<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2018-0467</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Chunming Zhang  
Shengji Jia and  
Yongfeng Wu |
| **Corresponding Author** | Chunming Zhang |
| **E-mail** | cmzhang@stat.wisc.edu |

Notice: Accepted version subject to English editing.
On simultaneous calibration of two-sample $t$-tests for high-dimension low-sample-size data

Chunming Zhang, Shengji Jia, Yongfeng Wu

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

September 16, 2019

Abstract

The exact distribution of a two-sample $t$-statistic in a single test for equal population means is typically unavailable with either non-Gaussian samples or unequal population variances or unequal sample sizes $n_1$ and $n_2$, and thus a calibration method through a reference distribution offers a practically feasible substitute. This paper addresses new issues in simultaneously calibrating a diverging number $m$ of two-sample $t$-statistics for simultaneous inference of significance in high-dimension low-sample-size data. For the Gaussian calibration method, we demonstrate that (a) the simultaneous “general” two-sample $t$-statistics achieve the overall significance level, if $\log(m)$ increases at a strictly slower rate than $(n_1 + n_2)^{1/3}$ as $n_1 + n_2$ diverges; (b) however, directly applying the same calibration method to simultaneous “pooled” two-sample $t$-statistics may substantially lose the overall level accuracy. The proposed “adaptively pooled” two-sample $t$-statistics overcome such incoherence, whereas operate as simply as but perform as well as the “general” two-sample $t$-statistics.
Moreover, we propose a “two-stage” $t$-test procedure to effectively alleviate the skewness effects commonly encountered from various two-sample $t$-statistics in practice, thus enhancing the calibration accuracy. Implications of these results are illustrated using both 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-dimension low-sample-size data, with wide applications in biology, genetics, astronomy, economics and neuroscience research among others. This problem is characterized by carrying out a large number of hypothesis tests simultaneously, whereas each test involves a data vector whose length is comparatively smaller. For example, in the 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. Those procedures implicitly assume that some marginal quantities, such as significance levels (or type-I error rates) and $p$-values, can be calculated exactly for each of the simultaneous tests. In practice, such assumption may not be realistic when the exact distributions of test statistics in finite-sample cases are not directly available. Those cases motivate the need of estimating the distributions from which marginal quantities are computed, but it is unclear how good the approximation must be for simultaneous inference to be feasible.

This paper concerns the performance of conducting a diverging number $m$ of two-sample $t$-tests simultaneously for the equality between mean effects of two groups, where
m frequently far exceeds the sample sizes $n_1$ and $n_2$ in two groups, while the combined sample size $n = n_1 + n_2$ is still moderately large. Three issues arise naturally from analyzing such matrix type of data. (i) It is well-known that the exact distribution of an individual two-sample $t$-statistic for comparing population means is typically unavailable with either non-Gaussian samples or unequal population variances or unequal sample sizes. Indeed, this issue remains to be one of the unsolved problems in the statistical literature, the so-called Behrens-Fisher problem [26, 27]. In statistical practice, a calibration method through a reference distribution, for example the standard Gaussian distribution $N(0, 1)$, serves as a feasible substitute, provided that the approximation accuracy suffices for finite-sample sizes. (ii) The two-sample problem is more important in certain sense, but more complex and challenging than the one-sample problem. Moreover, unlike the one-sample $t$-statistic, the way of choosing a two-sample $t$-statistic is not unique in statistical tools. The two most common choices are the “general” two-sample $t$-statistic and “pooled” two-sample $t$-statistic. Nonetheless, the discussion on whether the calibration method for the simultaneous “general” choice and simultaneous “pooled” choice is immediately equally applicable is not available in the literature. (iii) In practice, asymmetric populations appear frequently, but reduce the accuracy of a single two-sample $t$-statistic. No discussion is available in the literature for simultaneous inference based on a diverging number of two-sample $t$-statistics.

Due to the popularity of two-sample $t$-tests in research practices, 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 paper addresses three new issues for two-sample $t$-statistics allowing both independent as well as dependent data.

**Issue 1**: We demonstrate that for the Gaussian calibration method, the overall significance 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. Besides, we demonstrate that \((m, n_1, n_2)\) supports the controlling of false discovery rates (FDR) of some multiple testing procedures based on calibrated \(p\)-values.

**Issue 2**: In contrast, the “pooled” two-sample \(t\)-statistics may behave substantially differently from the “general” two-sample \(t\)-statistics, particularly in the case where a “composite-variance-quantity” (to be defined in \((2.7)\) of Section 2) exceeds one. The proposed “adaptively pooled” two-sample \(t\)-statistics in Section 3.2 operate as simply as 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 enhancing the calibration accuracy.

In the case of simultaneous one-sample \(t\)-statistics under independence and positive regression dependence on subsets [2], calibration through the Gaussian or Student \(t\) distribution and bootstrap method was studied in [13], assuming that the number \(m_0\) of true null hypotheses is identical to \(m\), i.e. \(m_0 = m\), which is restrictive in applications. In the current paper, we address 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 non-random quantity. We also apply the factor model to deal with several practically motivated dependence models including the jointly Gaussian distributed test statistics, so that FDR can be controlled asymptotically.

The rest of the article is organized as follows. Section 2 formulates the overall significance level of simultaneous two-sample \(t\)-statistics for comparing means of two populations. Section 3 addresses the above issues 1–3 in details. Section 4 discusses the effect of dependence among observations on the calibration method. Section 5 and Section 6 illustrate simulation studies and real data examples, respectively. Section 7 concludes the paper. All
technical details, figures and tables are relegated to Appendices (in the online supplement).

2 Model structure and significance testing

In many applications, data coming from two groups, such as a normal control group and a cancer patient group, will be compared. 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:

\[
\begin{align*}
X_{i,j} &= \mu_{X,i} + \varepsilon_{i,j}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_1, \\
Y_{i,j} &= \mu_{Y,i} + e_{i,j}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_2, 
\end{align*}
\]

(2.1)

where the index \(i\) refers to the \(i\)th test (for example, gene or brain voxel), \(j\) indicates the \(j\)th 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 \(i\)th test, and \(\varepsilon_{i,j}\) and \(e_{i,j}\) are random errors. Some basic assumptions are collected in conditions A1–A3 for statistical analysis.

The hypotheses we wish to test specify

\[
H_{0,i} : \mu_{X,i} = \mu_{Y,i} \quad \text{against} \quad H_{1,i} : \mu_{X,i} \neq \mu_{Y,i},
\]

(2.2)

for \(1 \leq i \leq m\) simultaneously. 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, have been widely used in statistical applications. One version is formed by the “general” two-sample t-statistic ([26], equation (2)),

\[
T_{i;n_1,n_2}^{\text{general}} = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{s_{X;i}^2/n_1 + s_{Y;i}^2/n_2}},
\]

(2.3)

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

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

(a2) In other cases, the exact distribution of $T_{i;n_1,n_2}^{\text{general}}$ under $H_{0,i}$ is typically unavailable, but the central limit theory (CLT) and Slutsky’s theorem \[10\] give

$$T_{i;n_1,n_2}^{\text{general}} \xrightarrow{D} N(0,1), \quad \text{under } H_{0,i}, \quad (2.4)$$

as $n_1 \to \infty$ and $n_2 \to \infty$, where $\xrightarrow{D}$ denotes converges 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; \[5\]),

$$T_{i;n_1,n_2}^{\text{pool}} = \frac{\bar{X}_i - \bar{Y}_i}{s_{\text{pool};X,Y;i} \sqrt{1/n_1 + 1/n_2}}, \quad (2.5)$$

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

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

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

$$T_{i;n_1,n_2}^{\text{pool}} \xrightarrow{D} N(0, \sigma_{\rho,\theta_{(\varepsilon,e)},i}^2), \quad \text{under } H_{0,i}, \quad (2.6)$$

where

$$\sigma_{\rho,\theta_{(\varepsilon,e)},i}^2 = \frac{(1-\rho) + \rho \theta_{(\varepsilon,e)}^{(\varepsilon,e);i}}{\rho + (1-\rho) \theta_{(\varepsilon,e)}^{(\varepsilon,e);i}}, \quad \text{with } \theta_{(\varepsilon,e);i} = \sigma_{e,i}^2/\sigma_{\varepsilon,i}^2. \quad (2.7)$$
The derivation of (2.6) is relegated to the Appendix A. We call \( \sigma^2_{\rho;\theta(\varepsilon,e);i} \) the “composite-variance-quantity” (CVQ), which aggregates the ratio of sample sizes with the ratio of population variances. Clearly, \( \sigma^2_{\rho;\theta(\varepsilon,e);i} = 1 \) holds only in either Case I or Case II below:

**Case I**: \( \rho = 1/2, \) i.e. equal sample sizes with \( n_1 = n_2; \)  

\[
\text{Case I : } \quad \rho = 1/2, \text{ i.e. equal sample sizes with } n_1 = n_2; \quad (2.8)
\]

**Case II**: \( \theta(\varepsilon,e);i = 1, \) i.e. equal population variances with \( \sigma^2_{\varepsilon;i} = \sigma^2_{e;i}. \)  

\[
\text{Case II : } \quad \theta(\varepsilon,e);i = 1, \text{ i.e. equal population variances with } \sigma^2_{\varepsilon;i} = \sigma^2_{e;i}. \quad (2.9)
\]

Also note that \( \sigma^2_{\rho;\theta(\varepsilon,e);i} > 1 \) holds only if either “\( n_1 < n_2 \) and \( \sigma^2_{\varepsilon;i} > \sigma^2_{e;i} \)” or “\( n_1 > n_2 \) and \( \sigma^2_{\varepsilon;i} < \sigma^2_{e;i} \).” In general, the limiting distribution in (2.6) could not be used directly, since population variances \( \sigma^2_{\varepsilon;i} \) and \( \sigma^2_{e;i} \) in \( \theta(\varepsilon,e);i \) are 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 overall significance level is entailed for controlling some aspects of the overall error rates. We first use the “general” two-sample \( t \)-statistics \( \{T_{i;n_1,n_2}^{\text{general}}\}_{i=1}^{m} \) to introduce some necessary notations. Discussions on extensions to alternative choices \( \{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^{m} \) will be postponed to Section 3.2. For a critical value \( t \), the significance level of the \( i \)th test is

\[
\alpha_{i;n_1,n_2}(t) = P_{H_0,i}(|T_{i;n_1,n_2}^{\text{general}}| > t), \quad (2.10)
\]

where \( P_{H_0,i} \) denotes the probability calculated when \( H_{0,i} \) is true. When testing \( m \) null hypotheses simultaneously, indices of true null hypotheses are collected in the set \( \mathcal{I}_0 = \{ i : H_{0,i} \text{ is true} \} \), with the cardinality \( m_0 = |\mathcal{I}_0| \). The overall significance level is captured by the family-wise-error-rate (abbreviated as FWER or FWER), \( \text{FWER}(t) = \mathbb{P}(V_m(t) \geq 1) \), where \( V_m(t) = \sum_{i=1}^{m} 1(H_{0,i} \text{ is true}, |T_{i;n_1,n_2}^{\text{general}}| > t) \) denotes the number of false rejections, with an indicator operator \( 1(\cdot) \). More generally, for integers \( k \geq 1, \) \( \text{FWER}_k(t) = \mathbb{P}(V_m(t) \geq k) \) denotes the \( k \)-fold family-wise-error-rate (abbreviated as FWER, see [21]).
Recall from Section 2.1 that exact values of $\alpha_{i:n_1,n_2}(t)$ based on the exact null distribution of $T^\text{general}_{i:n_1,n_2}$ are unavailable in many practical settings. However, when $n_1 \to \infty$ and $n_2 \to \infty$, the null distribution of $T^\text{general}_{i:n_1,n_2}$ can be approximated by $\mathcal{N}(0,1)$, as seen in (2.4).

This result motivates the approximation by $\mathcal{N}(0,1)$ random variables $\{T^a_i\}_{i=1}^m$. It is thus natural to utilize quantities,

$$
\alpha^a_i(t) = P(|T^a_i| > t), \quad V^a_m(t) = \sum_{i \in I_0} I(|T^a_i| > t),
$$

FWER^a(t) = P(V^a_m(t) \geq 1), \quad FWER^a_k(t) = P(V^a_m(t) \geq k),
$$

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

In this paper, we will study the relation between the number $m$ of tests and sample sizes $n_1$ and $n_2$ within each test, such that appropriate choices of critical values $t^a_{\alpha;m}$ and $t^a_{\alpha;m;k}$ obtained from the calibrated distributions (through $\{T^a_i\}_{i=1}^m$), when applied to two-sample $t$-statistics $\{T^\text{general}_{i:n_1,n_2}\}_{i=1}^m$ and $\{T^\text{bool}_{i:n_1,n_2}\}_{i=1}^m$, will guarantee that

$$
\text{FWER}_1(t^a_{\alpha;m}) \leq \alpha + o(1), \quad (2.11)
$$

$$
\text{FWER}_k(t^a_{\alpha;m;k}) \leq \alpha + o(1), \quad (2.12)
$$

as $m \to \infty$, $n_1 \to \infty$ and $n_2 \to \infty$, where $\alpha$ is the control level. Similarly, it is ideally desirable to control FDR based on certain threshold $\tau_{\alpha;m;n}$ for true $p$-values $\{P_i\}$, i.e. $\text{FDR}(\tau_{\alpha;m;n}) \leq \alpha + o(1)$, where $\text{FDR}(\tau) = \mathbb{E}\left[\frac{\sum_{i \in I_0} I(P_i \leq \tau)}{\sum_{i=1}^m I(P_i \leq \tau)}\right]$, with $a \lor b = \max\{a, b\}$.

