) \leq \alpha + o(1). \tag{2.13}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

3 Error controls with independent data

3.1 "General" two-sample t-tests for (2.2)

We first discuss error controls using the "**general**" two-sample t-statistics $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$, for which we require additional assumptions A4–A7. We further assume that the rates of growth of m, n_1, and n_2 are connected via

$$\log(m) = o(n^{1/3}), \tag{3.1}$$

with the combined sample size $n = n_1 + n_2$.

3.1.1 Controlling $\text{FWER}_1(\text{t}^{\text{a}}_{\alpha;m})$ in (2.11) and $\text{FWER}_k(\text{t}^{\text{a}}_{\alpha;m;k})$ in (2.12)

The validity of the calibration method is supported by (3.3) of Proposition 1, which states that the overall significance level converges to a limit that does not exceed the nominal level, the desirable property in (2.11).

Proposition 1 (control $\text{FWER}_1(\text{t}^{\text{a}}_{\alpha;m})$ under independence between tests) *Assume model* (2.1) *and that conditions* A1–A7 *hold. For* $\alpha \in (0,1)$, $m_0/m \to \pi_0 \in (0,1]$, $m \to \infty$, *and* $n \to \infty$, *if the general two-sample t-statistics* $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$ *are used,* (m,n) *satisfies* (3.1), *and*

$$\text{t}^{\text{a}}_{\alpha;m} = \Phi^{-1}(\{1+(1-\alpha)^{1/m}\}/2), \tag{3.2}$$

where Φ *denotes the cumulative distribution function* (C.D.F.) *of an* $\mathbb{N}(0,1)$ *variable, then*

$$\begin{aligned} \text{FWER}_1(\text{t}^{\text{a}}_{\alpha;m}) &= \text{FWER}^{\text{a}}_1(\text{t}^{\text{a}}_{\alpha;m}) + o(1), \\ \text{FWER}^{\text{a}}_1(\text{t}^{\text{a}}_{\alpha;m}) &= 1-(1-\alpha)^{m_0/m} \le \alpha. \end{aligned} \tag{3.3}$$

Similarly, (3.6) of Proposition 2 implies that $\text{FWER}_k(\text{t}^{\text{a}}_{\alpha;m;k}) \le \alpha + o(1)$, which is desirable in (2.12). A common feature of Propositions 1–2 is that as the proportion π_0 of true nulls approaches one, $\text{FWER}(\text{t}^{\text{a}}_{\alpha;m})$ and $\text{FWER}_k(\text{t}^{\text{a}}_{\alpha;m;k})$ approach the control level α, and hence the inequalities in (2.11)–(2.12) become equalities.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

Proposition 2 (control $\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k})$ **under independence between tests)** *Assume model* (2.1) *and that conditions* A1–A7 *hold. For* $k \geq 2$, $\alpha \in (0,1)$, $m_0/m \to \pi_0 \in (0,1]$, $m \to \infty$, *and* $n \to \infty$, *if the general two-sample t-statistics* $\{T^{\mathrm{general}}_{i;n_1,n_2}\}_{i=1}^m$ *are used,* (m,n) *satisfies* (3.1), *and*

$$\mathrm{t}^{\mathrm{a}}_{\alpha;m;k} = \Phi^{-1}(1-(\beta_{k;\alpha}/2)/m), \tag{3.4}$$

where $\beta_{k;\alpha}$ *denotes the solution of equation*

$$G_k(\beta_{k;\alpha}) = \alpha, \tag{3.5}$$

with $G_k(\beta) = 1 - \sum_{j=0}^{k-1} \beta^j/j!e^{-\beta}$ *for* $\beta \in (0,\infty)$, *then*

$$\begin{aligned} \mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}) &= \mathrm{FWER}^{\mathrm{a}}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}) + o(1), \\ \mathrm{FWER}^{\mathrm{a}}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}) &= G_k(\pi_0\beta_{k;\alpha}) + o(1) \leq \alpha + o(1). \end{aligned} \tag{3.6}$$

3.1.2 Controlling the FDR in (2.13) for multiple testing procedures

Similarly to the marginal significance levels, the true marginal p-values $\{P_i\}$ are unknown in advance or are not directly available when the exact distributions of the two-sample t-statistics are unknown, and thus need to be approximated from the calibrated distribution. The practical implication is that using the approximate p-values $\{P^{\mathrm{a}}_i\}$ means the resulting multiple testing procedure, such as the Bonferroni correction, is still valid. This because the FDR under the conditions of Proposition 1 is asymptotically bounded by the level α if the approximation errors of the p-values are $o(1/m)$.

Analogously, consider the Benjamini–Hochberg (BH) multiple testing procedure [1], which rejects the null hypotheses $H_{0,i}$ when $P_i \leq P_{(\widehat{k})}$, where $\widehat{k} = \max\{j : P_{(j)} \leq \alpha j/m\}$, and $P_{(1)} \leq \cdots \leq P_{(m)}$ denote the ordered p-values $\{P_i\}$. Then, $\mathrm{FDR}_{\mathrm{BH}} = \mathrm{E}(\frac{V_{\mathrm{BH}}}{R_{\mathrm{BH}} \vee 1})$ gives the FDR of the BH procedure, where $V_{\mathrm{BH}} = \sum_{i \in \mathcal{I}_0} \mathrm{I}(P_i \leq P_{(\widehat{k})})$ and $R_{\mathrm{BH}} = \widehat{k}$. For the calibration method, applying the approximate p-values $\{P^{\mathrm{a}}_i\}$ instead of $\{P_i\}$ to the BH procedure yields the number $V^{\mathrm{a}}_{\mathrm{BH}}$ of false rejections and the number $R^{\mathrm{a}}_{\mathrm{BH}}$ of total rejections,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

and the corresponding FDR defined by $\mathrm{FDR}^{\mathrm{a}}_{\mathrm{BH}} = \mathrm{E}(\frac{V^{\mathrm{a}}_{\mathrm{BH}}}{R^{\mathrm{a}}_{\mathrm{BH}} \vee 1})$. More generally, for the p-values $\{P_i\}$ used in the BH procedure, $\mathrm{FDR}_{\mathrm{BH}} = \mathrm{FDR}(\tau_{\alpha;m;n})$ ([24], Lemma 1), where $\mathrm{FDR}(t) = \mathrm{E}\{\frac{V_{P;m}(t)}{R_{P;m}(t) \vee 1}\}$, for $t \in [0,\, 1]$, and $\tau_{\alpha;m;n} = \sup\{t : \widehat{\mathrm{FDR}}(t) \le \alpha\}$, with $\widehat{\mathrm{FDR}}(t) = m\, t / R_{P;m}(t)$, $V_{P;m}(t) = \sum_{i \in \mathcal{I}_0} \mathrm{I}(P_i \le t)$, and $R_{P;m}(t) = \sum_{i=1}^{m} \mathrm{I}(P_i \le t)$. Similarly, for the approximate p-values $\{P^{\mathrm{a}}_i\}$, define $\mathrm{FDR}^{\mathrm{a}}(t) = \mathrm{E}\{\frac{V^{\mathrm{a}}_{P;m}(t)}{R^{\mathrm{a}}_{P;m}(t) \vee 1}\}$ and $\tau^{\mathrm{a}}_{\alpha;m;n} = \sup\{t : \widehat{\mathrm{FDR}}^{\mathrm{a}}(t) \le \alpha\}$, where $\widehat{\mathrm{FDR}}^{\mathrm{a}}(t) = m\, t / R^{\mathrm{a}}_{P;m}(t)$, $V^{\mathrm{a}}_{P;m}(t) = \sum_{i \in \mathcal{I}_0} \mathrm{I}(P^{\mathrm{a}}_i \le t)$, and $R^{\mathrm{a}}_{P;m}(t) = \sum_{i=1}^{m} \mathrm{I}(P^{\mathrm{a}}_i \le t)$.

Proposition 3 shows that the resulting $\mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n})$ can be controlled under mild conditions; Figure 13 presents simulation evaluations. Additional assumptions A5′, A7′, A8–A10 are needed.

