International Journal of Advanced Statistics and Probability, 1 (2) (2013) 9-24
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJASP

Tests of linear hypotheses in the ANOVA under heteroscedasticity

Jin-Ting Zhang*

Department of Statistics and Applied Probability, National University of Singapore
**Corresponding author E-mail: stazjt@nus.edu.sg*

Abstract

It is often of interest to undertake a general linear hypothesis testing (GLHT) problem in the one-way ANOVA without assuming the equality of the group variances. When the equality of the group variances is valid, it is well known that the GLHT problem can be solved by the classical F-test. The classical F-test, however, may lead to misleading conclusions when the variance homogeneity assumption is seriously violated since it does not take the group variance heteroscedasticity into account. To our knowledge, little work has been done for this heteroscedastic GLHT problem except for some special cases. In this paper, we propose a simple approximate Hotelling T^2 (AHT) test. We show that the AHT test is invariant under affine-transformations, different choices of the coefficient matrix used to define the same hypothesis, and different labeling schemes of the group means. Simulations and real data applications indicate that the AHT test is comparable with or outperforms some well-known approximate solutions proposed for the k-sample Behrens-Fisher problem which is a special case of the heteroscedastic GLHT problem.

Keywords: *Approximate Hotelling T^2 test, ANOVA under heteroscedasticity, Linear hypothesis test, k-sample Behrens-Fisher problem, Wishart-approximation.*

1 Introduction

Given k independent samples with the l-th sample being $x_{lj}, j = 1, 2, \cdots, n_l, \overset{i.i.d}{\sim} N(\mu_l, \sigma_l^2)$ for $l = 1, 2, \cdots, k$, where and throughout, $N(\mu, \sigma^2)$ denotes a univariate normal distribution with mean μ and variance σ^2, we want to test the following general linear hypothesis testing (GLHT) problem:

$$H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{c}, \quad \text{vs} \quad H_1 : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{c}, \tag{1}$$

without assuming the equality of the group variances $\sigma_l^2, l = 1, \cdots, k$, where $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_k]^T$, $\mathbf{C} : q \times k$ is a known coefficient matrix with $\text{rank}(\mathbf{C}) = q \leq k$, and $\mathbf{c} : q \times 1$ is a known constant vector. For convenience, the above problem may be referred to as the heteroscedastic GLHT problem. When the equality of the group variances $\sigma_l^2, l = 1, 2, \cdots, k$ is assumed, it is well known that the GLHT problem (1) can be solved by the classical F-test. For the heteroscedastic GLHT problem (1), however, the classical F-test may lead to misleading conclusions since it does not take the group variance heteroscedasticity into account. To our knowledge, little work has been done for this heteroscedastic GLHT problem except for some special cases. In this paper, we propose a simple approximate Hotelling T^2 (AHT) test. The AHT test uses the Wald-type test statistic with its null distribution approximated by a Hotelling T^2 distribution with parameters q and d. The parameter d can be easily estimated from the data using

a simple formula. Moreover, it is shown that the AHT test is invariant under affine-transformations, different choices of the matrix $\mathbf{C}$ for the same hypothesis, and different labeling schemes of the group means $\mu_1, \mu_2, \cdots, \mu_k$.

The heteroscedastic GLHT problem (1) is very general and it has wide applications. Many important special cases can be obtained via properly specifying the coefficient matrix $\mathbf{C}$ and the constant vector $\mathbf{c}$. For example, when we set $\mathbf{c} = c$, a scalar, and set $\mathbf{C} = \boldsymbol{\lambda}^T$, a $1 \times k$ row vector, the heteroscedastic GLHT problem (1) reduces to the problem of testing $H_0 : \boldsymbol{\lambda}^T\boldsymbol{\mu} = c$ versus $H_1 : \boldsymbol{\lambda}^T\boldsymbol{\mu} \neq c$. In this case, the AHT test is equivalent to an approximate t-test defined in a similar manner; see Section 2.2 for details. Using this approximate t-test, we can also construct approximate confidence intervals for $\boldsymbol{\lambda}^T\boldsymbol{\mu} - c$ if desired. Another important special case is obtained when we set $\mathbf{c} = \mathbf{0}$ and $\mathbf{C} = [\mathbf{I}_{k-1}, -\mathbf{1}_{k-1}]$ where and throughout $\mathbf{I}_a$ denotes the identity matrix of size a and $\mathbf{1}_a$ the column vector of ones of length a. In this case, testing the heteroscedastic GLHT problem (1) reduces to testing the following well-known heteroscedastic one-way ANOVA (ANalysis Of VAriance) problem:

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \quad H_0 \text{ is not true}, \tag{2}$$

without assuming the equality of the group variances $\sigma_l^2, l = 1, 2, \cdots, k$. This problem is also known as the k-sample Behrens-Fisher (BF) problem.