When the exact $\{P_i\}$ are unavailable, it is more realistic to control the corresponding FDR based on some threshold $\tau^a_{\alpha;m;n}$ for the calibrated $p$-values $\{P^a_i\}_{i=1}^m$, such that

$$
\text{FDR}(\tau^a_{\alpha;m;n}) \leq \alpha + o(1). \quad (2.13)
$$
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 \(\{T_{\text{general}}^i\}_{i=1}^m\), for which additional assumptions A4–A7 are made. We also assume that rates of growth of \(m\), \(n_1\) and \(n_2\) are connected via

\[
\log(m) = o(n^{1/3}),
\]

with the combined sample size \(n = n_1 + n_2\).

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

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

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

\[
t_{\alpha;m}^a = \Phi^{-1}\left(\left\{1 + (1 - \alpha)^{1/m}\right\}/2\right),
\]

where \(\Phi\) denotes the cumulative distribution function (C.D.F.) of a \(N(0,1)\) variable, then

\[
\begin{align*}
\text{FWER}_1(t_{\alpha;m}^a) &= \text{FWER}_1^a(t_{\alpha;m}^a) + o(1), \\
\text{FWER}_1^a(t_{\alpha;m}^a) &= 1 - (1 - \alpha)^{m_0/m} \leq \alpha.
\end{align*}
\]

Likewise, (3.6) of Proposition 2 below implies that FWER\(_k(t_{\alpha;m;k}^a)\) \(\leq \alpha + o(1)\), which is desirable in (2.12). A common feature of Propositions 1–2 is that as the proportion \(\pi_0\) of true nulls approaches one, FWER\(_1(t_{\alpha;m}^a)\) and FWER\(_k(t_{\alpha;m;k}^a)\) approach the control level \(\alpha\), and hence inequalities in (2.11)–(2.12) become equalities.
Proposition 2 (control FWER \( k \) under independence among tests) Assume model \([2.1]\) and conditions A1–A7. For \( k \geq 2, \alpha \in (0,1), m_0/m \to \pi_0 \in (0,1], m \to \infty, n \to \infty, \) if the general two-sample t-statistics \( \{T_{i,n_1,n_2}^{\text{general}}\}_{i=1}^m \) are used, \((m,n)\) satisfies \([3.1]\), and

\[
t_{\alpha;m;k}^a = \Phi^{-1}(1 - (\beta_{k;\alpha}/2)/m), \tag{3.4}
\]

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

\[
\begin{align*}
\text{FWER}_k(t_{\alpha;m;k}^a) &= \text{FWER}_k^a(t_{\alpha;m;k}^a) + o(1), \\
\text{FWER}_k^a(t_{\alpha;m;k}^a) &= G_k(\pi_0 \beta_{k;\alpha}) + o(1) \leq \alpha + o(1). \tag{3.6}
\end{align*}
\]

3.1.2 Controlling FDR in \([2.13]\) of multiple testing procedures

Similar to the marginal significance levels, true marginal \( p \)-values \( \{P_i\} \) are unknown in advance or not directly available, when the exact distributions of 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 \( \{P_i^a\} \), the resulting multiple testing procedure, such as the Bonferroni correction, is still valid, in the sense that the false discovery rate (FDR) under conditions of Proposition 1 will be asymptotically bounded by the level \( \alpha \), if the approximation errors of \( p \)-values are \( o(1/m) \).

Analogously, consider the Benjamini-Hochberg (BH) multiple testing procedure \([1]\) which rejects null hypotheses \( H_{0,i} \) with \( P_i \leq P_{(k)} \), where \( \hat{k} = \max\{j : P_{(j)} \leq \alpha j/m\} \), and \( P_{(1)} \leq \cdots \leq P_{(m)} \) denote the ordered \( p \)-values \( \{P_i\} \). Then \( \text{FDR}_{\text{BH}} = E(\frac{V_{\text{BH}}}{\bar{R}_{\text{BH}}} \vee 1) \) gives the FDR of the BH procedure, where \( V_{\text{BH}} = \sum_{i \in I_0} I(P_i \leq P_{(k)}) \) and \( \bar{R}_{\text{BH}} = \hat{k} \). For the calibration method, applying the approximate \( p \)-values \( \{P_i^a\} \) instead of \( \{P_i\} \) to the BH procedure yields the number \( V_{\text{BH}}^a \) of false rejections and the number \( R_{\text{BH}}^a \) of total rejections, and the corresponding FDR defined by \( \text{FDR}_{\text{BH}}^a = E(\frac{V_{\text{BH}}^a}{R_{\text{BH}}^a} \vee 1) \). More generally, for
Proposition 3 (control FDR(τα;m;n) of the BH procedure under independence among tests)

Assume model (2.1) and conditions A1–A5, A5′, A6, A7′, A8–A10. Define by $F_P(·; n)$ and $f_P(·; n)$ the C.D.F. and p.d.f. of approximate p-values $\{P_i\}^m_{i=1}$. For $\alpha \in (0, 1)$, let

$$
\varsigma_{α;n} = \sup\{t : H(t; n) \leq \alpha\}, \quad \varsigma_{α;n}^a = \sup\{t : H^a(t; n) \leq \alpha\},
$$

where $H(t; n) = t/F_P(t; n)$ and $H^a(t; n) = t/F^a_P(t; n)$. Suppose $H'(t; n)$ is bounded below for $t$ in an open interval with endpoints $\varsigma_{α;n}$ and $\varsigma_{α;n}^a$, and $f_P(\varsigma_{α;n}^a; n) < α^{-1} < f_P(0; n)$. If the general two-sample t-statistics $\{T_{i\ell}^m\}^m_{i=1}$ are used, and

$$
\Phi^{-1}(1 - \varsigma_{α;n}^a) \in (0, o(n^{1/6})],
$$

(3.7)

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

$$
\text{FDR}(\tau_{α;m;n}^a) \leq \alpha + o(1).
$$

(3.8)

Remark 1 Similar to Lemma A.1 of [13], we obtain $\tau_{α;m;n}^a = \varsigma_{α;n}^a + O_P(m^{-1/2})$, where $\tau_{α;m;n}^a$ gives the threshold for approximate p-values. So condition (3.7) becomes

$$
\Phi^{-1}(1 - \tau_{α;m;n}^a + O_P(m^{-1/2})) \in (0, o(n^{1/6})],
$$

(3.9)

which implicitly describes the relationship between $m$ and $n$. For example, if $\tau_{α;m;n}^a$ is of order $m^{-b}$ with probability tending to 1, 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 \( \liminf_{(n,m) \to \infty} \text{FDR}(\tau_{\alpha;m;n}^a) \geq \beta \) with a constant \( \beta > \alpha \). In particular, if \( \log(m)/n^{1/3} \to \infty \), we obtain \( \text{FDR}(\tau_{\alpha;m;n}^a) \to 1 \), implying that FDR is out of control 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 \( \log(m) = o(n^{1/2}) \), if we assume more conditions, such as the symmetric errors and a stronger large deviation result for two-sample t-tests \( T_{\text{general};n_1,n_2}^{\text{general}}: \mathbb{P}_{H_{0,i}}(T_{\text{general};n_1,n_2}^{\text{general}} \geq x)/(1 - \Phi(x)) = \exp(-3^{-1}\kappa_3 x^3 n^{-1/2})\{1 + \theta(1 + x)^2/n^{1/2}\} \), where \( \kappa_3 = \frac{\mathbb{E}\{(X_{i,1} - \mu_{X,i})^3\} + \mathbb{E}\{(Y_{i,1} - \mu_{Y,i})^3\}}{\mathbb{E}\{\sigma_{e,i}^2 + \sigma_{e,i}^2\}^{3/2}} \), and \( \theta = \theta(x,n) \) satisfies \( |\theta(x,n)| \leq C \) uniformly in \( x \in (0,o(n^{1/4})) \). The justification of this large deviation result is beyond the scope of this article.

(iii) The condition A5', “two-sample t-statistics corresponding to true non-nulls are identically distributed”, merely simplifies the technical proof for Proposition 3. For simulation studies in Section 5 where the differences \( (\mu_{X,i} - \mu_{Y,i}) \) under true non-nulls vary with \( i \), numerical evidences in Figure 13 indicates that Proposition 3 continues to hold in cases where condition A5' is relaxed.

Remark 3 In Proposition 13, the Gaussian distribution is used to approximate the distribution of test statistics \( T_{i,n_1,n_2}^{\text{general}} \). These results can be easily generalized to the t-distribution approximation, if we replace \( \Phi(\cdot) \) by the C.D.F. of the \( t_{n_1+n_2-2} \) distribution.
3.2 Proposed “adaptively pooled” two-sample $t$-tests for (2.2)

We now discuss error controls using the “pooled” two-sample $t$-statistics $\{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^m$. Recall from (A.11) and (A.26) in the Appendix A that conclusions of Propositions 1–2 rely on the tail distribution of $T_{i;n_1,n_2}^{\text{general}}$ under the null $H_{0,i}$ approximated by that of the $N(0,1)$ distribution, fulfilling

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

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

$$|P_{H_{0,i}}(T_{i;n_1,n_2}^{\text{pool}} \geq x)/(1 - \Phi(x)) - 1| \to 0 \quad (3.10)$$

holds uniformly up to the point $x$ of order $o(n^{1/6})$, then (2.11) and (2.12) are also applicable to $\{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^m$. Indeed, condition (3.10) holds when CVQ equals one, i.e., $\sigma_{p\theta(e,e)} = 1$, in either Case I with $n_1 = n_2$ as discussed in (2.8) or Case II with $\sigma_{\varepsilon;i}^2 = \sigma_{e;i}^2$ as discussed in (2.9). Numerical evidences are given in Figure 3 with $\sigma_{\varepsilon;i}^2 = \sigma_{e;i}^2$, 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).

For more realistic situations where the CVQ differs from one, how would (3.10) be affected? If the original form (2.5) of $T_{i;n_1,n_2}^{\text{pool}}$ is used, then the result in (2.6) indicates

$$P_{H_{0,i}}(T_{i;n_1,n_2}^{\text{pool}} \geq x)/(1 - \Phi(x)) = \{1 - \Phi(x/\sigma_{p\theta(e,e)}^{1/2})\}/\{1 - \Phi(x)\}\{1 + o(1)\}. \quad (3.11)$$

To analyze the ratio on the right hand side of (3.11), 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 1 when $\sigma < 1$. Such difference ultimately affects (3.10) in the following ways.

(i) If $\sigma_{p\theta(e,e)} > 1$, then the maximum value of $|P_{H_{0,i}}(T_{i;n_1,n_2}^{\text{pool}} \geq x)/(1 - \Phi(x)) - 1|$ will always be much larger than 0.
(ii) If $\sigma_{\rho \theta(e,e):i} < 1$, then the maximum value of $|P_{H_0,i}(T_{i/m_1,n_2}^{\text{pool}} \geq x)/\{1 - \Phi(x)\} - 1|$ potentially will approach 0, particularly when $\sigma_{\rho \theta(e,e):i}$ approaches 1.

Hence, condition (3.10) may fail if $\sigma_{\rho \theta(e,e):i} > 1$, and the overall level accuracy may be lost from directly applying the calibration method to simultaneous “pooled” two-sample $t$-statistics $T_{i/m_1,n_2}^{\text{pool}}$. See numerical illustrations in Figure 5 associated with $\sigma_{\rho \theta(e,e):i} > 1$.

To circumvent the incoherence of $T_{i/m_1,n_2}^{\text{pool}}$ with $T_{i/m_1,n_2}^{\text{general}}$, particularly in the case of CVQ > 1, we propose to develop an “adaptively pooled” version, which has approximately a N(0, 1) distribution under the null. Following (2.6), a natural choice is given by

$$T_{i/m_1,n_2}^{\text{pool;A}} = \frac{T_{i/m_1,n_2}^{\text{pool}}}{\hat{\rho} \hat{\theta}_{(e,e):i}},$$

(3.12)

where $\hat{\theta}_{(e,e):i} = s_{X;i}/s_{Y;i}$ serves as an estimate of $\theta_{(e,e):i} = \sigma_{e;i}/\sigma_{e;i}$. Simulation results in Section 5 support that the performance of the calibration method applied to the “adaptively pooled” choice $\{T_{i/m_1,n_2}^{\text{pool;A}}\}_m^i$ is comparable to that applied to the “general” choice $\{T_{i/m_1,n_2}^{\text{general}}\}_m^i$.