Proposition 3 (control $\mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n})$ of the BH procedure under independence between tests) *Assume model* (2.1) *and that conditions* A1–A5, A5′, A6, A7′, *and* A8–A10 *hold. Define by* $F^{\mathrm{a}}_P(\cdot; n)$ *and* $f^{\mathrm{a}}_P(\cdot; n)$ *the* C.D.F. *and* p.d.f., *respectively, of the approximate p-values* $\{P^{\mathrm{a}}_i\}_{i=1}^m$. *For* $\alpha \in (0, 1)$, *let*

$$\varsigma_{\alpha;n} = \sup\{t : H(t; n) \le \alpha\}, \quad \varsigma^{\mathrm{a}}_{\alpha;n} = \sup\{t : H^{\mathrm{a}}(t; n) \le \alpha\},$$

where $H(t; n) = t / F_P(t; n)$ *and* $H^{\mathrm{a}}(t; n) = t / F^{\mathrm{a}}_P(t; n)$. *Suppose* $H'(t; n)$ *is bounded below for* t *in an open interval with endpoints* $\varsigma_{\alpha;n}$ *and* $\varsigma^{\mathrm{a}}_{\alpha;n}$, *and* $f^{\mathrm{a}}_P(\varsigma^{\mathrm{a}}_{\alpha;n}; n) < \alpha^{-1} < f^{\mathrm{a}}_P(0; n)$. *If the general two-sample t-statistics* $\{T^{\mathrm{general}}_{i;n_1,n_2}\}_{i=1}^m$ *are used and*

$$\Phi^{-1}(1 - \varsigma^{\mathrm{a}}_{\alpha;n}) \in (\, 0,\, o(n^{1/6})\,), \tag{3.7}$$

then as $m \to \infty$ *and* $n \to \infty$,

$$\mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n}) \le \alpha + o(1). \tag{3.8}$$

Remark 1 *Similarly to Lemma* A.1 *of* [15], *we obtain* $\tau^{\mathrm{a}}_{\alpha;m;n} = \varsigma^{\mathrm{a}}_{\alpha;n} + O_{\mathrm{P}}(m^{-1/2})$, *where* $\tau^{\mathrm{a}}_{\alpha;m;n}$ *gives the threshold for the approximate p-values. Therefore, condition* (3.7) *becomes*

$$\Phi^{-1}(1 - \tau^{\mathrm{a}}_{\alpha;m;n} + O_{\mathrm{P}}(m^{-1/2})) \in (\, 0,\, o(n^{1/6})\,), \tag{3.9}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

which implicitly describes the relationship between m *and* n*. For example, if* $\tau^{\mathrm{a}}_{\alpha;m;n}$ *is of order* m^{-b} *with probability tending to one, where* $0 < b \leq 1/2$*, then a sufficient condition for* (3.9) *is* $\log(m) = o(n^{1/3})$*, as characterized by* (3.1)*.*

Remark 2

(i) *Using similar arguments for Corollary* 2.1 *in* [23]*, we can show that* $\log(m) = o(n^{1/3})$ *is also a necessary condition for controlling* FDR *asymptotically. More precisely, if* $\log(m) \geq c_0\, n^{1/3}$ *for some constant* $c_0 > 0$*, we obtain* $\liminf_{(n,m)\to\infty} \mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n}) \geq \beta$*, with a constant* $\beta > \alpha$*. In particular, if* $\log(m)/n^{1/3} \to \infty$*, we obtain* $\mathrm{FDR}(\tau^{\mathrm{a}}_{\alpha;m;n}) \to 1$*, implying that the* FDR *is not controlled as* $m \to \infty$ *and* $n \to \infty$*.*

(ii) *On the other hand, the condition* $\log(m) = o(n^{1/3})$ *can be relaxed to a better rate* $\log(m) = o(n^{1/2})$ *with additional conditions, such as that of symmetric errors and a stronger large deviation result for the two-sample t-tests* $T^{\mathrm{general}}_{i;n_1,n_2}$*:* $\mathrm{P}_{H_{0,i}}(T^{\mathrm{general}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} = \exp(-3^{-1}\kappa_{3,i}x^3 n^{-1/2})\{1+\theta(1+x)^2/n^{1/2}\}$*, where* $\kappa_{3,i} = [\mathrm{E}\{(X_{i,1}-\mu_{X;i})^3\}/\rho^2 - \mathrm{E}\{(Y_{i,1}-\mu_{Y;i})^3\}/(1-\rho)^2]/\{\sigma^2_{X;i}/\rho + \sigma^2_{Y;i}/(1-\rho)\}^{3/2}$*, and* $\theta = \theta(x,n)$ *satisfies* $|\theta(x,n)| \leq C$ *uniformly in* $x \in (0, o(n^{1/4}))$*. The justification for this large deviation result is beyond the scope of this study. See Section* 3.3 *for a related discussion.*

(iii) *The condition* A5′*, "two-sample t-statistics corresponding to true non-nulls are identically distributed," simplifies the technical proof for Proposition* 3*. In the simulation studies in Section* 5*, where the differences* $(\mu_{X;i} - \mu_{Y;i})$ *under the true non-nulls vary with* i*, Figure* 13 *indicates that Proposition* 3 *continues to hold in cases where condition* A5′ *is relaxed.*

Remark 3 *In Propositions* 1–3*, the Gaussian distribution is used to approximate the distribution of the test statistics* $T^{\mathrm{general}}_{i;n_1,n_2}$*. These results can be easily generalized to the t-distribution approximation by replacing* $\Phi(\cdot)$ *with the* C.D.F. *of the* $t_{n_1+n_2-2}$ *distribution.*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

3.2 Proposed "adaptively pooled" two-sample t-tests for (2.2)

We now discuss error controls using the "pooled" two-sample t-statistics $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$. Recall from (A.11) and (A.26) in the Appendix A that the conclusions of Propositions 1–3 rely on the tail distribution of $T^{\text{general}}_{i;n_1,n_2}$ under the null $H_{0,i}$, approximated by that of the $\mathbb{N}(0,1)$ distribution, fulfilling

$$|\mathrm{P}_{H_{0,i}}(T^{\text{general}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} - 1| \to 0$$

uniformly in x up to a point of order $o(n^{1/6})$. Applying similar derivations to the "pooled" version of the test statistics $T^{\text{pool}}_{i;n_1,n_2}$, we observe that if the condition

$$|\mathrm{P}_{H_{0,i}}(T^{\text{pool}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} - 1| \to 0 \tag{3.10}$$

holds uniformly up to the point x of order $o(n^{1/6})$, then (2.11) and (2.12) are also applicable to $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$. Indeed, condition (3.10) holds when the CVQ is equal to one, that is, $\sigma_{\rho;\theta_{(\varepsilon,e);i}} = 1$, in either Case I with $n_1 = n_2$, as discussed in (2.8), or Case II with $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i}$, as discussed in (2.9). Numerical evidence is provided in Figure 3 with $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i}$, where the performance of the calibration method applied to the "pooled" choices (in the second column panels) is nearly identical to that applied to the "**general**" choices (in the first column panels).

Next, we examine the effect on (3.10) if the CVQ is allowed to differ from one. If the original form (2.5) of $T^{\text{pool}}_{i;n_1,n_2}$ is used, then the result in (2.6) indicates

$$\mathrm{P}_{H_{0,i}}(T^{\text{pool}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} = \{1-\Phi(x/\sigma_{\rho;\theta_{(\varepsilon,e);i}})\}/\{1-\Phi(x)\}\{1+o(1)\}. \tag{3.11}$$

To analyze the ratio on the right-hand side of (3.11), the panels of Figure 1 plot the function $\{1-\Phi(x/\sigma)\}/\{1-\Phi(x)\}$, which behaves very differently in the cases of $\sigma > 1$ and $\sigma < 1$. The maximum value of $\{1-\Phi(x/\sigma)\}/\{1-\Phi(x)\}$ is unbounded when $\sigma > 1$, but is at most one when $\sigma < 1$. This difference ultimately affects (3.10) in the following ways.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

(i) If $\sigma_{\rho;\theta_{(\varepsilon,e);i}} > 1$, then the maximum value of $|\mathrm{P}_{H_{0,i}}(T^{\mathrm{pool}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} - 1|$ will always be much larger than zero.

(ii) If $\sigma_{\rho;\theta_{(\varepsilon,e);i}} < 1$, then the maximum value of $|\mathrm{P}_{H_{0,i}}(T^{\mathrm{pool}}_{i;n_1,n_2} \geq x)/\{1-\Phi(x)\} - 1|$ will potentially approach zero, particularly when $\sigma_{\rho;\theta_{(\varepsilon,e);i}}$ approaches one.

Hence, condition (3.10) may fail if $\sigma_{\rho;\theta_{(\varepsilon,e);i}} > 1$, and the overall level accuracy may be lost by directly applying the calibration method to the simultaneous "pooled" two-sample t-statistics $T^{\mathrm{pool}}_{i;n_1,n_2}$. See the numerical illustrations in Figure 5 associated with $\sigma_{\rho;\theta_{(\varepsilon,e);i}} > 1$.