When $k = 2$, the problem (2) reduces to the well-known two-sample BF problem, which can be dated back to Behrens [2] and Fisher [8]. This two-sample BF problem has been well addressed by various authors including Welch [26], Aspin [1], Lee and Gurland [21], among others. In particular, the Welch test is the most popular and accurate solution ([15]). A good comprehensive review about the two-sample BF problem was given by Kim and Cohen [13]. In the recent decade, this two-sample BF problem continuously produced some interesting literature including Gupta and Wang [10], Ruben [24], and Chang and Pal [5] among others. In particular, Chang and Pal [5]'s parametric bootstrap (PB) test (also known as computational approach test) is very powerful and easy to implement. It can be easily extended to solving the k-sample BF problem ([15]) or comparing several population means for normal data or non-normal data ([6], [4]).

For the general k-sample BF problem (2), exact solutions are generally intractable but many approximate solutions are indeed available in the literature. Well-known approximate solutions include Welch's [28] approximate degrees of freedom (ADF) test, James' [12] second order series expansion solution, Brown and Forsythe's [3] modified F-test, and Krishnamoorthy, Lu and Mathew's [15] PB test, among others. These approximate solutions represent a number of similar approximate solutions proposed in the literature for the k-sample BF problem (2). Other well-known approximate solutions for the k-sample BF problem (2) include Welch [27], James [11], Krutchkoff [18], Wilcox ([29], [30]), Rice and Gaines [23], Weerahandi [25], Lee and Ahn [20], among others. It is not possible to give a complete list of this literature even for this k-sample BF problem, not mentioning that there is a vast literature about many approximate solutions to heteroscedastic two-way ANOVA, one-way MANOVA (Multivariate ANOVA) and two-way MANOVA problems. The interested reader is referred to the comprehensive review by Coombs *et al.* [7] and references therein. For more recent work in these areas, the reader is referred to Krishnamoorthy, Lu and Mathew [15], Chang and Pal [5], Krishnamoorthy and Lu [14] and references therein.

The AHT test proposed in this paper performs well in terms of size and power at least for the k-sample BF problem (2). Simulations presented in Section 3 in this paper show that for the k-sample BF problem (2) for normal data with various sample sizes and parametric configurations, the proposed AHT test is comparable with Krishnamoorthy, Lu and Mathew's [15] PB test and it generally outperforms Welch's [28] ADF test in terms of size and power. When k is large, the Welch test is very liberal, i.e., its empirical sizes are much larger than the given nominal size but this is not the case for the PB and AHT tests. Further simulation results presented in Zhang [32], an earlier version of this paper, show that for the k-sample BF problem (2) for normal and non-normal data with small samples and large k, the proposed AHT test are comparable with the PB test, but generally outperforms Welch's [28] ADF test, James' [12] second order series expansion solution, and Brown and Forsythe's [3] modified F-test.

Notice that although the PB test ([15]) is powerful and easy to implement as mentioned earlier, it requires substantially more computations than the AHT test. For a single case as in real data application, given the fast and affordable computational resources available nowadays, this may not be a concern. However, for mass computation as in simulation studies, the computation work is a big burden. In fact, in the simulation studies presented in Section 3, the time spent by the PB test is about 10000 times of the time spent by the AHT test. Therefore, we recommend the AHT test for the heteroscedastic GLHT problem (1) especially when a number of normal population means are involved and when a quick test result is desired.

The good performance of the AHT test is not totally a surprise. When $k = 2$, the AHT test coincides with Welch's [27] test which, as mentioned earlier, is known to be the most accurate approximate solution to the two-sample BF problem ([15]). As shown in Section 2, the AHT test is developed for the heteroscedastic GLHT problem (1) in a similar manner as the modified Nel and Van der Merwe's [22] test was developed for the two-sample multivariate BF problem by Krishnamoorthy and Yu [17] and they called the new test as the MNV test. Intensive simulations in the literature ([17], [31], [16], and among others) show that the MNV test performs quite well for various sample sizes and parameter configurations. It is then natural to expect that the AHT test will have a similar nice performance for the heteroscedastic GLHT problem (1).