### 3.3 Proposed “two-stage” $t$-test procedure for (2.2)

In practice, $T_{i/m_1,n_2}^{\text{general}}$ and $T_{i/m_1,n_2}^{\text{pool;A}}$ could be skewly distributed under $H_{0,i}$, yielding slower convergence rate to N(0, 1) and lower calibration accuracy by N(0, 1). See also Remark 2(ii). For $T_{i/m_1,n_2}^{\text{general}}$, its theoretical form of skewness “adjusted” two-sample $t$-statistic,

$$T_{i/m_1,n_2}^{\text{adjust;T}} = \frac{\left(\bar{X}_i - \bar{Y}_i\right)}{\frac{\mu_3X,i/n_i^2 - \mu_3Y,i/n_i^2}{6(s_{X;i}/n_1 + s_{Y;i}/n_2)} + \frac{\mu_3X,i/n_i^2 - \mu_3Y,i/n_i^2}{3(s_{X;i}/n_1 + s_{Y;i}/n_2)^2} \left(\bar{X}_i - \bar{Y}_i\right)^2} \cdot \sqrt{s_{X;i}/n_1 + s_{Y;i}/n_2},$$

(3.13)

can be derived from [10] used for the “adjusted” one-sample $t$-statistic, where

$$\mu_3X,i = E\{(X_{i,1} - \mu_{X,i})^3\}, \quad \mu_3Y,i = E\{(Y_{i,1} - \mu_{Y,i})^3\}. \quad \text{(3.14)}$$

A form similar to (3.13) is in equation (2.16) of [9]. As expected, $T_{i/m_1,n_2}^{\text{adjust;T}}$ could alleviate the skewness effects from $T_{i/m_1,n_2}^{\text{general}}$, and thus will be more symmetric under $H_{0,i}$. Clearly, if
\[
\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0, \text{ then } T^{adj;\text{E}}_{i:n_1,n_2} \text{ reduces to } T^{\text{general}}_{i:n_1,n_2}. \text{ Hence, the quantity}
\]
\[
\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = E[(X_i - \bar{X}_i) - E(X_i - \bar{X}_i)]^3,
\]  
(3.15)

serves as a valid measure of skewness of \(T^{\text{general}}_{i:n_1,n_2}\), assuming conditions A1, A2 and A3. In practice, \(T^{adj;\text{E}}_{i:n_1,n_2}\) is infeasible for skewness adjustment, because the quantity \(\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2\) is unknown, but can be estimated via sample third moments, leading to the empirical form of skewness “adjusted” two-sample \(t\)-statistic,
\[
T^{adj;\text{E}}_{i:n_1,n_2} = \frac{(\bar{X}_i - \bar{Y}_i) + \frac{\bar{\mu}_{3,X;i}/n_1^2 - \bar{\mu}_{3,Y;i}/n_2^2}{6(s_{X;i}/n_1 + s_{Y;i}/n_2)} + \frac{\bar{\mu}_{3,X;i}/n_1^2 - \bar{\mu}_{3,Y;i}/n_2^2}{3(s_{X;i}/n_1 + s_{Y;i}/n_2)}^2 (\bar{X}_i - \bar{Y}_i)^2}{\sqrt{s_{X;i}^2/n_1 + s_{Y;i}^2/n_2}},
\]  
(3.16)

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

Regarding the choice between \(T^{\text{general}}_{i:n_1,n_2}\) and \(T^{adj;\text{E}}_{i:n_1,n_2}\), 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 \(T^{adj;\text{E}}_{i:n_1,n_2}\), and also outperform \(T^{adj;\text{E}}_{i:n_1,n_2}\) (due 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 0, then \(T^{adj;\text{E}}_{i:n_1,n_2}\) will be effective in correcting the skewness, whereas \(T^{\text{general}}_{i:n_1,n_2}\) may not be.

Hence, before selecting between \(T^{\text{general}}_{i:n_1,n_2}\) and \(T^{adj;\text{E}}_{i:n_1,n_2}\), it will be more natural to first assess the adequacy of
\[
H^{(1)}_{0,i} : \mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0.
\]  
(3.17)

Note that (3.14) and (3.17) motivate us to consider the \(t\)-statistic:
\[
\frac{\bar{\mu}_{3,X;i}/n_1^2 - \bar{\mu}_{3,Y;i}/n_2^2}{\sqrt{\hat{s}^2_{3,X;i}/n_1^2 + \hat{s}^2_{3,Y;i}/n_2^5}}
\]  
(3.18)

where \(\hat{s}^2_{3,X;i}\) and \(\hat{s}^2_{3,Y;i}\) denote sample variances of \(\{(X_{i,j} - \bar{X}_i)^3\}_{j=1}^{n_1}\) and \(\{(Y_{i,j} - \bar{Y}_i)^3\}_{j=1}^{n_2}\) respectively. Under the null hypothesis (3.17), (3.18) \(\overset{D}{\rightarrow}\) \(N(0,1)\) by the CLT and Slutsky’s theorem, assuming the finite 6th moments of \(X_{i,1}\) and \(Y_{i,1}\).

To improve the efficiency in testing (2.2), we propose the “two-stage” \(t\)-test procedure.
1st-stage: For each $i = 1, \ldots, m$, apply the 1st-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 2nd-stage two-sample $t$-statistic $T_{i:n_1,n_2}^{2,\text{stage}}$ by

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

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

As illustrated in simulation studies in Section 5, $T_{\text{adjust};T}^{\text{stage}}$ always performs the best, but is practically infeasible. The proposed $T_{i:n_1,n_2}^{2,\text{stage}}$ is as good as the better of $T_{\text{general}}^{\text{stage};i}$ and $T_{\text{adjust};E}^{\text{stage};i}$.

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 $\sigma_{\hat{\theta}(\varepsilon,e)} \sqrt{1/n_1 + 1/n_2}$.

4 Error controls allowing dependent data

In practice, dependence in datasets may arise either among different tests, or between the $X$-group and $Y$-group, or within the same $X$-group or within the same $Y$-group. Section 4.1 considers types of dependence among tests, Sections 4.2–4.3 explore modes (4.10) and (4.12) incorporating dependence structure between two groups and within a same group, and Appendix B discusses extensions of (4.10) and (4.12).

4.1 Dependence among tests

Recall that Propositions 1–2 rely on condition A7 which assumes independence among test statistics corresponding to true nulls. Section 4.1.1 evaluates the impact of general dependency among test statistics on the control of overall significance level, where Propositions...
4–5 will remove condition A7. Section 4.1.2 considers test statistics which are asymptotically jointly Gaussian.

4.1.1 General dependence among tests

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

$$\text{FWER}_1(t^a_{\alpha;m}) \leq \pi_0 \beta_{1;\alpha} + o(1),$$

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 $\alpha$ with any $\pi_0 \leq \alpha/\beta_{1;\alpha}$, 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 $\alpha$. For example, a level $\alpha = 0.05$ allows any choice of $\pi_0$ in the range $(0, 0.9748]$, wide enough for realistic applications. Interestingly, even in the special case of $\pi_0 = 1$ (which is rare in practice), the fact of $\pi_0 \beta_{1;\alpha} = \beta_{1;\alpha}$ and $\alpha \leq \beta_{1;\alpha}$ (with negligible difference between $\alpha$ and $\beta_{1;\alpha}$ particularly when $\alpha$ is small as illustrated in the right panel of Figure 2) indicates that the critical value $t^a_{\alpha;m}$ in (3.2) offers an asymptotically slightly conservative level $\beta_{1;\alpha}$ for the resulting FWER($t^a_{\alpha;m}$).

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

**Proposition 5 (control FWER$_k(t^a_{\alpha;m;k})$ under general dependence among tests)** Assume model (2.1) and conditions A1–A6. For $k \geq 2$, $\alpha \in (0, 1)$, $m_0/m \to \pi_0 \in (0, 1]$, $m \to \infty$, $n \to \infty$, if the general two-sample $t$-statistics $\{T_{i:n_1,n_2}^{\text{general}}\}_{i=1}^m$ are used, with $t^a_{\alpha;m;k}$ given in
and \((m, n)\) satisfying (3.1), then
\[
\text{FWER}_k(t_{a,m;k}^\alpha) \leq \pi_0 \beta_{k;\alpha}/k + o(1),
\]
(4.2)
where \(\beta_{k;\alpha}\) solves (3.5).

As a comparison to Proposition 2, the upper bound \(\pi_0 \beta_{k;\alpha}/k\) in (4.2), with \(k \geq 2\), is controlled by the nominal level \(\alpha\), only when the proportion \(\pi_0\) does not exceed \(\alpha/(\beta_{k;\alpha}/k)\), which equals 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 \(\alpha\). This reflects the cost of generalizing Proposition 2 from mutually independent tests to cases allowing general dependency.

4.1.2 Jointly Gaussian distributed test statistics

Consider a specific factor model for observations \(\{X_{i,j}\}\) and \(\{Y_{i,j}\}\):
\[
X_{i,j} = \mu_{X,i} + \beta_{X;i}^T u_j + \varepsilon_{i,j}, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_1,
\]
\[
Y_{i,j} = \mu_{Y,i} + \beta_{Y;i}^T v_j + e_{i,j}, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_2.
\]
(4.3)
where \(u_j\) are unobserved \(d_u\)-dimensional random vectors, with \(\{u_1, \ldots, u_{n_1}\} \overset{i.i.d.}{\sim} N(0, \Sigma_u)\); \(v_j\) are unobserved \(d_v\)-dimensional random vectors, with \(\{v_1, \ldots, v_{n_2}\} \overset{i.i.d.}{\sim} N(0, \Sigma_v)\); \(u_1, \ldots, u_{n_1}\) and \(v_1, \ldots, v_{n_2}\) are independent. For example, gene expressions \(\{X_{i,j} : 1 \leq i \leq m\}\) of the \(j\)th subject may be influenced by some common factors \(u_j\), say, the age or other variables of the \(j\)th subject. Also, assume \(\{\varepsilon_{i,j}\}\) and \(\{e_{i,j}\}\) are identical to those in model (4.10) below; \(\{\varepsilon_{i,j}\}\), \(\{e_{i,j}\}\), \(\{u_j\}\) and \(\{v_j\}\) are independent.

For model (4.3), the dependence among the two-sample \(t\)-statistics,
\[
T_{i;\!n_1,n_2}^{\text{general}} = \frac{(\mu_{X,i} - \mu_{Y,i}) + (\bar{\varepsilon}_i - \bar{e}_i) + (\beta_{X;i}^T \bar{u} - \beta_{Y;i}^T \bar{v})}{\sqrt{s_{X;i}/n_1 + s_{Y;i}/n_2}}, \quad i = 1, \ldots, m,
\]
(4.4)
are caused by factors \(\bar{u} = \sum_{j=1}^{n_1} u_j/n_1\) and \(\bar{v} = \sum_{j=1}^{n_2} v_j/n_2\) common to all tests. It follows that the two-sample \(t\)-statistics can be rewritten as
\[
(T_{1;\!n_1,n_2}^{\text{general}}, \ldots, T_{m;\!n_1,n_2}^{\text{general}})^T = D \bullet U,
\]
(4.5)
where \( D = (X_1 - Y_1, \ldots, X_m - Y_m)^T \); the operator \( \bullet \) in (4.5) indicates component-wise multiplication; \( U = (U_1, \ldots, U_m)^T \) with \( U_i = (s_{X,i}^2/n_1 + s_{Y,i}^2/n_2)^{-1/2} \). For fixed \( m \), the CLT gives
\[
\sqrt{n_1 + n_2} D \xrightarrow{P} (W_1, \ldots, W_m)^T, \tag{4.6}
\]
as \( n_1 \to \infty \) and \( n_2 \to \infty \), where \( (W_1, \ldots, W_m)^T \sim N(\nu, \Omega) \) for some \( \nu \in \mathbb{R}^m \) and positive definite matrix \( \Omega = (\omega_{ij})_{1 \leq i,j \leq m} \). Similarly, the law of large numbers gives \( s_{X,i}^2 \xrightarrow{P} \beta^T_{X,i} \Sigma_u \beta_{X,i} + \sigma^2_{\varepsilon,i} \) and \( s_{Y,i}^2 \xrightarrow{P} \beta^T_{Y,i} \Sigma_v \beta_{Y,i} + \sigma^2_{\varepsilon,i} \), implying
\[
(n_1 + n_2)^{-1/2} U_i \xrightarrow{P} c_i, \quad 1 \leq i \leq m, \tag{4.7}
\]
where \( c_i = \{(\beta^T_{X,i} \Sigma_u \beta_{X,i} + \sigma^2_{\varepsilon,i})/\rho + (\beta^T_{Y,i} \Sigma_v \beta_{Y,i} + \sigma^2_{\varepsilon,i})/(1-\rho)\}^{-1/2} \), and thus
\[
(n_1 + n_2)^{-1/2} U \xrightarrow{P} c, \tag{4.8}
\]
with \( c = (c_1, \ldots, c_m)^T \). By Slutsky’s theorem [10], (4.5), (4.6) and (4.8) imply that
\[
(T_{1;n_1,n_2}^{\text{general}}, \ldots, T_{m;n_1,n_2}^{\text{general}})^T \xrightarrow{P} (Z_1, \ldots, Z_m)^T \sim N(\tilde{\nu}, \tilde{\Omega}), \tag{4.9}
\]
where \( Z_i = c_i W_i, \tilde{\nu} = c \bullet \nu \) and \( \tilde{\Omega} = (c_i c_j \omega_{ij})_{1 \leq i,j \leq m} \).

The joint Gaussianity of test statistics in (4.9) makes it feasible to apply the factor model method in [14] to decompose \( \tilde{\Omega} \) and then to control FDP and FDR asymptotically. On the other hand, this method relies on knowing \( \tilde{\Omega} \) in advance, and thus techniques of estimating high-dimensional covariance matrices will be required to estimate \( \tilde{\Omega} \). Our Gaussian calibration helps to simplify its diagonal entries to be 1.