To circumvent the incoherence of $T^{\mathrm{pool}}_{i;n_1,n_2}$ with $T^{\mathrm{general}}_{i;n_1,n_2}$, particularly in the case of CVQ $>$ 1, we propose an "**adaptively pooled**" version, which follows an approximately $\mathbb{N}(0,1)$ distribution under the null. Following (2.6), a natural choice is given by

$$T^{\mathrm{pool;A}}_{i;n_1,n_2} = \frac{T^{\mathrm{pool}}_{i;n_1,n_2}}{\sigma_{\rho;\widehat{\theta}_{(\varepsilon,e);i}}}, \tag{3.12}$$

where $\widehat{\theta}_{(\varepsilon,e);i} = s^2_{Y;i}/s^2_{X;i}$ serves as an estimate of $\theta_{(\varepsilon,e);i} = \sigma^2_{e;i}/\sigma^2_{\varepsilon;i}$. The simulation results in Section 5 support that the performance of the calibration method applied to the "**adaptively pooled**" choice $\{T^{\mathrm{pool;A}}_{i;n_1,n_2}\}_{i=1}^m$ is comparable to that applied to the "**general**" choice $\{T^{\mathrm{general}}_{i;n_1,n_2}\}_{i=1}^m$.

3.3 Proposed "two-stage" t-test procedure for (2.2)

In practice, $T^{\mathrm{general}}_{i;n_1,n_2}$ and $T^{\mathrm{pool;A}}_{i;n_1,n_2}$ could be skewly distributed under $H_{0,i}$, yielding a slower convergence rate to $\mathbb{N}(0,1)$ and a lower calibration accuracy by $\mathbb{N}(0,1)$. See also Remark 2(ii). For $T^{\mathrm{general}}_{i;n_1,n_2}$, its theoretical form of the skewness-"adjusted" two-sample t-statistic,

$$T^{\mathrm{adjust;T}}_{i;n_1,n_2} = \frac{\left(\overline{X}_i - \overline{Y}_i\right) + \frac{\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2}{6(s^2_{X;i}/n_1 + s^2_{Y;i}/n_2)} + \frac{\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2}{3(s^2_{X;i}/n_1 + s^2_{Y;i}/n_2)^2}\left(\overline{X}_i - \overline{Y}_i\right)^2}{\sqrt{s^2_{X;i}/n_1 + s^2_{Y;i}/n_2}}, \tag{3.13}$$

can be derived from [16], used for the "adjusted" one-sample t-statistic, where

$$\mu_{3,X;i} = \mathrm{E}\{(X_{i,1} - \mu_{X;i})^3\}, \qquad \mu_{3,Y;i} = \mathrm{E}\{(Y_{i,1} - \mu_{Y;i})^3\}. \tag{3.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

A form similar to (3.13) can be found in equation (2.16) of [9]. As expected, $T^{\text{adjust;T}}_{i;n_1,n_2}$ alleviates the skewness effects from $T^{\text{general}}_{i;n_1,n_2}$ and, thus, is more symmetric under $H_{0,i}$. Clearly, if $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0$, then $T^{\text{adjust;T}}_{i;n_1,n_2}$ reduces to $T^{\text{general}}_{i;n_1,n_2}$. Hence, the quantity

$$\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = \mathrm{E}[\{(\overline{X}_i - \overline{Y}_i) - \mathrm{E}(\overline{X}_i - \overline{Y}_i)\}^3], \tag{3.15}$$

serves as a valid measure of skewness of $T^{\text{general}}_{i;n_1,n_2}$, assuming conditions A1–A3. In practice, $T^{\text{adjust;T}}_{i;n_1,n_2}$ is infeasible for the skewness adjustment, because the quantity $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$ is unknown. However, it can be estimated using the sample's third moments, leading to the empirical form of the skewness-"adjusted" two-sample t-statistic,

$$T^{\text{adjust;E}}_{i;n_1,n_2} = \frac{\left(\overline{X}_i - \overline{Y}_i\right) + \frac{\widehat{\mu}_{3,X;i}/n_1^2 - \widehat{\mu}_{3,Y;i}/n_2^2}{6(s^2_{X;i}/n_1 + s^2_{Y;i}/n_2)} + \frac{\widehat{\mu}_{3,X;i}/n_1^2 - \widehat{\mu}_{3,Y;i}/n_2^2}{3(s^2_{X;i}/n_1 + s^2_{Y;i}/n_2)^2}\left(\overline{X}_i - \overline{Y}_i\right)^2}{\sqrt{s^2_{X;i}/n_1 + s^2_{Y;i}/n_2}}, \tag{3.16}$$

where $\widehat{\mu}_{3,X;i} = \sum_{j=1}^{n_1}(X_{i,j} - \overline{X}_i)^3/n_1$ and $\widehat{\mu}_{3,Y;i} = \sum_{j=1}^{n_2}(Y_{i,j} - \overline{Y}_i)^3/n_2$.

With regard to the choice between $T^{\text{general}}_{i;n_1,n_2}$ and $T^{\text{adjust;E}}_{i;n_1,n_2}$, we discuss two cases. If $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0$ exactly or approximately, then $T^{\text{general}}_{i;n_1,n_2}$ is expected to be more symmetrically distributed under $H_{0,i}$ than is $T^{\text{adjust;E}}_{i;n_1,n_2}$, and will outperform $T^{\text{adjust;E}}_{i;n_1,n_2}$ (owing to the variability of sample third moments). On the other hand, if $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$ is far from zero, then $T^{\text{adjust;E}}_{i;n_1,n_2}$ will be effective in correcting the skewness, whereas $T^{\text{general}}_{i;n_1,n_2}$ may not be.

Hence, before selecting $T^{\text{general}}_{i;n_1,n_2}$ or $T^{\text{adjust;E}}_{i;n_1,n_2}$, we first need to assess the adequacy of

$$H^{(1)}_{0,i}: \ \ \mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0. \tag{3.17}$$

Note that (3.14) and (3.17) motivate us to consider the t-statistic

$$\frac{\widehat{\mu}_{3,X;i}/n_1^2 - \widehat{\mu}_{3,Y;i}/n_2^2}{\sqrt{\widehat{\sigma}^2_{3,X;i}/n_1^5 + \widehat{\sigma}^2_{3,Y;i}/n_2^5}}, \tag{3.18}$$

where $\widehat{\sigma}^2_{3,X;i}$ and $\widehat{\sigma}^2_{3,Y;i}$ denote the sample variances of $\{(X_{i,j} - \overline{X}_i)^3\}_{j=1}^{n_1}$ and $\{(Y_{i,j} - \overline{Y}_i)^3\}_{j=1}^{n_2}$, respectively. Under the null hypothesis (3.17), (3.18) $\xrightarrow{\mathcal{D}} \mathbb{N}(0,1)$, by the CLT and Slutsky's theorem, assuming finite sixth moments of $X_{i,1}$ and $Y_{i,1}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

To improve the efficiency of testing (2.2), we propose a "**two-stage**" t-test procedure:

1st**-stage:** For each $i = 1, \ldots, m$, apply the first-stage two-sample t-statistic (3.18) to test, individually, for the null hypothesis $H_{0,i}^{(1)}$ in (3.17).

2nd**-stage:** For each $i = 1, \ldots, m$, define the second-stage two-sample t-statistic $T_{i;n_1,n_2}^{2\_\text{stage}}$ by

$$T_{i;n_1,n_2}^{2\_\text{stage}} = \begin{cases} T_{i;n_1,n_2}^{\text{adjust;E}} \text{ in (3.16)}, & \text{if (3.18) rejects (3.17)}, \\ T_{i;n_1,n_2}^{\text{general}} \text{ in (2.3)}, & \text{if (3.18) retains (3.17)}. \end{cases} \tag{3.19}$$

Use $\{T_{i;n_1,n_2}^{2\_\text{stage}}\}_{i=1}^{m}$ to perform the multiple testing procedure for (2.2).

As illustrated in the simulation studies in Section 5, $T^{\text{adjust;T}}$ always performs best, but is practically infeasible. The proposed T^{2_stage} is as good as the better of T^{general} and $T^{\text{adjust;E}}$.

Remark 4 *For the "**adaptively pooled**" two-sample t-statistic $T_{i;n_1,n_2}^{\text{pool;A}}$, the skewness adjustment is similar to* (3.16) *for $T_{i;n_1,n_2}^{\text{general}}$, except that the denominator is $\sigma_{\rho;\widehat{\theta}_{(\varepsilon,e);i}} s_{\text{pool}_{X;Y};i} \sqrt{1/n_1 + 1/n_2}$.*

4 Error controls allowing dependent data

In practice, dependence in data sets may arise from different tests, between the X-group and Y-group, or within the same X-group or within the same Y-group. Section 4.1 considers the types of dependence between tests, Sections 4.2–4.3 explore models (4.10) and (4.12), respectively, incorporating the dependence structure between two groups and within the same group, respectively. Appendix B discusses extensions of (4.10) and (4.12).