The rest of the paper is organized as follows. In Section 2, the AHT test is developed. Simulation studies are presented in Section 3. Applications to a real data set are given in Section 4. Some concluding remarks are given in Section 5. Technical proofs of the main results are outlined in the Appendix.

2 Main results

2.1 The Wald-type test statistic

We now start to construct the test statistic for the heteroscedastic GLHT problem (1). Let $\hat{\mu}_l = \bar{x}_l = n_l^{-1}\sum_{j=1}^{n_l} x_{lj}$ and $\hat{\sigma}_l^2 = (n_l-1)^{-1}\sum_{j=1}^{n_l}(x_{lj}-\hat{\mu}_l)(x_{lj}-\hat{\mu}_l)^T$ be the usual unbiased sample mean and variance of the l-th sample for $l = 1, 2, \cdots, k$. Set $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_k]^T$. Then $\hat{\boldsymbol{\mu}} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}\left(\frac{\sigma_1^2}{n_1}, \frac{\sigma_2^2}{n_2}, \cdots, \frac{\sigma_k^2}{n_k}\right)$. Since $\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c} \sim N_q(\mathbf{C}\boldsymbol{\mu} - \mathbf{c}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$, we can then construct the following Wald-type test statistic:

$$T = (\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c})^T \left(\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^T\right)^{-1} (\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c}), \tag{3}$$

where $\hat{\boldsymbol{\Sigma}} = \text{diag}(\frac{\hat{\sigma}_1^2}{n_1}, \frac{\hat{\sigma}_2^2}{n_2}, \cdots, \frac{\hat{\sigma}_k^2}{n_k})$. Let $n_{\min} = \min_{l=1}^k n_l$ and $n_{\max} = \max_{l=1}^k n_l$. Let χ_q^2 denote a χ^2-distribution with q degrees of freedom and $F_{r,m}$ an F-distribution with r and m degrees of freedom.

Remark 2.1 *When the variance homogeneity is assumed and the sample variances $\hat{\sigma}_l^2$ are replaced by their pooled sample variance $\sum_{l=1}^k (n_l-1)\hat{\sigma}_l^2/(N-k)$ where $N = \sum_{l=1}^k n_l$ denotes the total sample size of the k samples, it is easy to show that $T/q \sim F_{q,N-k}$. However, when this homogeneity assumption is violated, the distribution of T is complicated and its closed-form expression is less tractable unless $n_{\min}$ is very large.*

Remark 2.2 *Assume that the sample sizes $n_1, n_2, \cdots, n_k$ tend to infinity proportionally. That is,*

$$n_l/n_{\min} \to r_l < \infty,\ l = 1, 2, \cdots, k, \quad \text{as } n_{\min} \to \infty. \tag{4}$$

Then it is standard to show that as $n_{\min} \to \infty$, T asymptotically follows χ_q^2. However, we can show that the convergence rate of T to χ_q^2 is of order $n_{\min}^{-1/2}$ which is rather slow when $n_{\min}$ is small or moderate. Thus, the resulting χ^2-test is hardly useful for the heteroscedastic GLHT problem (1).

2.2 The AHT test

To overcome the problem mentioned in Remark 2.2, we here propose the so-called AHT test based on T. The key idea of the AHT test is to approximate the null distribution of T by a Hotelling T^2 distribution with some proper parameters. For this end, following Krishnamoorthy and Yu [17], we re-express T as:

$$T = \mathbf{z}^T\mathbf{W}^{-1}\mathbf{z}, \tag{5}$$

where $\mathbf{z} = (\mathbf{C\Sigma C}^T)^{-1/2}(\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c})$ and $\mathbf{W} = (\mathbf{C\Sigma C}^T)^{-1/2}(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)(\mathbf{C\Sigma C}^T)^{-1/2}$. Notice that $\mathbf{z} \sim N_q(\boldsymbol{\mu}_z, \mathbf{I}_q)$, where $\boldsymbol{\mu}_z = (\mathbf{C\Sigma C}^T)^{-1/2}(\mathbf{C}\boldsymbol{\mu} - \mathbf{c})$. Thus, under the null hypothesis, $\mathbf{z} \sim N_q(\mathbf{0}, \mathbf{I}_q)$. Let $W_q(m, \mathbf{V})$ denote a Wishart distribution with m degrees of freedom and covariance matrix $\mathbf{V}$ and let $T^2_{q,d}$ denote the well-known Hotelling T^2 distribution with parameters q and d. If we can show that for some $d > q + 3$,