### 4.2 Dependence between groups and within a group: Model I

Consider observations \( \{X_{i,j}\} \) and \( \{Y_{i,j}\} \) following Model I,
\[
X_{i,j} = \mu_{X,i} + \varepsilon_{i,j} + w_i/2, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_1, \tag{4.10}
\]
\[
Y_{i,j} = \mu_{Y,i} + e_{i,j} + w_i/2, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_2, \tag{4.10}
\]
where errors \( \{w_1, \ldots, w_m\} \sim \text{i.i.d. } N(0, \sigma_w^2) \) with \( \sigma_w^2 \in (0, \infty) \); for each \( i \), errors \( \{\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i}\} \sim \text{i.i.d. } N(0, \sigma_{\varepsilon_i}^2) \), errors \( \{\varepsilon_{i,1}, \ldots, e_{i,n_i}\} \sim \text{i.i.d. } N(0, \sigma_{\varepsilon_i}^2) \), and \( \{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i}), (e_{i,1}, \ldots, e_{i,n_i}), w_i\} \) are mutually independent; \( \{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i}; e_{i,1}, \ldots, e_{i,n_i}; w_i) : i \in I_0\} \) are independent. It follows that two-sample \( t \)-statistics reduce to the following forms,

\[
T_{i;n_1,n_2}^{\text{general}} = \frac{\bar{\varepsilon}_i - \bar{\varepsilon}_i}{\sqrt{s_{\varepsilon_i}^2/n_1 + s_{\varepsilon_i}^2/n_2}}, \quad T_{i;n_1,n_2}^{\text{pool}} = \frac{\bar{\varepsilon}_i - \bar{\varepsilon}_i}{s_{\text{pool}, \varepsilon, i}^{1/n_1 + 1/n_2}}, \quad T_{i;n_1,n_2}^{\text{pool;A}} = \frac{T_{i;n_1,n_2}^{\text{pool}}}{\sigma_2(\hat{\alpha}, \varepsilon, e, i)}.
\]

(4.11)

It is interesting to note that this dataset involves dependence between different groups, as well as within a same group, but test statistics (using either \( T_{i;n_1,n_2}^{\text{general}} \) or \( T_{i;n_1,n_2}^{\text{pool}} \) or \( T_{i;n_1,n_2}^{\text{pool;A}} \)) associated with true nulls are independent. Moreover, Model I in the case of \( \sigma_w^2 = 0 \) reduces to the counterpart of model (2.1).

Regarding Model I, we can show two distributional results below for the “general” two-sample \( t \)-statistic \( T_{i;n_1,n_2}^{\text{general}} \) under \( H_{0,i} \):

(c1) if \( \sigma_{\varepsilon_i}^2 = \sigma_{\varepsilon_i}^2 \) 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 \( T_{i;n_1,n_2}^{\text{general}} \sim N(0,1) \).

Hence, 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 make two conclusions below.

(d1) If \( \sigma_{\varepsilon_i}^2 = \sigma_{\varepsilon_i}^2 \), 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 \( \{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^m \).

(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

\( T_{i;n_1,n_2}^{\text{pool}} \sim N(0, \sigma_2(\rho, \varepsilon, e, i)) \). Similar to the discussion in Section 3.2, there will be no guarantee in the case of \( \sigma_2(\rho, \varepsilon, e, i) > 1 \) for achieving level bounds \( \alpha \) in (2.11) and (2.12) using \( \{T_{i;n_1,n_2}^{\text{pool}}\}_{i=1}^m \).
But according to (4.11), the “adaptively pooled” version satisfies $T^\text{pool;A}_{i;n_1,n_2} \overset{D}{\rightarrow} N(0, 1)$, and thus the $N(0, 1)$ 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 which is similar to Model I, except that the signs of error terms $w_i/2$ in $X_{i,j}$ are switched to be negative, yielding Model II,

\[
X_{i,j} = \mu_{X,i} + \varepsilon_{i,j} - w_i/2, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_1, \\
Y_{i,j} = \mu_{Y,i} + e_{i,j} + w_i/2, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_2.
\]  

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 two-sample $t$-statistics can be derived as follows,

\[
T^\text{general}_{i;n_1,n_2} = \frac{\bar{Z}_i - \bar{E}_i - w_i}{\sqrt{s_{\varepsilon_i}^2/n_1 + s_{e_i}^2/n_2}}, \quad T^\text{pool}_{i;n_1,n_2} = \frac{\bar{Z}_i - \bar{E}_i - w_i}{s^\text{pool}_{\varepsilon_i} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad T^\text{pool;A}_{i;n_1,n_2} = \frac{T^\text{pool}_{i;n_1,n_2}}{\sigma_{p;\hat{\theta}_{(e,e,i)}}},
\]  

which differ from those in (4.11). Again, dependence between different groups, as well as within a same group, exist in the dataset, where the extent of dependence is captured by the magnitude of $\sigma_w^2$, but two-sample $t$-statistics associated with 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 $\sigma_{\varepsilon_i}^2 = \sigma_{e_i}^2 = \sigma_i^2$ and $n_1 = n_2$, then $T^\text{general}_{i;n_1,n_2} \sim t_{2n_1-2} \times f_1$, where $f_1 = \sqrt{1 + \frac{n_1}{2} \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}}$;

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

\[
T^\text{general}_{i;n_1,n_2} = Z \times f_2 \{1 + o_p(1)\} \overset{P}{\rightarrow} \infty, \quad \text{where} \quad f_2 = \sqrt{1 + \frac{n_1 n_2 \sigma_w^2}{n_2 \sigma_{\varepsilon_i}^2 + n_1 \sigma_e^2}},
\]  

$Z \sim N(0, 1)$ and $\overset{P}{\rightarrow}$ denotes converges in probability.
We can also show that $T_{i:n_1,n_2}^{pool:A}$ has the same limit null distribution as $T_{i:n_1,n_2}^{general}$. For the null distribution of the "pooled" two-sample $t$-statistic $T_{i:n_1,n_2}^{pool}$, we make two conclusions below.

(f1) If $\sigma_{e_i}^2 = \sigma_{e_i}^2 = \sigma_v^2$, then $T_{i:n_1,n_2}^{pool} \sim t_{n_1+n_2-2} \times f_3$, where $f_3 = \sqrt{1 + \frac{n_1 n_2}{n_1+n_2} \frac{\sigma_v^2}{\sigma_e^2}}$.

(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

$$T_{i:n_1,n_2}^{pool} = Z \times f_4\{1 + o_P(1)\} \to \infty,$$

where $f_4 = \sqrt{\frac{(1-\rho) + \rho \sigma_v^2 / \sigma_e^2 + \frac{n_1 n_2}{n_1+n_2} \frac{\sigma_v^2}{\sigma_e^2}}{\rho (1-\rho) \sigma_v^2 / \sigma_e^2}}$. (4.15)

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

5 Simulation study

We assess the finite sample performance of the calibration method applied to two-sample $t$-test statistics $\{T_{i:n_1,n_2}^{general}\}_{i=1}^m$, $\{T_{i:n_1,n_2}^{pool}\}_{i=1}^m$, $\{T_{i:n_1,n_2}^{pool:A}\}_{i=1}^m$, $\{T_{i:n_1,n_2}^{adjust:T}\}_{i=1}^m$, $\{T_{i:n_1,n_2}^{adjust:E}\}_{i=1}^m$, $\{T_{i:n_1,n_2}^{2,stage}\}_{i=1}^m$, as the total sample size $n = n_1 + n_2$ varies. For each $k \in \{1,2\}$, conduct the simulation 1000 times. In each simulation, we calculate the numbers of false rejections $V_m(t_{\alpha;m}^a)$ and $V_m(t_{\alpha;m;k}^a)$. The empirical estimates of $FWER(t_{\alpha;m}^a)$ and $FWER_k(t_{\alpha;m;k}^a)$ are the proportion of times that $\{V_m(t_{\alpha;m}^a) \geq 1\}$ and $\{V_m(t_{\alpha;m;k}^a) \geq k\}$ occur in 1000 simulations. Set $\alpha = 0.05$ to be the control level. The "two-stage" $t$-tests use level 0.05 in the 1st-stage. A range of sample sizes are considered, with $n_1 = 10c$ and $n_2 = 20c$, $c \in \{1,2,\ldots,10\}$, 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

$$X_{i,j} = \mu_{X_{i,j}} + \varepsilon_{i,j} + \text{sign}X_{i,j} w_i/2, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_1,$$

$$Y_{i,j} = \mu_{Y_{i,j}} + e_{i,j} + \text{sign}Y_{i,j} w_i/2, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_2,$$

(5.1)
where $\mu_{X,i} = \mu_{Y,i} = 1$ for $i = 1, \ldots, m$, whereas 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 $i = m + 1, \ldots, m$, and $\{\varepsilon_{i,j}\}$ are independent of $\{e_{i,j}\}$. Also, errors $\{w_1, \ldots, w_m\} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$ as described below (4.10).

Note that (5.1) includes models (2.1), (4.10) and (4.12):

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

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

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

In model (5.1), schemes for errors $\{\varepsilon_{i,j}\}$ and $\{e_{i,j}\}$ are considered in Examples 1–5 as follows. **Example 1**: $\{\varepsilon_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$; $\{e_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, with $\sigma = 1.0$. **Example 2**: $\{\varepsilon_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)\}; \{e_{i,j}\} \overset{i.i.d.}{\sim} t_4$. **Example 3**: $\{\varepsilon_{i,j}\} \overset{i.i.d.}{\sim} t_4; \{e_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. **Example 4**: $\{\varepsilon_{i,j}\} \overset{i.i.d.}{\sim} \chi^2_2 - 2; \{e_{i,j}\} \overset{i.i.d.}{\sim} - (\chi^2_2 - 2)$. **Example 5**: $\{\varepsilon_{i,j}\} \overset{i.i.d.}{\sim} \chi^2_4 - 4; e_{i,j} = (2b_i - 1) u_{i,j}$, where $\{u_{i,j}\} \overset{i.i.d.}{\sim} \{\exp(1/\lambda) - \lambda\}$ with $\lambda = 4$, and coefficients $b_i$ are non-random, equal to sampled values of $b_i^*$, with $\{b_1^*, \ldots, b_m^*\} \overset{i.i.d.}{\sim} \text{Ber}(1/2)$. Here, Examples 4 and 5 intend to assess the skewness effects of the two-sample $t$-tests on the calibration methods.

Moreover, in model (5.1), variances $\sigma_w^2$ of 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, a combination of errors $\{\varepsilon_{i,j}, e_{i,j}\}$ and errors $\{w_i\}$ in model (5.1) yields 15 examples denoted as follows:

“Example 1”, \ldots, “Example 5”: for independent data;

“Example 1(I), \ldots, “Example 5(I)” for dependent data;

“Example 1(II), \ldots, “Example 5(II)” for dependent data.

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

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

In Example 1 with Gaussian errors, the top row panels of Figure 3 indicate that the estimated FWER($t_{i; n_1, n_2}^a$) of $\{T_{i; n_1, n_2}^{\text{general}}\}$ gets closer to 0.05 as the sample size $n$ increases. The N(0, 1) calibration applied to $\{T_{i; n_1, n_2}^{\text{pool}}\}$ performs similarly to that of $\{T_{i; n_1, n_2}^{\text{general}}\}$, due to equal population variances such that $\sigma_{\epsilon_i}^2 = 1$ in Example 1. In this case, there is also no adverse effect of using the “adaptively pooled” version $\{T_{i; n_1, n_2}^{\text{pool}; A}\}$. The calibration methods applied to $\{T_{i; n_1, n_2}^{\text{adjust}; T}\}$, $\{T_{i; n_1, n_2}^{\text{adjust}; E}\}$ and $\{T_{i; n_1, n_2}^{2, \text{stage}}\}$ perform similarly to that of $\{T_{i; n_1, n_2}^{\text{general}}\}$, due to symmetric distributions of $\{\epsilon_{i,j}\}$ and $\{\epsilon_{i,j}\}$.

In addition, recall from part (b1) in Section 2.1 that $T_{i; n_1, n_2}^{\text{pool}}$ in Example 1 exactly follows the $t_{n_1+n_2-2}$-distribution under the null. Hence, the second column panels of Figure 3 and Figure 8 also overlay the true values (using read lines) of FWER($t_{i; n_1, n_2}^a$) and FWER$_k(t_{i; n_1, n_2}^a)$ respectively, which match well with their empirical counterparts. This supports the validity of the simulation set-ups. Likewise, the left column of Figure 13 compares the false discovery proportion (abbreviated as FDP, defined as the number of false rejections divided by the number of rejections) of the Benjamini-Hochberg (BH) multiple testing procedure implemented in the following ways: approximate $p$-values calculated from the approximate N(0, 1)-distributions for $T_{i; n_1, n_2}^{\text{general}}$, $T_{i; n_1, n_2}^{\text{pool}}$, $T_{i; n_1, n_2}^{\text{pool}; A}$, $T_{i; n_1, n_2}^{\text{adjust}; T}$, $T_{i; n_1, n_2}^{\text{adjust}; E}$ and $T_{i; n_1, n_2}^{2, \text{stage}}$, exact $p$-values calculated from the exact $t_{n_1+n_2-2}$-distribution for $T_{i; n_1, n_2}^{\text{pool}}$. As we observe, when $n$ approaches 100 and larger, the FDPs using the N(0, 1) calibration mimic that using the exact distribution.

In Examples 2–3 with non-Gaussian errors, population variances are $\sigma_{\epsilon_i}^2 < \sigma_{\epsilon_i}^2$ in Example 2, while $\sigma_{\epsilon_i}^2 > \sigma_{\epsilon_i}^2$ in Example 3. Figures 4 and 5 indicate that within each example, there is little difference between performances of the calibration method applied
to test statistics \(\{T_{i:n_1,n_2}^{\text{general}}\}, \{T_{i:n_1,n_2}^{\text{pool;A}}\}, \{T_{i:n_1,n_2}^{\text{adjust;T}}\}, \text{and} \{T_{i:n_1,n_2}^{2,\text{stage}}\}\). However, \(\{T_{i:n_1,n_2}^{\text{pool}}\}\) behave substantially differently in Example 2 and Example 3, where FWERs are conservatively controlled in Example 2 (as seen in the top row, second column panel of Figure 4), but out of control 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 \(\sigma^2_{\theta_{(e,c);i}} < 1\) in Example 2 with \(n_1 < n_2\) and \(\sigma^2_{\varepsilon;i} < \sigma^2_{\varepsilon;i}\), whereas \(\sigma^2_{\theta_{(e,c);i}} > 1\) in Example 3 with \(n_1 < n_2\) and \(\sigma^2_{\varepsilon;i} > \sigma^2_{\varepsilon;i}\). The comparison thus supports that the “adaptively pooled” version \(T_{i:n_1,n_2}^{\text{pool;A}}\) is a valid substitute for the originally “pooled” version \(T_{i:n_1,n_2}^{\text{pool}}\) and compares as well as the “general” version \(T_{i:n_1,n_2}^{\text{general}}\).