4.1 Dependence between tests

Recall that Propositions 1–2 rely on condition A7, which assumes independence between the test statistics corresponding to the true nulls. Section 4.1.1 evaluates the effect of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

general dependency between the test statistics on the control of the overall significance level; Propositions 4–5 remove condition A7. Section 4.1.2 considers test statistics that are asymptotically jointly Gaussian.

4.1.1 General dependence between tests

Proposition 4 (control $\mathrm{FWER}_1(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$ under general dependence between tests) *Assume model* (2.1) *and that conditions* A1–A6 *hold. For* $\alpha \in (0,1)$, $m_0/m \to \pi_0 \in (0,1]$, $m \to \infty$, *and* $n \to \infty$, *if the general two-sample t-statistics* $\{T^{\mathrm{general}}_{i;n_1,n_2}\}_{i=1}^m$ *are used, with* $\mathrm{t}^{\mathrm{a}}_{\alpha;m}$ *given in* (3.2) *and* (m,n) *satisfying* (3.1), *then*

$$\mathrm{FWER}_1(\mathrm{t}^{\mathrm{a}}_{\alpha;m}) \le \pi_0\beta_{1;\alpha} + o(1), \tag{4.1}$$

where $\beta_{1;\alpha} = -\log(1-\alpha)$.

In view of (4.1), the limiting overall significance level continues to be bounded by the nominal level α, for any $\pi_0 \le \alpha/\beta_{1;\alpha}$, when m tests are allowed to be dependent. See the left panel of Figure 2 for the plot of $\alpha/\beta_{1;\alpha}$ with respect to α. For example, a level $\alpha = 0.05$ allows any choice of π_0 in the range $(0, 0.9748]$, which is wide enough for realistic applications. Interestingly, even in the special case of $\pi_0 = 1$ (which is rare, in practice), that $\pi_0\beta_{1;\alpha} = \beta_{1;\alpha}$ and $\alpha \le \beta_{1;\alpha}$ (with a negligible difference between α and $\beta_{1;\alpha}$, particularly when α is small, as illustrated in the right panel of Figure 2) indicates that the critical value $\mathrm{t}^{\mathrm{a}}_{\alpha;m}$ in (3.2) offers an asymptotically slightly conservative $\beta_{1;\alpha}$ for the resulting $\mathrm{FWER}(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$.

Proposition 5 states that, when $k = 1$, the upper bound achievable for $\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k})$ reduces to that for $\mathrm{FWER}(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$.

Proposition 5 (control $\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k})$ under general dependence between tests) *Assume model* (2.1) *and that conditions* A1–A6 *hold. For* $k \ge 2$, $\alpha \in (0,1)$, $m_0/m \to \pi_0 \in$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

$(0, 1]$, $m \to \infty$, *and* $n \to \infty$, *if the general two-sample t-statistics* $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$ *are used, with* $\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}$ *given in* (3.4) *and* (m, n) *satisfying* (3.1), *then*

$$\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k}) \le \pi_0 \beta_{k;\alpha}/k + o(1), \tag{4.2}$$

where $\beta_{k;\alpha}$ *solves* (3.5).

Compared with Proposition 2, the upper bound $\pi_0\beta_{k;\alpha}/k$ in (4.2), with $k \ge 2$, is controlled by the nominal level α only when the proportion π_0 does not exceed $\alpha/(\beta_{k;\alpha}/k)$, which is equal to 0.2821 for $\alpha = 0.05$ and $k = 2$. In the extreme case of $\pi_0 = 1$, we can show that $\pi_0\beta_{k;\alpha}/k = \beta_{k;\alpha}/k$ is invariably at least as large as α. This reflects the cost of generalizing Proposition 2 from mutually independent tests to cases allowing for general dependency.

4.1.2 Jointly Gaussian distributed test statistics

Consider a specific factor model for observations $\{X_{i,j}\}$ and $\{Y_{i,j}\}$:

$$\begin{aligned} X_{i,j} &= \mu_{X;i} + \boldsymbol{\beta}_{X;i}^T \boldsymbol{u}_j + \varepsilon_{i,j}, \quad 1 \le i \le m,\ 1 \le j \le n_1, \\ Y_{i,j} &= \mu_{Y;i} + \boldsymbol{\beta}_{Y;i}^T \boldsymbol{v}_j + e_{i,j}, \quad 1 \le i \le m,\ 1 \le j \le n_2, \end{aligned} \tag{4.3}$$

where $\boldsymbol{u}_j$ are unobserved d_u-dimensional random vectors, with $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_1}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(\boldsymbol{0}, \Sigma_{\boldsymbol{u}})$; $\boldsymbol{v}_j$ are unobserved d_v-dimensional random vectors, with $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_2}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(\boldsymbol{0}, \Sigma_{\boldsymbol{v}})$; and $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_1})$ and $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_2})$ are independent. For example, the gene expressions $\{X_{i,j} : 1 \le i \le m\}$ of the jth subject may be influenced by common factors $\boldsymbol{u}_j$, for example, the age or other variables of the jth subject. In addition, assume $\{\varepsilon_{i,j}\}$ and $\{e_{i,j}\}$ are identical to those in model (4.10); $\{\varepsilon_{i,j}\}$, $\{e_{i,j}\}$, $\{\boldsymbol{u}_j\}$, and $\{\boldsymbol{v}_j\}$ are independent.

For model (4.3), the dependence between the two-sample t-statistics,

$$T^{\text{general}}_{i;n_1,n_2} = \frac{(\mu_{X;i} - \mu_{Y;i}) + (\overline{\varepsilon}_i - \overline{e}_i) + (\boldsymbol{\beta}_{X;i}^T \overline{\boldsymbol{u}} - \boldsymbol{\beta}_{Y;i}^T \overline{\boldsymbol{v}})}{\sqrt{s^2_{X;i}/n_1 + s^2_{Y;i}/n_2}}, \qquad i = 1, \ldots, m, \tag{4.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

is caused by factors $\overline{\boldsymbol{u}} = \sum_{j=1}^{n_1} \boldsymbol{u}_j/n_1$ and $\overline{\boldsymbol{v}} = \sum_{j=1}^{n_2} \boldsymbol{v}_j/n_2$, which are common to all tests. It follows that the two-sample t-statistics can be rewritten as

$$(T^{\text{general}}_{1;n_1,n_2}, \ldots, T^{\text{general}}_{m;n_1,n_2})^T = \boldsymbol{D} \bullet \boldsymbol{U}, \tag{4.5}$$

where $\boldsymbol{D} = (\overline{X}_1 - \overline{Y}_1, \ldots, \overline{X}_m - \overline{Y}_m)^T$; the operator $\bullet$ in (4.5) indicates component-wise multiplication; and $\boldsymbol{U} = (U_1, \ldots, U_m)^T$, with $U_i = (s^2_{X;i}/n_1 + s^2_{Y;i}/n_2)^{-1/2}$. For fixed m, the CLT gives

$$\sqrt{n_1 + n_2}\ \boldsymbol{D} \xrightarrow{\mathcal{D}} (W_1, \ldots, W_m)^T, \tag{4.6}$$

as $n_1 \to \infty$ and $n_2 \to \infty$, where $(W_1, \ldots, W_m)^T \sim \mathbb{N}(\boldsymbol{\nu}, \Omega)$, for some $\boldsymbol{\nu} \in \mathbb{R}^m$ and positive-definite matrix $\Omega = (\omega_{ij})_{1 \le i,j \le m}$. Similarly, the law of large numbers gives $s^2_{X;i} \xrightarrow{\text{P}} \boldsymbol{\beta}^T_{X;i} \Sigma_{\boldsymbol{u}} \boldsymbol{\beta}_{X;i} + \sigma^2_{\varepsilon;i}$ and $s^2_{Y;i} \xrightarrow{\text{P}} \boldsymbol{\beta}^T_{Y;i} \Sigma_{\boldsymbol{v}} \boldsymbol{\beta}_{Y;i} + \sigma^2_{e;i}$, implying

$$(n_1 + n_2)^{-1/2}\ U_i \xrightarrow{\text{P}} c_i, \quad 1 \le i \le m, \tag{4.7}$$

where $c_i = \{(\boldsymbol{\beta}^T_{X;i} \Sigma_{\boldsymbol{u}} \boldsymbol{\beta}_{X;i} + \sigma^2_{\varepsilon;i})/\rho + (\boldsymbol{\beta}^T_{Y;i} \Sigma_{\boldsymbol{v}} \boldsymbol{\beta}_{Y;i} + \sigma^2_{e;i})/(1-\rho)\}^{-1/2}$; thus,

$$(n_1 + n_2)^{-1/2}\ \boldsymbol{U} \xrightarrow{\text{P}} \boldsymbol{c}, \tag{4.8}$$

with $\boldsymbol{c} = (c_1, \ldots, c_m)^T$. By Slutsky's theorem [10], (4.5), (4.6), and (4.8) imply that

$$(T^{\text{general}}_{1;n_1,n_2}, \ldots, T^{\text{general}}_{m;n_1,n_2})^T \xrightarrow{\mathcal{D}} (Z_1, \ldots, Z_m)^T \sim \mathbb{N}(\widetilde{\boldsymbol{\nu}}, \widetilde{\Omega}), \tag{4.9}$$

where $Z_i = c_i\, W_i$, $\widetilde{\boldsymbol{\nu}} = \boldsymbol{c} \bullet \boldsymbol{\nu}$, and $\widetilde{\Omega} = (c_i\, c_j\, \omega_{ij})_{1 \le i,j \le m}$.