$$\mathbf{W} \sim W_q(d, \mathbf{I}_q/d) \quad \text{approximately}, \tag{6}$$

then under the null hypothesis and (5), we have

$$T \sim T^2_{q,d} \quad \text{approximately}. \tag{7}$$

Remark 2.3 *To make the AHT test work, it is generally required that the first two moments of T and $T^2_{q,d}$ be finite. The condition "$d > q + 3$" guarantees that the first two moments of $T^2_{q,d}$ are finite as seen from the moment expression (20) given in Section 2.4.*

Thus, the main task of the AHT test is to show that we can approximate the random matrix $\mathbf{W}$ by some Wishart random matrix with d degrees of freedom and covariance matrix $\mathbf{I}_q/d$. This task may be easily conducted if we can show that $\mathbf{W}$ is a Wishart mixture, i.e., a sum of several independent Wishart random matrices since a Wishart mixture may be well approximated by a single Wishart random matrix as shown by Nel and Van der Merwe [22].

We now show that $\mathbf{W}$ is a Wishart mixture. For this end, write $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k]$ and $\mathbf{H} = (\mathbf{C\Sigma C}^T)^{-1/2}\mathbf{C} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_k]$, where $\mathbf{c}_l$ and $\mathbf{h}_l$ are the l-th columns of $\mathbf{C}$ and $\mathbf{H}$ respectively, with $\mathbf{h}_l = (\mathbf{C\Sigma C}^T)^{-1/2}\mathbf{c}_l, l = 1, 2, \cdots, k$.

Proposition 2.4 *We have*

$$\mathbf{W} = \mathbf{H}\hat{\mathbf{\Sigma}}\mathbf{H}^T = \sum_{l=1}^{k}\mathbf{W}_l, \;\; \mathbf{W}_l = \frac{\hat{\sigma}_l^2}{n_l}\mathbf{h}_l\mathbf{h}_l^T \sim W_q(n_l - 1, \frac{\mathbf{\Omega}_l}{n_l - 1}), \tag{8}$$

where $\mathbf{W}_l, l = 1, 2, \cdots, k$ are independent with $E(\mathbf{W}_l) = \mathbf{\Omega}_l = \frac{\sigma_l^2}{n_l}\mathbf{h}_l\mathbf{h}_l^T$. Furthermore,

$$E(\mathbf{W}) = \sum_{l=1}^{k}\mathbf{\Omega}_l = \mathbf{I}_q, \;\; Etr(\mathbf{W} - E\mathbf{W})^2 = 2\sum_{l=1}^{k}(n_l - 1)^{-1}\delta_l^2, \tag{9}$$

where

$$\delta_l = tr(\mathbf{\Omega}_l) = \frac{\sigma_l^2}{n_l}\mathbf{c}_l^T(\mathbf{C\Sigma C}^T)^{-1}\mathbf{c}_l, l = 1, 2, \cdots, k. \tag{10}$$

By Proposition 2.4, $\mathbf{W}$ is a Wishart mixture and hence its distribution can be approximated by that of a single Wishart random matrix, say, $\mathbf{R} \sim W_q(d, \mathbf{\Omega})$ ([22]). Proposition 2.4 also gives the first moment

and the total variation of $\mathbf{W}$ which will be used to determine d and $\mathbf{\Omega}$ later. The total variation of a random matrix $\mathbf{X} = (x_{ij}) : m \times m$ is defined as $\text{Etr}(\mathbf{X} - \text{E}\mathbf{X})^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \text{var}(x_{ij})$, i.e., the sum of the variances of all the entries of $\mathbf{X}$.