Moreover, in Example 3, since the 6th moment doesn’t exist for the \(t_4\)-distribution, \(\hat{\mu}_{3,X}/n_1^2 - \hat{\mu}_{3,Y}/n_2^2\) performs poorly in estimating \(\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2\). Thus, \(T_{i:n_1,n_2}^{\text{adjust;E}}\) deviates much from \(T_{i:n_1,n_2}^{\text{adjust;T}}\), as seen in Figure 5. Nonetheless, \(T_{i:n_1,n_2}^{2,\text{stage}}\) is as good as \(T_{i:n_1,n_2}^{\text{general}}\).

Recall that for Examples 1–3, \(\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2 = 0\) (as in Table 1), and thus \(T_{i:n_1,n_2}^{\text{adjust;T}}\) and \(T_{i:n_1,n_2}^{\text{general}}\) are identical and also the best, and \(T_{i:n_1,n_2}^{2,\text{stage}}\) compares well with \(T_{i:n_1,n_2}^{\text{general}}\). 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 \(T_{i:n_1,n_2}^{\text{adjust;E}}\) is better than \(T_{i:n_1,n_2}^{\text{general}}\), and \(T_{i:n_1,n_2}^{2,\text{stage}}\) is close to the better one of \(T_{i:n_1,n_2}^{\text{general}}\) and \(T_{i:n_1,n_2}^{\text{adjust;E}}\). The theoretical \(T_{i:n_1,n_2}^{\text{adjust;T}}\) can still control FWER in the best way. 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 dependence mechanism in Examples \(\ell(I), \ell = 1, \ldots, 5\), it is apparent that top row panels and middle row panels of Figures 3, 12 are nearly indistinguishable, regardless of the magnitude of \(\sigma_w > 0\). This agrees with the analysis in Section
By the same argument, the calculated FDPs of the BH procedure in the left column panels of Figure 13 resemble those in the middle column panels of Figure 13.

In striking contrast, for Examples ℓ(II), ℓ = 1, . . . , 5, with dependence mechanism described by Model II, loss of controlling FWER₁ and FWER₂ is noticeably shown in bottom row panels 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 panels of Figure 13 reveal that the FDPs based on the N(0, 1) calibration for approximating p-values no longer mimic the actuals ones. Again, this is caused by the fact that when data are generated from Model II, the variances in the asymptotic distribution of T_{general}^{i; n_1, n_2} (as well as T_{pool; A}^{i; n_1, n_2} and T_{pool}^{i; n_1, n_2}) 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_{pool}^{i; n_1, n_2} in Example 1(II), continues to perform well.

6 Real data examples

We apply the Gaussian calibration for two-sample t-tests to the analysis of three real data sets. As expected, Table 2 reveals the discrepancy between results delivered by the “pooled” and “general” versions. Nonetheless, results based on the “adaptively pooled” version always agree well with those of the “general” version. This lends further support 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 dataset of [12], which contains genetic expression levels of 6033 genes, obtained for 102 men, with 50 normal control subjects, and 52 prostate cancer patients. The primary goal of this study was to discover a small number of “interesting” genes, whose expression levels differ between the prostate and normal subjects. Using the BH multiple testing procedure, Table 2 compares the number of genes which are detected to be significant, where p-values are calculated from approaches: the N(0, 1)-
distribution for $T_{i:n_1,n_2}^{\text{general}}$, the $t_{n_1+n_2-2}$-distribution for $T_{i:n_1,n_2}^{\text{pool}}$, the $\mathcal{N}(0,1)$-distributions for $T_{i:n_1,n_2}^{\text{pool}}$, $T_{i:n_1,n_2}^{\text{pool;A}}$ and $T_{i:n_1,n_2}^{2,\text{stage}}$. Recall that 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 (through the $t$-distribution) and 51 and 50 (through $\mathcal{N}(0,1)$ calibration methods) could be caused by either non-Gaussian samples or unequal population variances so that $T_{i:n_1,n_2}^{\text{pool}}$ 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 dataset includes gene expression for $m = 8648$ genes using prostate cell populations from low-grade (with $n_1 = 27$) and high-grade (with $n_2 = 17$) samples of cancerous tissue. Using the BH multiple testing procedure, where $p$-values are calculated according to the $\mathcal{N}(0,1)$-distribution for $T_{i:n_1,n_2}^{\text{general}}$, the $t_{n_1+n_2-2}$-distribution for $T_{i:n_1,n_2}^{\text{pool}}$, the $\mathcal{N}(0,1)$-distributions for $T_{i:n_1,n_2}^{\text{pool}}$, $T_{i:n_1,n_2}^{\text{pool;A}}$ and $T_{i:n_1,n_2}^{2,\text{stage}}$, the numbers of genes declared to be significant are 565, 196, 436, 563, 565 respectively. See Table 2. In this example, the detection difference between using the $t$-distribution and approximate $\mathcal{N}(0,1)$-distribution could be caused by either non-Gaussian samples or unequal population variances so that the $t_{n_1+n_2-2}$-distribution may not be valid for $T_{i:n_1,n_2}^{\text{pool}}$. It may also be due to the sample size $n = 44$ not large enough for the Gaussian calibration. Interestingly, the “adaptively pooled” two-sample $t$-statistics $\{T_{i:n_1,n_2}^{\text{pool;A}}\}$ continue to detect a comparable number of significant genes to the “general” counterparts $\{T_{i:n_1,n_2}^{\text{general}}\}$.

As a third illustration, we analyze the Acute Lymphoblastic Leukemia (ALL) dataset. Refer to [5] for details of the ALL data, with 12625 genes measured for two groups of samples sizes 37 and 42. Table 2 presents the number of genes called differentially expressed in the BCR/ABL versus NEG comparison, for the four methods. The “pooled” two-sample
t-statistics $T_{i:n_1,n_2}^{\text{pool}}$ through the $t_{n_1+n_2-2}$-distribution identify 169 genes (identical to that given in Table S2 of [5]), but differ from detection results using 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 continue to be comparable.

7 Discussion

We examined the validity of the calibration method when 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 $m$ of tests. Relationship between $m$ and $(n_1, n_2)$ is studied to ensure the control of the overall level accuracy, as well as to support control of FDR for some multiple testing procedures. Distinction between the choice of “general” and “pooled” two-sample t-statistics is made in cases where the typical form of independence assumption among tests either holds or is violated. The proposed “adaptively pooled” two-sample t-statistics, when used simultaneously in the calibration method, perform as 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 errors are symmetric, but outperforms when 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 “general”, “two-stage” and “adaptively pooled” two-sample t-statistics.

Dependence structure poses challenges to controlling the overall significance level, and
FDR. In Section 4, we demonstrated that FWER and FWER$_k$ could be controlled under arbitrary dependence among tests, but FDR would be out of control if we just 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, so that both FDP and FDR can be controlled asymptotically. Besides, we also addressed explicitly the performances of “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 groups and within a group.

Several issues are desirable for future research. First, the bootstrap method provides an alternative way 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 much heavier computational cost. Second, the issue on how to improve the power of a given multiple testing procedure, when the $p$-values need to be approximated, may require a study 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 nulls”. In practice, information of $m_0$ or $\pi_0$ can be either learned from prior knowledge or estimated via empirical procedures [3, 18, 21]. If the resulting $\pi_0$ is close to 0, it is more reasonable to utilize other approaches which suit the dense case well.

**Supplementary materials:** All technical details, figures and tables are relegated to the supplementary material which is available online.

**Acknowledgements:** The authors thank the editor, associate editor, and an anonymous referee for insightful comments. C. Zhang’s research is supported by the U.S. NSF grant DMS–1712418, and Wisconsin Alumni Research Foundation.
References


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


Appendix A: Conditions and Proofs of Main Results

Conditions. The following technical conditions are not the weakest possible, but facilitate the derivations.

A1. For each \(i, \varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}\) are i.i.d. with \(E(\varepsilon_{i,1}) = 0\) and \(\text{var}(\varepsilon_{i,1}) = \sigma_{\varepsilon_{i,1}}^2 \in (0, \infty)\).

A2. For each \(i, e_{i,1}, \ldots, e_{i,n_2}\) are i.i.d. with \(E(e_{i,1}) = 0\) and \(\text{var}(e_{i,1}) = \sigma_{e_{i,1}}^2 \in (0, \infty)\).

A3. For each \(i, (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1})\) is independent of \((e_{i,1}, \ldots, e_{i,n_2})\).

A4. For each \(i, E(|\varepsilon_{i,1}|^3) < \infty\) and \(E(|e_{i,1}|^3) < \infty\).

A5. Two-sample \(t\)-statistics corresponding to true nulls are identically distributed.

A5’. Two-sample \(t\)-statistics corresponding to true non-nulls are identically distributed.

A6. There are constants \(c_1\) and \(c_2\) satisfying \(0 < c_1 \leq c_2 < \infty\), such that \(c_1 \leq n_1/n_2 \leq c_2\).

A7. Two-sample \(t\)-statistics corresponding to true nulls are independent.

A7’. Two-sample \(t\)-statistics are independent.

A8. Let \(F_{0;T}(\cdot; n)\) and \(F_{1;T}(\cdot; n)\) denote the C.D.F. of two-sample \(t\)-statistics under the true null and non-null, respectively.

A9. The marginal C.D.F. and p.d.f. of two-sample \(t\)-statistics are \(F_T(\cdot; n) = \pi_0 F_{0;T}(\cdot; n) + (1 - \pi_0) F_{1;T}(\cdot; n)\) and \(f_T(\cdot; n) = F_T'(\cdot; n)\), where \(f_T(t; n)\) is Lipschitz continuous in \(t\) uniformly in \(n\).

A10. The marginal C.D.F. of true \(p\)-values \(\{P_i\}\) is \(F_P(\cdot; n) = \pi_0 F_{0,P}(\cdot; n) + (1 - \pi_0) F_{1,P}(\cdot; n)\), where \(F_{0,P}(\cdot; n)\) is the C.D.F. of the standard uniform distribution, and \(F_{1,P}(\cdot; n)\) is the C.D.F. of \(\{P_i\}\) under the true non-null. Assume \(F_{1,P}(t; n)\) is continuous in \(t\).
Note that condition A5 is valid when \( \{ \varepsilon_{i,1} : i \in I_0 \} \) are identically distributed and \( \{ e_{i,1} : i \in I_0 \} \) are identically distributed. Condition A7 holds if \( \{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}; e_{i,1}, \ldots, e_{i,n_2}) : i \in I_0 \} \) are independent.

We first present Lemma 1, which will be used in proving Propositions 1, 2, 4 and 5.

**Notation.** For sequences \( \{a_n\}_{n \geq 1} \) and \( \{b_n\}_{n \geq 1} \), \( a_n \asymp b_n \) denotes \( \lim_{n \to \infty} a_n/b_n = 1 \).

**Lemma 1** Assume model (2.1) and conditions A1–A6, and the general two-sample \( t \)-statistics \( \{T_{i; n_1, n_2}^{\text{general}}\}_{i = 1}^m \) are used. Assume \( \alpha \in (0, 1) \), \( m_0/m \to \pi_0 \in (0, 1) \), \( m \to \infty \), \( n \to \infty \), and \((m, n)\) satisfies (3.1).

(i) If \( t_{\alpha;m}^a \) is given in (3.2) and \((m, n)\) satisfies (3.1), then

\[
\max_{i \in I_0} \alpha_{i; n_1, n_2}(t_{\alpha;m}^a) \leq \frac{\beta_{1;\alpha}}{m} \{1 + o(1)\}, \quad \text{and} \quad \sum_{i \in I_0} \alpha_{i; n_1, n_2}(t_{\alpha;m}^a) \to \pi_0 \beta_{1;\alpha}, \quad (A.1)
\]

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

(ii) If \( t_{\alpha;m;k}^a \) is given in (3.4) and \((m, n)\) satisfies (3.1), then for \( k \geq 2 \),

\[
\max_{i \in I_0} \alpha_{i; n_1, n_2}(t_{\alpha;m;k}^a) \leq \frac{\beta_{k;\alpha}}{m} \{1 + o(1)\}, \quad \text{and} \quad \sum_{i \in I_0} \alpha_{i; n_1, n_2}(t_{\alpha;m;k}^a) \to \pi_0 \beta_{k;\alpha}, \quad (A.2)
\]

where \( \beta_{k;\alpha} \) solves the equation (3.5).