The joint Gaussianity of the test statistics in (4.9) makes it feasible to apply the factor model method in [14] to decompose $\widetilde{\Omega}$, and then to control the false discovery proportion (FDP; defined as the number of false rejections divided by the number of rejections) and FDR asymptotically. On the other hand, this method relies on knowing $\widetilde{\Omega}$ in advance. Thus, we need the techniques used to estimate high-dimensional covariance matrices to estimate $\widetilde{\Omega}$. Our Gaussian calibration helps to simplify its diagonal entries to ones.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

4.2 Dependence between groups and within a group: Model I

Consider observations $\{X_{i,j}\}$ and $\{Y_{i,j}\}$ following Model I,

$$\begin{aligned} X_{i,j} &= \mu_{X;i} + \varepsilon_{i,j} + w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_1, \\ Y_{i,j} &= \mu_{Y;i} + e_{i,j} + w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_2, \end{aligned} \tag{4.10}$$

where the errors $\{w_1, \ldots, w_m\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_w^2)$, with $\sigma_w^2 \in (0, \infty)$. For each i, the errors $\{\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_{\varepsilon;i}^2)$, the errors $\{e_{i,1}, \ldots, e_{i,n_2}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_{e;i}^2)$, and $\{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}), (e_{i,1}, \ldots, e_{i,n_2}), w$ are mutually independent. Furthermore, $\{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}; e_{i,1}, \ldots, e_{i,n_2}; w_i) : i \in \mathcal{I}_0\}$ are independent. It follows that the two-sample t-statistics reduce to the following forms:

$$T_{i;n_1,n_2}^{\text{general}} = \frac{\overline{\varepsilon}_i - \overline{e}_i}{\sqrt{s_{\varepsilon;i}^2/n_1 + s_{e;i}^2/n_2}}, \quad T_{i;n_1,n_2}^{\text{pool}} = \frac{\overline{\varepsilon}_i - \overline{e}_i}{s_{\text{pool}_{\varepsilon;e};i}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad T_{i;n_1,n_2}^{\text{pool;A}} = \frac{T_{i;n_1,n_2}^{\text{pool}}}{\sigma_{\rho;\widehat{\theta}_{(\varepsilon,e);i}}}. \tag{4.11}$$

Note that this data set involves dependence between different groups, and within the same group; however, the test statistics (using $\{T_{i;n_1,n_2}^{\text{general}}\}$, $\{T_{i;n_1,n_2}^{\text{pool}}\}$ or $T_{i;n_1,n_2}^{\text{pool;A}}$) associated with the true nulls are independent. Moreover, Model I in the case of $\sigma_w^2 = 0$ reduces to the counterpart of model (2.1).

With regard to Model I, we can show two distributional results for the "**general**" two-sample t-statistic $T_{i;n_1,n_2}^{\text{general}}$ under $H_{0,i}$:

(c1) if $\sigma_{\varepsilon;i}^2 = \sigma_{e;i}^2$ and $n_1 = n_2$, then $T_{i;n_1,n_2}^{\text{general}} \sim t_{2n_1-2}$;

(c2) if $n_1 \to \infty$ and $n_2 \to \infty$, then $T_{i;n_1,n_2}^{\text{general}} \xrightarrow{\mathcal{D}} \mathbb{N}(0,1)$.

Hence, the conclusions of Propositions 1–2 carry through to the "**general**" two-sample t-statistics $\{T_{i;n_1,n_2}^{\text{general}}\}_{i=1}^m$.

As a comparison, for the "pooled" two-sample t-statistic $T_{i;n_1,n_2}^{\text{pool}}$ under $H_{0,i}$, we draw the following two conclusions:

(d1) If $\sigma_{\varepsilon;i}^2 = \sigma_{e;i}^2$, then $T_{i;n_1,n_2}^{\text{pool}} \sim t_{n_1+n_2-2}$. In this case, the results in Propositions 1–2 continue to apply for the "pooled" choice $\{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^m$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

(d2) If $n_1 \to \infty$ and $n_2 \to \infty$, such that $n_1/(n_1+n_2) \to \rho \in (0,1)$, then (2.6) gives $T^{\text{pool}}_{i;n_1,n_2} \xrightarrow{\mathcal{D}} \mathbb{N}(0, \sigma^2_{\rho;\theta_{(\varepsilon,e);i}})$. Similarly to the discussion in Section 3.2, there is no guarantee in the case of $\sigma_{\rho;\theta_{(\varepsilon,e);i}} > 1$ that we can achieve level bounds α in (2.11) and (2.12) using $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$.

However, according to (4.11), the "**adaptively pooled**" version satisfies $T^{\text{pool;A}}_{i;n_1,n_2} \xrightarrow{\mathcal{D}} \mathbb{N}(0,1)$ and, thus, the $\mathbb{N}(0,1)$ calibration remains valid for $\{T^{\text{pool;A}}_{i;n_1,n_2}\}_{i=1}^m$.

4.3 Dependence between groups and within a group: Model II

Consider an alternative model similar to Model I, except that the signs of the error terms $w_i/2$ in $X_{i,j}$ are negative, yielding Model II:

$$\begin{aligned} X_{i,j} &= \mu_{X;i} + \varepsilon_{i,j} - w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_1, \\ Y_{i,j} &= \mu_{Y;i} + e_{i,j} + w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_2. \end{aligned} \tag{4.12}$$

Model (4.12) is motivated from a two-sample microarray testing example in Section 4 of [11] and Section 6.4 of [12] with $n_1 = n_2$, where w_i are small disturbances caused by unequal effects of unobserved covariates on the X-group and Y-group. The explicit forms of the two-sample t-statistics are derived as follows:

$$T^{\text{general}}_{i;n_1,n_2} = \frac{\overline{\varepsilon}_i - \overline{e}_i - w_i}{\sqrt{s^2_{\varepsilon;i}/n_1 + s^2_{e;i}/n_2}}, \quad T^{\text{pool}}_{i;n_1,n_2} = \frac{\overline{\varepsilon}_i - \overline{e}_i - w_i}{s_{\text{pool}_{\varepsilon;e};i}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \quad T^{\text{pool;A}}_{i;n_1,n_2} = \frac{T^{\text{pool}}_{i;n_1,n_2}}{\sigma_{\rho;\widehat{\theta}_{(\varepsilon,e);i}}}, \tag{4.13}$$

which differ from those in (4.11). Again, dependence between and within groups exist in the data set, where the extent of the dependence is captured by the magnitude of σ^2_w, but the two-sample t-statistics associated with the true nulls remain independent.

In the context of Model II, we can show two results for the null distribution of the "**general**" two-sample t-statistic $T^{\text{general}}_{i;n_1,n_2}$:

(e1) if $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i} = \sigma^2_i$ and $n_1 = n_2$, then $T^{\text{general}}_{i;n_1,n_2} \sim t_{2n_1-2} \times f_1$, where $f_1 = \sqrt{1 + \frac{n_1}{2}\frac{\sigma^2_w}{\sigma^2_i}}$;

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

(e2) if $n_1 \to \infty$ and $n_2 \to \infty$, then

$$T^{\text{general}}_{i;n_1,n_2} = Z \times f_2\{1 + o_{\text{P}}(1)\} \xrightarrow{\text{P}} \infty, \quad \text{where } f_2 = \sqrt{1 + \frac{n_1 n_2 \sigma_w^2}{n_2 \sigma_\varepsilon^2 + n_1 \sigma_e^2}}, \tag{4.14}$$

where $Z \sim \mathbb{N}(0,1)$ and $\xrightarrow{\text{P}}$ denotes convergence in probability.