**Proof:** We first show part (i). For \( t_{\alpha;m}^a \) given in (3.2) and \( N(0, 1) \) random variables \( \{T_i^a\}_{i = 1}^m \), we obtain

\[
\alpha_i^a(t_{\alpha;m}^a) = P(|T_i^a| > t_{\alpha;m}^a) = 2\{1 - \Phi(t_{\alpha;m}^a)\} = 1 - (1 - \alpha)^{1/m}, \quad i = 1, \ldots, m, \quad (A.3)
\]

and as \( m \to \infty \),

\[
\max_{i \in I_0} \alpha_i^a(t_{\alpha;m}^a) = 1 - (1 - \alpha)^{1/m} = \frac{\beta_{1;\alpha}}{m} \{1 + o(1)\} = o(1), \quad (A.4)
\]

\[
\sum_{i \in I_0} \alpha_i^a(t_{\alpha;m}^a) = m_0 \{1 - (1 - \alpha)^{1/m}\} = \pi_0 \beta_{1;\alpha} \{1 + o(1)\}, \quad (A.5)
\]
where
\[ \beta_{1,a} \equiv - \log(1 - \alpha) \in (0, \infty). \]  
(A.6)

For \( \alpha_{i:1,n_2}(t) \), it can be rewritten as
\[ \alpha_{i:1,n_2}(t) = \alpha_i^a(t) + d_i(t), \]
where
\[
\begin{align*}
  d_i(t) &= \alpha_i^a(t) \left\{ \frac{\alpha_{i:1,n_2}(t)}{\alpha_i^a(t)} - 1 \right\}, \\
  |d_i(t)| &= \alpha_i^a(t) \left| \frac{\alpha_{i:1,n_2}(t)}{\alpha_i^a(t)} - 1 \right|.
\end{align*}
\]
(A.7)

This leads to
\[
\begin{align*}
  \alpha_{i:1,n_2}(t) &\leq \alpha_i^a(t) + |d_i(t)|, \\
  \max_{i \in I_0} \alpha_{i:1,n_2}(t) &\leq \max_{i \in I_0} \alpha_i^a(t) + \max_{i \in I_0} |d_i(t)|, \\
  \sum_{i \in I_0} \alpha_{i:1,n_2}(t) &= \sum_{i \in I_0} \alpha_i^a(t) + \sum_{i \in I_0} d_i(t).
\end{align*}
\]
(A.9)

Thus, if the condition
\[
\max_{i \in I_0} \left| \frac{\alpha_{i:1,n_2}(t_{a;m})}{\alpha_i^a(t_{a;m})} - 1 \right| = o(1)
\]
(A.11)
holds, then (A.7), (A.4), (A.11) and (A.8) imply that
\[
\max_{i \in I_0} |d_i(t_{a;m})| \leq \frac{\beta_{1,a}}{m} \{1 + o(1)\} o(1) = \frac{\beta_{1,a}}{m} o(1), \\
\sum_{i \in I_0} d_i(t_{a;m}) \leq \pi_0 \beta_{1,a} o(1),
\]
(A.12)

which combined with (A.9), (A.4), (A.10), and (A.5) gives
\[
\begin{align*}
  \max_{i \in I_0} \alpha_{i:1,n_2}(t_{a;m}) &\leq \frac{\beta_{1,a}}{m} \{1 + o(1)\} + \frac{\beta_{1,a}}{m} o(1) = \frac{\beta_{1,a}}{m} \{1 + o(1)\}, \\
  \sum_{i \in I_0} \alpha_{i:1,n_2}(t_{a;m}) &= \pi_0 \beta_{1,a} \{1 + o(1)\} + o(1) = \pi_0 \beta_{1,a} \{1 + o(1)\}.
\end{align*}
\]
(A.13)

Hence condition (A.11) indeed implies (A.1).

Now, we justify that (A.11) holds. Recall
\[ \alpha_i^a(t) = P(|T_i^a| > t) \]
\[
\alpha_{i;\mathcal{I}_2}(t) = P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma}) > t)
= P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t) + P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < -t).
\]

It follows that
\[
\frac{\alpha_{i;\mathcal{I}_2}(t)}{\alpha_{i}(t)} - 1 = \frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t) + P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < -t)}{2\{1 - \Phi(t)\}} - 1
= \frac{1}{2}\left\{\frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t) + P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < t)}{1 - \Phi(t)} - 2\right\}
= \frac{1}{2}\left\{\frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t)}{1 - \Phi(t)} - 1\right\} + \frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < t)}{1 - \Phi(t)} - 1\right]\right\},
\]

and thus
\[
\max_{i \in \mathcal{I}_0} \left| \frac{\alpha_{i;\mathcal{I}_2}(t)}{\alpha_{i}(t)} - 1 \right|
= \frac{1}{2} \max_{i \in \mathcal{I}_0} \left\{\frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t)}{1 - \Phi(t)} - 1\right\} + \frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < t)}{1 - \Phi(t)} - 1\right\}
\le \max_{i \in \mathcal{I}_0} \left| \frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} > t)}{1 - \Phi(t)} - 1\right| + \max_{i \in \mathcal{I}_0} \left| \frac{P_{H_0,i}(T_{i;\mathcal{I}_2}^{\gamma} < t)}{1 - \Phi(t)} - 1\right|.
\]

\[\text{(A.14)}\]

From (A.3), we observe that
\[
\Phi(t_{a;\alpha}^a) = 1 - 1 - (1 - \alpha)^{1/m} = 1 - \beta_{1;\alpha}/m \{1 + o(1)\} \rightarrow 1,
\]

as \(m \rightarrow \infty\), where \(\beta_{1;\alpha}\) is as defined in (A.6), and thus we conclude \(t_{\alpha;m}^a \rightarrow \infty\). To find the explicit convergence rate of \(t_{\alpha;m}^a\) defined in (3.2), we use the tail probability ([10], p. 655) of a \(\mathcal{N}(0, 1)\) distribution,
\[
1 - \Phi(t_{\alpha;m}^a) \asymp \frac{1}{t_{\alpha;m}^a \sqrt{2\pi}} e^{-(t_{\alpha;m}^a)^2/2}.
\]

\[\text{(A.16)}\]

Combining (A.15) and (A.16) gives
\[
\frac{1}{\sqrt{2\pi}} \frac{1}{t_{\alpha;m}^a \sqrt{2\pi}} e^{-(t_{\alpha;m}^a)^2/2} \asymp \frac{\beta_{1;\alpha}/m}{2},
\]

which is equivalent to
\[
\frac{m}{\beta_{1;\alpha}^2}.
\]

This gives \((t_{\alpha;m}^a)^2 = O(\log(m))\), i.e., \(t_{\alpha;m}^a = O((\log(m))^{1/2})\), which together with (3.1) gives \(t_{\alpha;m}^a = o(n^{1/6})\). An application of Theorem 1.2 of [6] to (A.14), together with
condition A5, give \( \max_{i \in I_0} \left| \frac{\alpha_i(t_{m;\alpha}) - \beta_{k;\alpha}}{m} \right| = o(1) \) as \( m \to \infty \) and \( n \to \infty \). Hence (A.11) is verified.

Next, we show part (ii). The critical values \( t_{\alpha,m;k}^{a} \) given in (3.4) satisfy

\[
\alpha_i^a(t_{\alpha,m;k}) = P(|T_i^a| > t_{\alpha,m;k}^a) = 2\{1 - \Phi(t_{\alpha,m;k}^a)\} = \frac{\beta_{k;\alpha}}{m}, \quad i = 1, \ldots, m, \tag{A.17}
\]

where (3.5) implies that \( \beta_{k;\alpha} \in (0, \infty) \). Thus, we obtain

\[
\max_{i \in I_0} \alpha_i^a(t_{\alpha,m;k}) = \frac{\beta_{k;\alpha}}{m}, \quad \sum_{i \in I_0} \alpha_i^a(t_{\alpha,m;k}) = \frac{m}{m} \beta_{k;\alpha} = \pi_0 \beta_{k;\alpha} + o(1).
\]

Also, \( \Phi(t_{\alpha,m;k}^a) = 1 - \frac{\beta_{k;\alpha}/2}{m} \to 1 \). The rest of the proof is similar to that used in part (i). \( \blacksquare \)

**Proof of Proposition 1**. From (2.4) and condition A7, it suffices to consider \( N(0,1) \) random variables \( \{T_i^a\}_{i=1}^m \), with \( \{T_i^a : i \in I_0\} \) being independent. Direct calculations give

\[
\text{FWER}_1^a(t_{\alpha;m}) = P\left( \sum_{i \in I_0} I(|T_i^a| > t_{\alpha;m}^a) \geq 1 \right) \\
= P\left( \bigcup_{i \in I_0} \{|T_i^a| > t_{\alpha;m}^a\} \right) \\
= 1 - P\left( \bigcap_{i \in I_0} \{|T_i^a| > t_{\alpha;m}^a\} \right) \\
= 1 - \prod_{i \in I_0} P\left( \{|T_i^a| > t_{\alpha;m}^a\} \right) \\
= 1 - \prod_{i \in I_0} \{1 - \alpha_i^a(t_{\alpha;m}^a)\} = 1 - (1 - \alpha)^{m_0/m}, \tag{A.18}
\]

where \( 1 - (1 - \alpha)^{m_0/m} \leq \alpha \). This shows the second part of (3.3).

To show the first part of (3.3), note that derivations similar to (A.18) together with condition A7 give \( \text{FWER}_1(t_{\alpha;m}) = 1 - \prod_{i \in I_0} \{1 - \alpha_{i;1,n_2(t_{\alpha;m})}\} \). It thus suffices to show

\[
\prod_{i \in I_0} \{1 - \alpha_{i;1,n_2(t_{\alpha;m})}\} - \prod_{i \in I_0} \{1 - \alpha_i^a(t_{\alpha;m}^a)\} = o(1). \tag{A.18}
\]

From (A.18), \( \prod_{i \in I_0} \{1 - \alpha_i^a(t_{\alpha;m}^a)\} = \alpha^{1 - \pi_0 + o(1)} = e^{-\pi_0(-\log(1-\alpha))} + o(1) = e^{-\pi_0 \beta_{1;\alpha}} + o(1) \), we thus will show that

\[
\prod_{i \in I_0} \{1 - \alpha_{i;1,n_2(t_{\alpha;m})}\} = e^{-\pi_0 \beta_{1;\alpha}} + o(1) \tag{A.19}
\]

as \( m \to \infty \) and \( n \to \infty \). According to \[19\] (Lemma 6.1.1, p. 125), (A.19) will be deduced from (A.1). The proof is completed. \( \blacksquare \)
Proof of Proposition [2]. Similar to the proof of Proposition [1] it suffices to consider
\( N(0, 1) \) random variables \( \{T_i^a\}_{i=1}^m \) with \( \{T_i^a : i \in I_\alpha\} \) being independent.

To show the first part of (3.6), note that
\[
\text{FWER}_k^a(t_{\alpha;m;k}) = \mathbb{P}(V_m^a(t_{\alpha;m;k}) \geq k) = \mathbb{P}\left( \sum_{i \in I_\alpha} \mathbb{I}(|T_i^a| > t_{\alpha;m;k}) \geq k \right)
\]
\[
= 1 - \mathbb{P}\left( \sum_{i \in I_\alpha} \mathbb{I}(|T_i^a| > t_{\alpha;m;k}) \leq k - 1 \right).
\]
\[
\text{FWER}_k(t_{\alpha;m;k}) = \mathbb{P}(V_m(t_{\alpha;m;k}) \geq k) = \mathbb{P}\left( \sum_{i \in I_\alpha} \mathbb{I}(|T_{i,n_1,n_2}^\text{general}| > t_{\alpha;m;k}) \geq k \right)
\]
\[
= 1 - \mathbb{P}\left( \sum_{i \in I_\alpha} \mathbb{I}(|T_{i,n_1,n_2}^\text{general}| > t_{\alpha;m;k}) \leq k - 1 \right).
\]

Define by \( \varphi_{m,t}^a(u) \) and \( \varphi_{m,t}^a(u) \) the characteristic functions of \( V_m(t) \) and \( V_m^a(t) \) respectively, where \( u \in \mathbb{R} \). It suffices to show that as \( m \to \infty \) and \( n \to \infty \),
\[
\varphi_{m,t}^a(u) - \varphi_{m,t}^a(u) = o(1). \tag{A.20}
\]

Direct calculations give
\[
\varphi_{m,t}^a(u) = \mathbb{E}\{e^{i u V_m^a(t)}\} = \prod_{i \in I_\alpha} \mathbb{E}\{e^{i u (|T_i^a| > t)}\}
\]
\[
= \prod_{i \in I_\alpha} [\alpha_i^a(t) e^{i u} + \{1 - \alpha_i^a(t)\}] = \prod_{i \in I_\alpha} \{1 + \alpha_i^a(t)(e^{i u} - 1)\},
\]
where \( i = \sqrt{-1} \) denotes the imaginary number. By (A.17),
\[
\max_{i \in I_\alpha} |\alpha_i^a(t_{\alpha;m;k})(e^{i u} - 1)| = \left| \frac{\beta_{k,\alpha}}{m} (e^{i u} - 1) \right| \leq 2 \times \frac{\beta_{k,\alpha}}{m} = o(1),
\]
\[
\sum_{i \in I_\alpha} |\alpha_i^a(t_{\alpha;m;k})(e^{i u} - 1)| = \sum_{i \in I_\alpha} \left| \frac{\beta_{k,\alpha}}{m} (e^{i u} - 1) \right| \leq 2 \times \frac{\beta_{k,\alpha}}{m} m_0 \leq 2\beta_{k,\alpha} < \infty,
\]
\[
\sum_{i \in I_\alpha} \alpha_i^a(t_{\alpha;m;k})(e^{i u} - 1) = \left\{ \sum_{i \in I_\alpha} \alpha_i^a(t_{\alpha;m;k}) \right\}(e^{i u} - 1) \leq \pi_0 \beta_{k,\alpha}(e^{i u} - 1) + o(1).
\]

According to [8] (a lemma on p. 208),
\[
\varphi_{m,t}^a(u) \to \exp\{\pi_0 \beta_{k,\alpha}(e^{i u} - 1)\} \tag{A.21}
\]
as \( m \to \infty \). Similarly,
\[
\varphi_{m,t}^a(u) = \mathbb{E}\{e^{i u V_m(t)}\} = \prod_{i \in I_\alpha} \mathbb{E}\{e^{i u (|T_{i,n_1,n_2}| > t)}\}
\]

38
\[
\prod_{\ell \in \mathcal{I}_0} \left[ \alpha_{\ell; n_1, n_2}(t) e^{i u} + \{1 - \alpha_{\ell; n_1, n_2}(t)\} \right] = \prod_{\ell \in \mathcal{I}_0} \left\{ 1 + \alpha_{\ell; n_1, n_2}(t) (e^{i u} - 1) \right\},
\]

Note that as \( m \to \infty \) and \( n \to \infty \), an application of (A.2) gives

\[
\max_{\ell \in \mathcal{I}_0} \left| \alpha_{\ell; n_1, n_2}(t^a_{\alpha; m; k}) (e^{i u} - 1) \right| = o(1),
\]
\[
\sum_{\ell \in \mathcal{I}_0} \left| \alpha_{\ell; n_1, n_2}(t^a_{\alpha; m; k}) (e^{i u} - 1) \right| \leq M < \infty,
\]
\[
\sum_{\ell \in \mathcal{I}_0} \alpha_{\ell; n_1, n_2}(t^a_{\alpha; m; k}) (e^{i u} - 1) \to \pi_0 \beta_k \alpha (e^{i u} - 1).
\]

Applying (a lemma on p. 208) again implies

\[
\varphi_{\nu_m(t^a_{\alpha; m; k})}(u) \to \exp \{ \pi_0 \beta_k \alpha (e^{i u} - 1) \}.
\]

Thus (A.21) and (A.22) imply (A.20).

To show the second part of (3.6), note that (A.21) yields

\[
V_m(t^a_{\alpha; m; k}) \overset{D}{\to} \text{Poisson}(\pi_0 \beta_k \alpha),
\]

where \( \text{Poisson}(\beta) \) denotes the Poisson random variable with the parameter \( \beta \). Thus as \( m \to \infty \),

\[
\text{FWER}^a_k(t^a_{\alpha; m; k}) = P(\text{Poisson}(\pi_0 \beta_k \alpha) \geq k) + o(1)
\]
\[
= G_k(\pi_0 \beta_k \alpha) + o(1).
\]

Since \( G_k(\beta) \) is monotone increasing in \( \beta \in (0, \infty) \), we obtain \( G_k(\pi_0 \beta_k \alpha) \leq G_k(\beta_k \alpha) \). This combined with (3.5) completes the proof. ■

Proof of Proposition 3. Consider \( H_{1,i} : \mu_{X,i} > \mu_{Y,i} \); the two-sided alternative can be treated similarly. It suffices to show

\[
s_{\alpha;n} - s_{\alpha:n} = o_n(1),
\]

\[
\hat{\text{FDR}}(\tau^a_{\alpha; m; n}) = \alpha + o_n(1) + O_P(m^{-1/2}),
\]

where \( o_n(1) \) denotes a term converging to zero as \( n \to \infty \).

To show (A.23), let \( c_n = F_{0; T}^{-1}(1 - s_{\alpha;n}^a; n) \) and \( d_n = \Phi^{-1}(1 - s_{\alpha;n}^a) \). Then

\[
1 - F_{0; T}(c_n; n) = s_{\alpha;n}^a = 1 - \Phi(d_n).
\]
By condition (3.7) and [6] (Theorem 1.2), we have
\[
\frac{1 - F_{0;T}(d_n; n)}{1 - \Phi(d_n)} \to 1.
\] (A.26)

Since \(T_{i,n_1,n_2}^{\text{general}} \overset{D}{\to} N(0, 1)\) under \(H_{0;T}\), we have \(1 - F_{0;T}(x; n) \to 1 - \Phi(x)\) for any \(x\). By [A.25],
\[
\lim_{n \to \infty} \frac{1 - \Phi(c_n)}{1 - \Phi(d_n)} = \lim_{n \to \infty} \frac{1 - F_{0;T}(c_n; n)}{1 - F_{0;T}(d_n; n)} = \lim_{n \to \infty} \frac{1 - F_{0;T}(c_n; n)}{1 - F_{0;T}(d_n; n)} = 1,
\]
which implies
\[
c_n - d_n = F_{0;T}^{-1}(1 - \zeta_{\alpha;n}^a; n) - \Phi^{-1}(1 - \zeta_{\alpha;n}^a) = o_n(1).
\] (A.27)

Then [15] (result (A.6)) together with (A.27) imply \(H(\zeta_{\alpha;n}^a; n) - H(\zeta_{\alpha;n}; n) = o_n(1)\). Since \(H'(t; n)\) is bounded below for \(t\) in an open interval with endpoints \(\zeta_{\alpha;n}\) and \(\zeta_{\alpha;n}^a\), \(\zeta_{\alpha;n}^a - \zeta_{\alpha;n} = o_n(1)\) holds.

We now show (A.24). By the definition of \(\tau_{\alpha;m;n}\), \(\widehat{\text{FDR}}(\tau_{\alpha;m;n}) = \alpha\), which yields
\[
\widehat{\text{FDR}}(\tau_{\alpha;m;n}) - \alpha = \widehat{\text{FDR}}(\tau_{\alpha;m;n}) - \widehat{\text{FDR}}(\zeta_{\alpha;n}^a) + \widehat{\text{FDR}}(\zeta_{\alpha;n}^a) - \widehat{\text{FDR}}(\tau_{\alpha;m;n}).
\] (A.28)

Utilizing [15] (results (A.10), (A.11) and (A.9)) yields
\[
\begin{align*}
\widehat{\text{FDR}}(\tau_{\alpha;m;n}) - \widehat{\text{FDR}}(\zeta_{\alpha;n}^a) &= O_P(m^{-1/2}), \\
\widehat{\text{FDR}}(\zeta_{\alpha;n}^a) - \widehat{\text{FDR}}(\zeta_{\alpha;n}) &= o_n(1) + O_P(m^{-1/2}), \\
\widehat{\text{FDR}}(\tau_{\alpha;m;n}) - \widehat{\text{FDR}}(\zeta_{\alpha;n}) &= O_P(m^{-1/2}),
\end{align*}
\] (A.29)

respectively, where the second equality also utilizes [A.23]. Substituting (A.29) into (A.28), we get (A.24).

Finally, an application of [24] (Theorem 6) shows that
\[
P(\text{FDR}(\tau_{\alpha;m;n}) \leq \widehat{\text{FDR}}(\tau_{\alpha;m;n})) \to 1.
\] (A.30)

By (A.30), together with (A.24), we obtain \(\text{FDR}(\tau_{\alpha;m;n}) \leq \alpha + o(1)\). This completes the proof. ■
Proof of Proposition 4. For the critical value $t_{\alpha;m}^a$ given in (3.2), we observe

\[ \text{FWER}_1(t_{\alpha;m}^a) = P(\bigcup_{i \in I_0} \{|T_{i;n_1,n_2}^{\text{general}}| > t_{\alpha;m}^a\}) \]

\[ \leq \sum_{i \in I_0} P(|T_{i;n_1,n_2}^{\text{general}}| > t_{\alpha;m}^a) \]

\[ = \sum_{i \in I_0} \alpha_i n_1 n_2 (t_{\alpha;m}^a) \]

\[ = \pi_0 \beta_{1;\alpha} + o(1), \]

where the last equality comes from (A.1). ■

Proof of Proposition 5. For the critical value $t_{\alpha;m;k}^a$ given in (3.4), an application of Markov inequality gives

\[ \text{FWER}_k(t_{\alpha;m;k}^a) \leq \frac{E\{V_m(t_{\alpha;m;k}^a)\}}{k} = \frac{\sum_{i \in I_0} P(|T_{i;n_1,n_2}^{\text{general}}| > t_{\alpha;m;k}^a)}{k} \]

\[ = \frac{\sum_{i \in I_0} \alpha_i n_1 n_2 (t_{\alpha;m;k}^a)}{k} \]

\[ = \pi_0 \beta_{k;\alpha}/k + o(1), \]

where the last equality is obtained from (A.2). ■

Derivation of (2.6). Under $H_{0,i}$ in (2.2), (2.5) becomes

\[ T_{\text{pool}}_{i;n_1,n_2} = \frac{\bar{z}_i - \bar{\varepsilon}_i}{s_{\text{pool}_{X,Y};i}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]

\[ = \frac{\bar{z}_i - \bar{\varepsilon}_i}{s_{\text{pool}_{X,Y};i}} \sqrt{\frac{\sigma_{\varepsilon,i}^2}{n_1} + \frac{\sigma_{\varepsilon,i}^2}{n_2}} \sqrt{\left(1 - \rho\right)\sigma_{\varepsilon,i}^2 + \rho \sigma_{\varepsilon,i}^2} \]

\[ \times (1 + o(1)) \]

\[ \text{(A.31)} \]

as $n_1 \to \infty$ and $n_2 \to \infty$, where $\bar{z}_i = \sum_{j=1}^{n_1} z_{i,j}/n_1$ and $\bar{\varepsilon}_i = \sum_{j=1}^{n_2} \varepsilon_{i,j}/n_2$.

(i) By CLT, \[ \frac{\bar{z}_i - \bar{\varepsilon}_i}{s_{\text{pool}_{X,Y};i}} \sqrt{\frac{\sigma_{\varepsilon,i}^2}{n_1} + \frac{\sigma_{\varepsilon,i}^2}{n_2}} \overset{D}{\to} N(0, 1). \]
(ii) By law of large numbers, \( s_{X;i}^2 \xrightarrow{P} \sigma_{\varepsilon;i}^2 \) and \( s_{Y;i}^2 \xrightarrow{P} \sigma_{\varepsilon;i}^2 \), and thus

\[
\frac{(n_1 - 1)s_{X;i}^2 + (n_2 - 1)s_{Y;i}^2}{(n_1 + n_2 - 2)} \xrightarrow{P} \rho \sigma_{\varepsilon;i}^2 + (1 - \rho)\sigma_{\varepsilon;i}^2.
\]

This combined with (A.31) and (2.7) implies that \( T_{i;n_1, n_2}^{\text{pool}} \xrightarrow{D} N(0, 1) \cdot \sqrt{\frac{(1 - \rho)\sigma_{\varepsilon;i}^2 + \rho \sigma_{\varepsilon;i}^2}{\rho \sigma_{\varepsilon;i}^2 + (1 - \rho)\sigma_{\varepsilon;i}^2}} = N(0, 1) \cdot \sigma_{\rho \theta(\varepsilon, \varepsilon;i)i} \). ■

**Appendix B: Extensions of Models (4.10) and (4.12)**

More generally, consider observations \( \{X_{i,j}\} \) and \( \{Y_{i,j}\} \) described by the following model:

\[
\begin{align*}
X_{i,j} &= \mu_{X,i} + \varepsilon_{i,j} + \gamma_{X,i}^T \mathbf{w}_i, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_1, \\
Y_{i,j} &= \mu_{Y,i} + e_{i,j} + \gamma_{Y,i}^T \mathbf{w}_i, \quad 1 \leq i \leq m, \ 1 \leq j \leq n_2,
\end{align*}
\]

(B.1)

where \( \mathbf{w}_i \) are unobserved \( d_w \)-dimensional random vectors, with \( \{\mathbf{w}_1, \ldots, \mathbf{w}_m\} \overset{i.i.d.}{\sim} N(\mathbf{0}, \Sigma_w) \); for each \( i \), errors \( \{\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}\} \overset{i.i.d.}{\sim} N(0, \sigma_{\varepsilon;i}^2) \), errors \( \{e_{i,1}, \ldots, e_{i,n_2}\} \overset{i.i.d.}{\sim} N(0, \sigma_{e;i}^2) \), and \( \{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}), (e_{i,1}, \ldots, e_{i,n_2}), \mathbf{w}_i\} \) are mutually independent; \( \{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_1}), e_{i,1}, \ldots, e_{i,n_2}; \mathbf{w}_i\) : \( i \in I_0 \) are independent. Clearly, the factor \( \mathbf{w}_i \) describes both the dependence between the \( X \)-group and \( Y \)-group, the dependence within the \( X \)-group, as well as the dependence within the \( Y \)-group, where the amount of the dependence is described by non-random parameters \( \gamma_{X,i} \) and \( \gamma_{Y,i} \). As seen from (B.2) and (B.3), test statistics (using either \( T_{i;n_1, n_2}^{\text{general}} \) or \( T_{i;n_1, n_2}^{\text{pool}} \)) or \( T_{i;n_1, n_2}^{\text{pool}} \) associated with true nulls continue to be independent.

**Case (i).** The case of \( \gamma_{X,i} = \gamma_{Y,i} \), which includes Model (4.10), indicates that the influence of common factors \( \mathbf{w}_i \) are identical between the \( X \)-group and \( Y \)-group. The conclusions on \( T_{i;n_1, n_2}^{\text{general}}, T_{i;n_1, n_2}^{\text{pool}} \) and \( T_{i;n_1, n_2}^{\text{pool}} \) are identical to those in Section 4.2.