We can also show that $T^{\text{pool;A}}_{i;n_1,n_2}$ has the same limit null distribution as $T^{\text{general}}_{i;n_1,n_2}$. For the null distribution of the "pooled" two-sample t-statistic $T^{\text{pool}}_{i;n_1,n_2}$, we draw two conclusions:

(f1) If $\sigma^2_{\varepsilon;i} = \sigma^2_{e;i} = \sigma^2_i$, then $T^{\text{pool}}_{i;n_1,n_2} \sim t_{n_1+n_2-2} \times f_3$, where $f_3 = \sqrt{1 + \frac{n_1 n_2}{n_1+n_2}\frac{\sigma_w^2}{\sigma_i^2}}$.

(f2) If $n_1 \to \infty$ and $n_2 \to \infty$, such that $n_1/(n_1+n_2) \to \rho \in (0,1)$, then

$$T^{\text{pool}}_{i;n_1,n_2} = Z \times f_4\{1 + o_{\text{P}}(1)\} \xrightarrow{\text{P}} \infty, \quad \text{where } f_4 = \sqrt{\frac{(1-\rho) + \rho\sigma_e^2/\sigma_\varepsilon^2 + \frac{n_1 n_2}{n_1+n_2}\frac{\sigma_w^2}{\sigma_\varepsilon^2}}{\rho + (1-\rho)\sigma_e^2/\sigma_\varepsilon^2}}. \tag{4.15}$$

Thus, the conclusions of Propositions 1–2 fail for the two-sample t-statistics $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$, because the factor f_2 in (4.14) invariably exceeds one. As a comparison, Propositions 1–2 may fail for $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$, particularly when the factor f_4 in (4.15) substantially exceeds one. In the case of $f_2 > f_4$, the "**adaptively pooled**" versions $\{T^{\text{pool;A}}_{i;n_1,n_2}\}_{i=1}^m$ do not ameliorate $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$.

5 Simulation study

We assess the finite-sample performance of the calibration method applied to the two-sample t-test statistics $\{T^{\text{general}}_{i;n_1,n_2}\}_{i=1}^m$, $\{T^{\text{pool}}_{i;n_1,n_2}\}_{i=1}^m$, $\{T^{\text{pool;A}}_{i;n_1,n_2}\}_{i=1}^m$, $\{T^{\text{adjust;T}}_{i;n_1,n_2}\}_{i=1}^m$, $\{T^{\text{adjust;E}}_{i;n_1,n_2}\}_{i=1}^m$, and $\{T^{\text{2-stage}}_{i;n_1,n_2}\}_{i=1}^m$, as the total sample size $n = n_1+n_2$ varies. For each $k \in \{1, 2\}$, we conduct the simulation 1000 times. In each simulation, we calculate the numbers of false rejections $V_m(\text{t}^{\text{a}}_{\alpha;m})$ and $V_m(\text{t}^{\text{a}}_{\alpha;m;k})$. The empirical estimates of $\text{FWER}(\text{t}^{\text{a}}_{\alpha;m})$ and $\text{FWER}_k(\text{t}^{\text{a}}_{\alpha;m;k})$ are the proportion of times that $\{V_m(\text{t}^{\text{a}}_{\alpha;m}) \geq 1\}$ and $\{V_m(\text{t}^{\text{a}}_{\alpha;m;k}) \geq k\}$, respectively, occur in the 1000 simulations. Set $\alpha = 0.05$ as the control level. The "**two-stage**" t-tests use level

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

0.05 in the first-stage. A range of sample sizes are considered, with $n_1 = 10c$ and $n_2 = 20c$, for $c \in \{1, 2, \ldots, 10\}$, yielding the combined sample size $n = 30c$. We set $m = 10000$, with $\pi_0 = m_0/m = 0.9$.

To generate data under either independence or dependence, we consider the model

$$\begin{aligned} X_{i,j} &= \mu_{X;i} + \varepsilon_{i,j} + \mathrm{sign}_{X;i}\, w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_1, \\ Y_{i,j} &= \mu_{Y;i} + e_{i,j} + \mathrm{sign}_{Y;i}\, w_i/2, \quad 1 \le i \le m,\ 1 \le j \le n_2, \end{aligned} \tag{5.1}$$

where $\mu_{X;i} = \mu_{Y;i} = 1$, for $i = 1, \ldots, m_0$. The values of $\mu_{X;i}$ and $\mu_{Y;i}$ are simulated from Uniform$(0.75, 1.25)$ and Uniform$(1.75, 2.25)$, respectively, for $i = m_0 + 1, \ldots, m$, and $\{\varepsilon_{i,j}\}$ are independent of $\{e_{i,j}\}$. In addition, the errors $\{w_1, \ldots, w_m\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma_w^2)$, as described below (4.10). Note that (5.1) includes models (2.1), (4.10), and (4.12):

if $\mathrm{sign}_{X;i} \equiv 0$ and $\mathrm{sign}_{Y;i} \equiv 0$, then model (5.1) reduces to model (2.1);

if $\mathrm{sign}_{X;i} \equiv +1$ and $\mathrm{sign}_{Y;i} \equiv +1$, then model (5.1) is Model I in (4.10);

if $\mathrm{sign}_{X;i} \equiv -1$ and $\mathrm{sign}_{Y;i} \equiv +1$, then model (5.1) is Model II in (4.12).

In model (5.1), the schemes for the errors $\{\varepsilon_{i,j}\}$ and $\{e_{i,j}\}$ are considered in **Examples** 1–5, as follows: **Example** 1: $\{\varepsilon_{i,j}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma^2)$, $\{e_{i,j}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, \sigma^2)$, with $\sigma = 1.0$; **Example** 2: $\{\varepsilon_{i,j}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, 1)$, $\{e_{i,j}\} \overset{\text{i.i.d.}}{\sim} t_4$; **Example** 3: $\{\varepsilon_{i,j}\} \overset{\text{i.i.d.}}{\sim} t_4$, $\{e_{i,j}\} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0, 1)$; **Example** 4: $\{\varepsilon_{i,j}\} \overset{\text{i.i.d.}}{\sim} \chi_2^2 - 2$, $\{e_{i,j}\} \overset{\text{i.i.d.}}{\sim} -(\chi_2^2 - 2)$; and **Example** 5: $\{\varepsilon_{i,j}\} \overset{\text{i.i.d.}}{\sim} \chi_4^2 - 4$, $e_{i,j} = (2b_i - 1)\, u_{i,j}$, where $\{u_{i,j}\} \overset{\text{i.i.d.}}{\sim} \{\mathrm{Exp}(1/\lambda) - \lambda\}$, with $\lambda = 4$, and the coefficients b_i are nonrandom and equal to the sampled values of b_i^*, with $\{b_1^*, \ldots, b_m^*\} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2)$. Here, **Examples** 4 and 5 assess the skewness effects of the two-sample t-tests on the calibration methods.

Moreover, in model (5.1), the variances σ_w^2 of the errors $\{w_i\}$ are considered for

model (2.1) with $\sigma_w = 0$; Model I with $\sigma_w = 0.5$; Model II with $\sigma_w = 0.1$.

Thus, the combination of errors $\{\varepsilon_{i,j}, e_{i,j}\}$ and errors $\{w_i\}$ in model (5.1) yields 15 examples,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

denoted as follows:

'Example 1',	..., **'Example** 5':	for independent data;
'Example 1(I)',	..., **'Example** 5(I)':	for dependent data;
'Example 1(II)',	..., **'Example** 5(II)':	for dependent data.

Graphical illustrations are displayed in Figures 3–7 for the empirical estimates of $\mathrm{FWER}(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$, in Figures 8–12 for the empirical estimates of $\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k})$ with $k = 2$, and in Figure 13 for the calculated FDP of the BH procedure.

5.1 Independent data

Recall that **Examples** 1–5 correspond to independent data. Table 1 summarizes the information on the CVQ and skewness of the error terms.