**Case (ii).** The case of \( \gamma_{X,i} \neq \gamma_{Y,i} \), which includes Model (4.12), indicates that the common factors \( \mathbf{w}_i \) in the \( X \)-group and \( Y \)-group are different. The conclusions on \( T_{i;n_1, n_2}^{\text{general}}, T_{i;n_1, n_2}^{\text{pool}} \) and \( T_{i;n_1, n_2}^{\text{pool}} \) are identical to those in Section 4.3.
Detailed discussions on the performance of \(T_{i,n_1,n_2}^{\text{general}}, T_{i,n_1,n_2}^{\text{pool}}\) and \(T_{i,n_1,n_2}^{\text{pool};A}\) are given below.

**Case (i): \(\gamma_{X,i} = \gamma_{Y,i}\) in model (B.1).** This case means that the influence of common factors \(w_i\) are identical between the \(X\)- and \(Y\)-groups. It follows that two-sample \(t\)-statistics under \(H_{0,i}\) reduce to the following forms,

\[
T_{i,n_1,n_2}^{\text{general}} = \frac{\bar{z}_i - \bar{e}_i}{\sqrt{s_{z;i}^2/n_1 + s_{e;i}^2/n_2}}, \quad T_{i,n_1,n_2}^{\text{pool}} = \frac{\bar{z}_i - \bar{e}_i}{s_{\text{pool};i}^2 \sqrt{1/n_1 + 1/n_2}}, \quad T_{i,n_1,n_2}^{\text{pool};A} = \frac{T_{i,n_1,n_2}^{\text{pool}}}{\sigma_{\hat{\theta}_{(e,e)}(i)}}. \tag{B.2}
\]

It is interesting to note that Case (i) involves dependence between different groups, as well as within a same group, but test statistics (using either \(\{T_{i,n_1,n_2}^{\text{general}}\}\) or \(\{T_{i,n_1,n_2}^{\text{pool}}\}\) or \(T_{i,n_1,n_2}^{\text{pool};A}\)) associated with true nulls are independent.

Under this special case, we can show two distributional results below for the “general” two-sample \(t\)-statistic \(T_{i,n_1,n_2}^{\text{general}}\) under \(H_{0,i}\):

(c1’) if \(\sigma_{z;i}^2 = \sigma_{e;i}^2\) 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 \(T_{i,n_1,n_2}^{\text{general}} \xrightarrow{D} N(0,1).\)

Hence, conclusions of Propositions 1–3 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 make two conclusions below.

(d1’) If \(\sigma_{z;i}^2 = \sigma_{e;i}^2\), then \(T_{i,n_1,n_2}^{\text{pool}} \sim t_{n_1+n_2-2}.\) In this case, the results in Propositions 1–3 continue to apply for the “pooled” choice \(\{T_{i,n_1,n_2}^{\text{pool}}\}_{i=1}^m\).

(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 \(T_{i,n_1,n_2}^{\text{pool}} \xrightarrow{D} N(0, \sigma_{\hat{\theta}_{(e,e)}(i)}^2).\) Similar to the discussion in Section 3.2, there will be no guarantee in the case of \(\sigma_{\hat{\theta}_{(e,e)}(i)} > 1\) for achieving level bounds \(\alpha\) in (2.11) and (2.12) using \(\{T_{i,n_1,n_2}^{\text{pool}}\}_{i=1}^m\).
But according to (B.2), the "adaptively pooled" version satisfies $T_{i:n_1,n_2}^{\text{pool;A}} \xrightarrow{D} \mathcal{N}(0, 1)$, and thus the $\mathcal{N}(0, 1)$ calibration remains valid for $\{T_{i:n_1,n_2}^{\text{pool;A}}\}_{i=1}^m$.

**Case (ii): $\gamma_{X,i} \neq \gamma_{Y,i}$ in model (B.1).** This case means that the common factors $w_i$ in the $X$-group and $Y$-group are different. The explicit forms of two-sample $t$-statistics can be derived as follows,

$$T_{i:n_1,n_2}^{\text{general}} = \frac{\bar{e}_i - \bar{w}_i + (\gamma_{X,i} - \gamma_{Y,i})^T w_i}{\sqrt{\frac{s_{\epsilon,i}^2}{n_1} + \frac{s_{\epsilon,i}^2}{n_2}}}, \quad T_{i:n_1,n_2}^{\text{pool}} = \frac{\bar{e}_i - \bar{w}_i + (\gamma_{X,i} - \gamma_{Y,i})^T w_i}{s_{\text{pool},i} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad T_{i:n_1,n_2}^{\text{pool;A}} = \frac{T_{i:n_1,n_2}^{\text{pool}}}{\sigma_p \hat{\theta}(e,e,i)}.$$  

(B.3)

which differ from those in (B.2). Again, dependence between different groups, as well as within a same group, exist in the dataset, where the extent of dependence is captured by the magnitude of $(\gamma_{X,i} - \gamma_{Y,i})^T w_i \sim \mathcal{N}(0, (\gamma_{X,i} - \gamma_{Y,i})^T \Sigma_{\epsilon} (\gamma_{X,i} - \gamma_{Y,i}))$, but two-sample $t$-statistics associated with true nulls remain independent.

In the context of (B.3), we can show two results for the null distribution of the "general" two-sample $t$-statistic $T_{i:n_1,n_2}^{\text{general}}$:

(e1') if $\sigma_{\epsilon,i}^2 = \sigma_{\epsilon,i}^2 = \sigma_i^2$ and $n_1 = n_2$, then

$$T_{i:n_1,n_2}^{\text{general}} \sim t_{2n_1 - 2} \times f_1', \text{ where } f_1' = \sqrt{1 + \frac{n_1 (\gamma_{X,i} - \gamma_{Y,i})^T \Sigma_{\epsilon} (\gamma_{X,i} - \gamma_{Y,i})}{\sigma_i^2}};$$

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

$$T_{i:n_1,n_2}^{\text{general}} = Z \times f_2'(1 + o_p(1)) \xrightarrow{P} \infty, \text{ where } f_2' = \sqrt{1 + \frac{n_1 n_2 (\gamma_{X,i} - \gamma_{Y,i})^T \Sigma_{\epsilon} (\gamma_{X,i} - \gamma_{Y,i})}{n_2 \sigma_{\epsilon}^2 + n_1 \sigma_i^2}}, \quad (B.4)$$

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

We can also show that $T_{i:n_1,n_2}^{\text{pool;A}}$ has the same limit null distribution as $T_{i:n_1,n_2}^{\text{general}}$. For the null distribution of the "pooled" two-sample $t$-statistic $T_{i:n_1,n_2}^{\text{pool}}$, we make two conclusions below.
(f1') If $\sigma^2_{e;i} = \sigma^2_{i;i} = \sigma^2_i$, then

$$T_{i;n_1,n_2}^{pool} \sim t_{n_1+n_2-2} \times f'_3,$$

where $f'_3 = \sqrt{1 + \frac{n_1n_2}{n_1 + n_2} \frac{(\gamma_{X;i} - \gamma_{Y;i})^T \Sigma w (\gamma_{X;i} - \gamma_{Y;i})}{\sigma^2_i}}$.

(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

$$T_{i;n_1,n_2}^{pool} = Z \times f'_4 \{1 + o_P(1)\} \xrightarrow{P} \infty,$$

where $f'_4 = \sqrt{\frac{\sigma^2_{p,\theta(e,e)} + n_1n_2}{n_1 + n_2} \frac{(\gamma_{X;i} - \gamma_{Y;i})^T \Sigma w (\gamma_{X;i} - \gamma_{Y;i})}{\rho \sigma^2_e + (1 - \rho) \sigma^2_i}}$.

Thus, conclusions of Propositions 1–3 will fail for two-sample $t$-statistics \{T_{i;n_1,n_2}^{general}\}_{i=1}^m, since the factor $f'_2$ in (B.4) invariably exceeds one. As a comparison, Propositions 1–3 may fail for \{T_{i;n_1,n_2}^{pool}\}_{i=1}^m, particularly when the factor $f'_4$ in (B.5) substantially exceeds one. In the case of $f'_2 > f'_4$, the “adaptively pooled” versions \{T_{i;n_1,n_2}^{pool;A}\}_{i=1}^m will not ameliorate \{T_{i;n_1,n_2}^{pool}\}_{i=1}^m.

Appendix C: Figures and Tables in the Paper

Figure 1: Plots of $\{1-\Phi(x/\sigma)/[1-\Phi(x)]$ versus $x$. Left panel: $\sigma > 1$; right panel: $\sigma < 1$. 

\[\sigma > 1\]

\[\sigma < 1\]
Figure 2: Left panel: plot of $\alpha/\beta_{1,\alpha}$ versus $\alpha$. Right panel: compare plots of $\beta_{1,\alpha}$ and $\alpha$ versus $\alpha$.

Figure 3: (Empirical estimates of $\mathrm{FWER}(t_{\alpha,m}^*)$ (using ◦).) The horizontal dashed line indicates $\alpha$. Two-sample $t$-tests in columns (a)–(f) are $T_{\text{general}}^*$ in (2.3), $T_{\text{pool}}^*$ in (2.5), $T_{\text{pool},A}^*$ in (3.12), $T_{\text{adjust},T}^*$ in (3.13), $T_{\text{adjust},E}^*$ in (3.16), $T_{\text{2-stage}}^*$ in (3.19). Top row panels: for Example 1; middle row panels: for Example 1(I); bottom row panels: for Example 1(II).
Figure 4: The caption is similar to that of Figure 3, except for Example 2, Example 2(I), Example 2(II).

Figure 5: The caption is similar to that of Figure 3, except for Example 3, Example 3(I), Example 3(II).
Figure 6: The caption is similar to that of Figure 3, except for Example 4, Example 4(I), Example 4(II).

Figure 7: The caption is similar to that of Figure 3, except for Example 5, Example 5(I), Example 5(II).
Figure 8: (Empirical estimates of FWER_{k}(t^a_m;k) (using ◦) with k = 2.) The horizontal dashed line indicates α. Two-sample t-tests in columns (a)–(f) are $T^\text{general}_{i;\nu_1,\nu_2}$ in (2.3), $T^\text{pool}_{i;\nu_1,\nu_2}$ in (2.5), $T^\text{pool;A}_{i;\nu_1,\nu_2}$ in (3.12), $T^\text{adjust; T}_{i;\nu_1,\nu_2}$ in (3.13), $T^\text{adjust; E}_{i;\nu_1,\nu_2}$ in (3.16), $T^\text{2 stage}_{i;\nu_1,\nu_2}$ in (3.19). Top row panels: for Example 1; middle row panels: for Example 1(I); bottom row panels: for Example 1(II).

Figure 9: The caption is similar to that of Figure 8, except for Example 2, Example 2(I), Example 2(II).
Figure 10: The caption is similar to that of Figure 8, except for Example 3, Example 3(I), Example 3(II).

Figure 11: The caption is similar to that of Figure 8, except for Example 4, Example 4(I), Example 4(II).
Figure 12: The caption is similar to that of Figure 8 except for Example 5, Example 5(I), Example 5(II).
Figure 13: (Calculated FDP of the BH procedure.) The p-values are calculated via the $N(0, 1)$ (using $-$ for $T_{i:n_1,n_2}^{\text{general}}$), (exact $t_{n_1+n_2-2}$-distribution (using red $-$) for $T_{i:n_1,n_2}^{\text{pool}}$ in Example 1), $N(0, 1)$ (using $-$ for $T_{i:n_1,n_2}^{\text{pool,A}}$), $N(0, 1)$ (using $-$ for $T_{i:n_1,n_2}^{\text{adjust,T}}$, $N(0, 1)$ (using $-$ for $T_{i:n_1,n_2}^{\text{adjust,E}}$, $N(0, 1)$ (using $-$ for $T_{i:n_1,n_2}^{2,\text{stage}}$). The horizontal dashed line indicates $\alpha$. 

Statistica Sinica: Newly accepted Paper (accepted author-version subject to English editing)
Table 1: Quantities in simulations examples in (5.1).

<table>
<thead>
<tr>
<th>Example</th>
<th>$\sigma_{\rho(x,e)}^2$</th>
<th>$\mu_{3,X;i}/n_1^2 - \mu_{3,Y;i}/n_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1.25</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>$16/n_1^2 + 16/n_2^2$</td>
</tr>
<tr>
<td>5</td>
<td>1.9769</td>
<td>$32/n_1^2 - (2b_i - 1) \times 128/n_2^2$</td>
</tr>
</tbody>
</table>

Table 2: Number of genes called differentially expressed at $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>data</th>
<th>[12]</th>
<th>[17]</th>
<th>[5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m; n_1; n_2$</td>
<td>6033; 50; 52</td>
<td>8648; 27; 17</td>
<td>12625; 37; 42</td>
</tr>
<tr>
<td>$T_{\text{general}}^{i;n_1,n_2}$ via $N(0,1)$</td>
<td>51</td>
<td>565</td>
<td>214</td>
</tr>
<tr>
<td>$T_{\text{pool}}^{i;n_1,n_2}$ via $t_{n_1+n_2-2}$</td>
<td>21</td>
<td>196</td>
<td>169</td>
</tr>
<tr>
<td>$T_{\text{pool}}^{i;n_1,n_2}$ via $N(0,1)$</td>
<td>51</td>
<td>436</td>
<td>210</td>
</tr>
<tr>
<td>$T_{\text{pool}A}^{i;n_1,n_2}$ via $N(0,1)$</td>
<td>51</td>
<td>563</td>
<td>215</td>
</tr>
<tr>
<td>$T_{\text{2stage}}^{i;n_1,n_2}$ via $N(0,1)$</td>
<td>50</td>
<td>565</td>
<td>213</td>
</tr>
</tbody>
</table>