In **Example** 1 with Gaussian errors, the top row of Figure 3 indicates that the estimated $\mathrm{FWER}(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$ of $\{T^{\mathrm{general}}_{i;n_1,n_2}\}$ gets closer to 0.05 as the sample size n increases. The $\mathbb{N}(0,1)$ calibration applied to $\{T^{\mathrm{pool}}_{i;n_1,n_2}\}$ performs similarly to that of $\{T^{\mathrm{general}}_{i;n_1,n_2}\}$, owing to the equal population variances, such that $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} = 1$ in **Example** 1. In this case, there is also no adverse effect of using the "**adaptively pooled**" version $\{T^{\mathrm{pool;A}}_{i;n_1,n_2}\}$. The calibration methods applied to $\{T^{\mathrm{adjust;T}}_{i;n_1,n_2}\}$, $\{T^{\mathrm{adjust;E}}_{i;n_1,n_2}\}$, and $\{T^{\mathrm{2\_stage}}_{i;n_1,n_2}\}$ perform similarly to that applied to $\{T^{\mathrm{general}}_{i;n_1,n_2}\}$, owing to the symmetric distributions of $\{\varepsilon_{i,j}\}$ and $\{e_{i,j}\}$.

In addition, recall from part (b1) in Section 2.1 that $T^{\mathrm{pool}}_{i;n_1,n_2}$ in **Example** 1 exactly follows the $t_{n_1+n_2-2}$-distribution under the null. The second columns of Figures 3 and 8 overlay the true values (using red lines) of $\mathrm{FWER}(\mathrm{t}^{\mathrm{a}}_{\alpha;m})$ and $\mathrm{FWER}_k(\mathrm{t}^{\mathrm{a}}_{\alpha;m;k})$, respectively, which match well with their empirical counterparts. This supports the validity of the simulations. Similarly, the left column of Figure 13 compares the FDP of the BH multiple testing procedure [1], implemented as follows: the approximate p-values calculated from the approximate $\mathbb{N}(0,1)$-distributions for $T^{\mathrm{general}}_{i;n_1,n_2}$, $T^{\mathrm{pool}}_{i;n_1,n_2}$, $T^{\mathrm{pool;A}}_{i;n_1,n_2}$, $T^{\mathrm{adjust;T}}_{i;n_1,n_2}$, $T^{\mathrm{adjust;E}}_{i;n_1,n_2}$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

$T^{2\_\text{stage}}_{i;n_1,n_2}$, and the exact p-values calculated from the exact $t_{n_1+n_2-2}$-distribution for $T^{\text{pool}}_{i;n_1,n_2}$. As shown, when n approaches 100 (or more), the FDPs using the $\mathbb{N}(0,1)$ calibration mimic that using the exact distribution.

In **Examples** 2–3, with nonGaussian errors, the population variances are $\sigma^2_{\epsilon;i} < \sigma^2_{e;i}$ in **Example** 2, and $\sigma^2_{\epsilon;i} > \sigma^2_{e;i}$ in **Example** 3. Figures 4 and 5 indicate that within each example, there is little difference in the performance of the calibration methods applied to the test statistics $\{T^{\text{general}}_{i;n_1,n_2}\}$, $\{T^{\text{pool;A}}_{i;n_1,n_2}\}$, $\{T^{\text{adjust;T}}_{i;n_1,n_2}\}$, and $\{T^{2\_\text{stage}}_{i;n_1,n_2}\}$. However, $\{T^{\text{pool}}_{i;n_1,n_2}\}$ behaves substantially differently in **Example** 2 and **Example** 3, where the FWERs are conservatively controlled in **Example** 2 (as seen in the top row, second column panel of Figure 4), but are not controlled in **Example** 3 (as seen in the top row, second column panel of Figure 5, even if n increases). Again, the difference is caused by the quantity $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} < 1$ in **Example** 2, with $n_1 < n_2$ and $\sigma^2_{\epsilon;i} < \sigma^2_{e;i}$, whereas $\sigma^2_{\rho;\theta_{(\varepsilon,e);i}} > 1$ in **Example** 3, with $n_1 < n_2$ and $\sigma^2_{\epsilon;i} > \sigma^2_{e;i}$. The comparison thus supports that the "**adaptively pooled**" version $T^{\text{pool;A}}_{i;n_1,n_2}$ is a valid substitute for the originally "pooled" version $T^{\text{pool}}_{i;n_1,n_2}$, and that its performance compares with that of the "**general**" version $T^{\text{general}}_{i;n_1,n_2}$.

Moreover, in **Example** 3, because the sixth moment does not exist for the t_4-distribution, $\widehat{\mu}_{3,X}/n_1^2 - \widehat{\mu}_{3,Y}/n_2^2$ performs poorly in estimating $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$. Thus, $T^{\text{adjust;E}}_{i;n_1,n_2}$ deviates significantly from $T^{\text{adjust;T}}_{i;n_1,n_2}$, as seen in Figure 5. Nonetheless, $T^{2\_\text{stage}}_{i;n_1,n_2}$ is as good as $T^{\text{general}}_{i;n_1,n_2}$.

Recall that for **Examples** 1–3, $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0$ (as shown in Table 1). Thus $T^{\text{adjust;T}}_{i;n_1,n_2}$ and $T^{\text{general}}_{i;n_1,n_2}$ are identical and the best, and $T^{2\_\text{stage}}_{i;n_1,n_2}$ compares well with $T^{\text{general}}_{i;n_1,n_2}$. As a comparison, **Examples** 4–5 assess the utility of the proposed "**two-stage**" t-test procedure in the presence of skewness. In **Example** 4, $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$ is relatively large. Figure 6 reveals that $T^{\text{adjust;E}}_{i;n_1,n_2}$ is better than $T^{\text{general}}_{i;n_1,n_2}$, and $T^{2\_\text{stage}}_{i;n_1,n_2}$ is close to the better of $T^{\text{general}}_{i;n_1,n_2}$ and $T^{\text{adjust;E}}_{i;n_1,n_2}$. The theoretical $T^{\text{adjust;T}}_{i;n_1,n_2}$ still controls the FWER in the best way.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

In **Example** 5, $\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$ depends on whether $b_i = 0$ or 1, as given in Table 1. In this case, we observe from Figure 7 that $T_{i;n_1,n_2}^{2\text{-stage}}$ outperforms both $T_{i;n_1,n_2}^{\text{general}}$ and $T_{i;n_1,n_2}^{\text{adjust;E}}$.

5.2 Dependent data

For Model I associated with the dependence mechanism in **Examples** ℓ(I), for $\ell = 1, \ldots, 5$, it is apparent that the top and middle rows of Figures 3–12 are nearly indistinguishable, regardless of the magnitude of $\sigma_w > 0$. This agrees with the analysis in Section 4.2. By the same argument, the calculated FDPs of the BH procedure in the left column of Figure 13 resemble those in the middle column of Figure 13.

In striking contrast, for **Examples** ℓ(II), for $\ell = 1, \ldots, 5$, with a dependence mechanism described by Model II, the loss of control over FWER_1 and FWER_2 is noticeable in the bottom rows of Figures 3–7 and Figures 8–12, even if σ_w is as low as 0.1, lending support to the discussion in Section 4.3. The right column of Figure 13 shows that the FDPs based on the $\mathbb{N}(0, 1)$ calibration for approximating the p-values no longer mimic the actuals proportions. Again, this is because when the data are generated from Model II, the variances in the asymptotic distributions of $T_{i;n_1,n_2}^{\text{general}}$ (as well as $T_{i;n_1,n_2}^{\text{pool;A}}$) and $T_{i;n_1,n_2}^{\text{pool}}$ escalate by factors f_2 in (4.14) and f_4 in (4.15), respectively. As anticipated, the exact $t_{n_1+n_2-2}$ calibration, available for $T_{i;n_1,n_2}^{\text{pool}}$ in **Example** 1(II), continues to perform well.

6 Real-data examples

We apply the Gaussian calibration for two-sample t-tests to analyze three real-data sets. As expected, Table 2 reveals a discrepancy between the results delivered by the "pooled" and "**general**" versions. Nonetheless, the results based on the "**adaptively pooled**" version always agree well with those of the "**general**" version. This lends further support

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

to the superiority of the "**adaptively pooled**" version to the "pooled" version in statistical practice. The proposed "**two-stage**" procedure resembles the "**general**" version.

First, we analyze the prostate cancer data set of [12], which contains genetic expression levels for 6033 genes, obtained for 102 men, comprising 50 normal control subjects and 52 prostate cancer patients. The primary goal of this study was to discover a small number of "interesting" genes that have expression levels that differ between the prostate and normal subjects. Using the BH multiple-testing procedure, Table 2 compares the number of genes detected as significant, where the p-values are calculated from the $\mathbb{N}(0,1)$-distribution for $T^{\text{general}}_{i;n_1,n_2}$, $t_{n_1+n_2-2}$-distribution for $T^{\text{pool}}_{i;n_1,n_2}$, and $\mathbb{N}(0,1)$-distributions for $T^{\text{pool}}_{i;n_1,n_2}$, $T^{\text{pool;A}}_{i;n_1,n_2}$, and $T^{\text{2\_stage}}_{i;n_1,n_2}$. Recall that the simulation studies in Figure 13 support the Gaussian calibration used in the BH procedure with independent data, with the combined sample size n around 100 and m as large as 10000. The difference between the detected numbers 21 (using the t-distribution) and 51 and 50 (using the $\mathbb{N}(0,1)$ calibration methods) could be caused by the nonGaussian samples or the unequal population variances; as a result, $T^{\text{pool}}_{i;n_1,n_2}$ may not follow the $t_{n_1+n_2-2}$-distribution.

Second, we apply the calibration method to the gene expression data produced by [17] in a study on prostate cancer progression. The study aims to identify genes that show evidence of differential expression in cancerous tumors. The data set includes gene expressions for $m = 8648$ genes using prostate cell populations from low-grade ($n_1 = 27$) and high-grade ($n_2 = 17$) samples of cancerous tissue. Using the BH multiple-testing procedure, where p-values are calculated from the $\mathbb{N}(0,1)$-distribution for $T^{\text{general}}_{i;n_1,n_2}$, $t_{n_1+n_2-2}$-distribution for $T^{\text{pool}}_{i;n_1,n_2}$, and $\mathbb{N}(0,1)$-distributions for $T^{\text{pool}}_{i;n_1,n_2}$, $T^{\text{pool;A}}_{i;n_1,n_2}$, and $T^{\text{2\_stage}}_{i;n_1,n_2}$, the numbers of genes declared to be significant are 565, 196, 436, 563, and 565, respectively; see Table 2. In this example, the detection difference between using the t-distribution and using the approximate $\mathbb{N}(0,1)$-distribution could be caused by the nonGaussian samples or the unequal

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

population variances; as a result, the $t_{n_1+n_2-2}$-distribution may not be valid for $T^{\text{pool}}_{i;n_1,n_2}$. The difference may also be because the sample size $n = 44$ is not large enough for the Gaussian calibration. Interestingly, the "**adaptively pooled**" two-sample t-statistics $\{T^{\text{pool;A}}_{i;n_1,n_2}\}$ continue to detect a comparable number of significant genes to those of its "**general**" counterparts $\{T^{\text{general}}_{i;n_1,n_2}\}$.

As a third illustration, we analyze the Acute Lymphoblastic Leukemia (ALL) data set. Refer to [5] for details of the ALL data set, containing data on 12625 genes measured for two groups of samples sizes, 37 and 42. Table 2 presents the number of genes differentially expressed in the BCR/ABL versus NEG comparison for the four methods. The "pooled" two-sample t-statistics $T^{\text{pool}}_{i;n_1,n_2}$ using the $t_{n_1+n_2-2}$-distribution identify 169 genes (identical to that given in Table S2 of [5]), which differs from the results of the other four calibration methods. Again, we observe that the numbers of genes identified by the "**two-stage**," "**adaptively pooled**," and "**general**" two-sample t-statistics are comparable.

7 Discussion

We have examined the validity of a calibration method used simultaneously in two-sample t-tests, the exact distributions of which are typically unknown in many practical applications. In that instance, the inaccuracy of the distributional approximation, associated with realistic samples sizes n_1 and n_2 will degrade the overall significance level, ultimately limiting the effective number of tests m. The relationship between m and (n_1, n_2) is studied to ensure control of the overall level accuracy, as well as to control the FDR for some multiple-testing procedures. A distinction is made between the choice of "**general**" and "pooled" two-sample t-statistics in cases where the typical form of the independence assumption between tests either holds or is violated. The proposed "**adaptively pooled**" two-sample t-statistics, when used simultaneously in the calibration method, perform as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

well as the simultaneous "**general**" version, whereas the original "pooled" version may behave abnormally. The proposed "**two-stage**" procedure compares well with the above methods when the errors are symmetric, but outperforms the others when the errors are skewed and is less sensitive to error asymmetry.

Simulation studies demonstrate that under appropriate independence assumptions, the calculated FDPs of some conventional multiple-testing procedures, such as the BH procedure, can be controlled when the p-values are approximated using the calibrated distribution for the "**general**," "**two-stage**," and "**adaptively pooled**" two-sample t-statistics.

The dependence structure poses challenges related to controlling the overall significance level and FDR. In Section 4, we demonstrated that the FWER and FWER_k can be controlled under arbitrary dependence between tests, but that the FDR would not be controlled if we simply followed the same procedure in Section 3 without any modification. To deal with the jointly Gaussian distributed test statistics, we introduce the factor model to decompose these dependent test statistics into nearly independent test statistics, such that the FDP and FDR can both be controlled asymptotically. In addition, we addressed explicitly the performance of the "**general**," "pooled," and "**adaptively pooled**" two-sample t-statistics in the more interesting and practically motivated models (4.10), (4.12), and (B.1), allowing dependence between and within groups.

Several issues are left to future research. First, the bootstrap method provides an alternative method for the calibrated distribution of the two-sample t-tests, potentially relaxing $\log(m) = o(n^{1/3})$ to $\log(m) = o(n^{1/2})$, at the expense of requiring more technical restrictions and a much heavier computational cost. Second, the power of a given multiple-testing procedure can be improved when the p-values need to be approximated, and should be studied on a case-by-case basis. Third, in Propositions 1, 2, 4, and 5, the condition $\pi_0 \in (0, 1]$ excludes $\pi_0 = 0$, which is the case of "dense true non-nulls." In practice,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

information on m_0 or π_0 can be learned from prior knowledge or estimated using empirical procedures [3, 18, 24]. If the resulting π_0 is close to zero, it is more reasonable to use other approaches that suit the dense case well.

Supplementary Material: All technical details, figures, and tables are relegated to the online Supplementary Material.

Acknowledgments The authors thank the editor, associate editor, and an anonymous referee for their insightful comments. Zhang's research was supported by the U.S. NSF grant DMS–1712418 and the Wisconsin Alumni Research Foundation.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser.* B, 57, 289–300.

[2] Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165–1188.

[3] Benjamini, Y., Krieger, A., Yekutieli, D. (2006). Adaptive linear stepup procedures that control the false discovery rate. *Biometrika*, 93, 491–507.

[4] Bickel, P. and Doksum, K. (2007). *Mathematical Statistics, Basic Ideas and Selected Topics*, vol. 1. Second edition. Pearson Prentice Hall.

[5] Bourgon, R., Gentleman, R. and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107, 9546–9551.

[6] Cao, H. (2007). Moderate deviations for two sample t-statistics. *ESAIM Probab. Stat.*, 11, 264–271.

[7] Cao, H. and Kosorok, M. (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli*, 17, 347–394.

[8] Chung, K.L. (2001). *A Course in Probability Theory*, Third Edition. Academic Press.

[9] Cressie, N. A. C., and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal*, 28(2), 131–148.

[10] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer.

[11] Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.*, 99, 96–104.

[12] Efron, B. (2010). *Large-Scale Inference. Empirical Bayes methods for estimation, testing, and prediction*. Institute of Mathematical Statistics (IMS) Monographs, 1. Cambridge University Press, Cambridge.

[13] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102, 1282–1288.

[14] Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.*, 107, 1019–1035.

[15] Jing, B., Kong, X., and Zhou, W. (2014). FDR control in multiple testing under non-normality. *Statist. Sinica*, 24, 1879–1899.

[16] Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, 73(363), 536–544.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

[17] Kim, J.H., Dhanasekaran, S.M., Mehra, R., Tomlins, S.A., Gu, W., Yu, J., Kumar-Sinha, C., Cao, X., Dash, A., Wang, L. et al. (2007). Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Research*, 67, 8229–8239.

[18] Kim, D. and Zhang, C.M. (2014). Adaptive linear step-up multiple testing procedure with the bias-reduced estimator. *Statistics and Probability Letters*, 87, 31–39.

[19] Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, N.Y..

[20] Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc Natl Acad Sci USA.*, 105, 18718-18723.

[21] Lehmann, E. L., and Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.*, 33, 1138–1154.

[22] Liang, K., Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B*, 74, 163–182.

[23] Liu, W.D. and Shao, Q.M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Annals of Statistics*, 42, 2003–2025.

[24] Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. Ser.* B, 66, 187–205.

[25] Wang, Q. (2005). Limit theorems for self-normalized large deviations. *Electronic Journal of Probability*, 10, 1260–1285.

[26] Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0467

[27] Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34, 28–35.

[28] Zhang, C.M., Fan, J. and Yu, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *Annals of Statistics*, 39, 613–642.

[29] Zhao, Z., Wang, W. and Wei, Z. (2013). An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Annals of Applied Statistics*, 7, 2229–2248.