To approximate the distribution of $\mathbf{W}$ by that of $\mathbf{R} \sim W_q(d, \mathbf{\Omega})$, Nel and Van der Merwe [22] determined d and $\mathbf{\Omega}$ via matching the first two moments of $\mathbf{W}$ and $\mathbf{R}$. They obtained a few different solutions to d, with the simplest one being the same as the one we shall obtain here. In a slightly different way from that used by Nel and Van der Merwe [22], we determine d and $\mathbf{\Omega}$ via matching the first moments and the total variations of $\mathbf{W}$ and $\mathbf{R}$. That is, we solve the following two equations for d and $\mathbf{\Omega}$:

$$\text{E}(\mathbf{W}) = \text{E}(\mathbf{R}) \text{ and } \text{Etr}(\mathbf{W} - \text{E}\mathbf{W})^2 = \text{Etr}(\mathbf{R} - \text{E}\mathbf{R})^2. \tag{11}$$

The solution is given in Proposition 2.5 below together with the lower and upper bounds of d.

Proposition 2.5 *The solution of (11) is given by* $\mathbf{\Omega} = \mathbf{I}_q/d$ *and*

$$d = \frac{q(q+1)/2}{\sum_{l=1}^{k}(n_l - 1)^{-1}\delta_l^2}. \tag{12}$$

Moreover, we have the following inequality:

$$\frac{q+1}{2}(n_{\min} - 1) \leq d \leq \frac{q+1}{2q}(N - k). \tag{13}$$

Remark 2.6 *Proposition 2.5 indicates that provided* $n_{\min} > 3 + 4/(q+1)$*, we always have* $d > q + 3$*, guaranteeing that the first two moments of* $T_{q,d}^2$ *are finite as required in Remark 2.3.*

Remark 2.7 *From (12) and (13), it is seen that when* $n_{\min}$ *becomes large,* d *generally becomes large; and when* $n_{\min} \to \infty$*, we have* $d \to \infty$ *so that* $T_{q,d}^2$ *weakly tends to* χ_q^2*, the limit distribution of* T *as pointed out in Remark 2.2.*

Remark 2.8 *When the assumption (4) is not satisfied, the ratio* $n_{\max}/n_{\min}$ *will tend to* ∞ *as* $n_{\min} \to \infty$ *so that the limit of* $n_{\min}\mathbf{\Sigma}$ *is not a full rank matrix and hence the limit of* $n_{\min}\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T$ *is not invertible. In this case, the test statistic* T *and the quantities* $\delta_l, l = 1, \cdots, k$ *are not well defined so that* T *may not have finite first two moments as required in Remark 2.3. In this case, the proposed AHT test may not perform well as demonstrated by some simulation results presented in Section 3.*

In real data application, the parameter d has to be estimated based on the data. A natural estimator of d is obtained via replacing $\delta_l, l = 1, 2, \cdots, k$ by their estimators:

$$\hat{\delta}_l = \frac{\hat{\sigma}_l^2}{n_l}\mathbf{c}_l^T(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)^{-1}\mathbf{c}_l,\ l = 1, 2, \cdots, k, \tag{14}$$

so that

$$\hat{d} = \frac{q(q+1)/2}{\sum_{l=1}^{k}(n_l - 1)^{-1}\hat{\delta}_l^2}. \tag{15}$$

Notice that $\sum_{l=1}^{k} \hat{\delta}_l = q$ so that the range of d given in (13) is also the range of $\hat{d}$.

Remark 2.9 *Under the assumption (4), it is standard to show that as* $n_{\min} \to \infty$*, we have* $\hat{d} \to d$*. In addition, we can show that* $E(T_{q,\hat{d}}^2) = E(T)[1 + O(n_{\min}^{-2})]$ *and* $Var(T_{q,\hat{d}}^2) = Var(T)[1 + O(n_{\min}^{-1})]$*. That is, the means of* T *and* $T_{q,\hat{d}}^2$ *are matched up to order* $n_{\min}^{-2}$ *while the variances of* T *and* $T_{q,\hat{d}}^2$ *are matched only up to order* $n_{\min}^{-1}$*. This indicates that when* $n_{\min}$ *is too small, the AHT test may not perform well as demonstrated by some simulation results presented in Section 3.*

In summary, the AHT test is based on approximating the distribution of the Wald-type test statistic T (5) by a Hotelling T^2 distribution $T^2_{q,\hat{d}}$. By the property of the Hotelling T^2 distribution, the AHT test can be conducted using the usual F-distribution since we can write

$$T \sim \frac{q\hat{d}}{\hat{d}-q+1} F_{q,\hat{d}-q+1} \quad \text{approximately.} \tag{16}$$

Provided $n_{\min} > 3+4/(q+1)$, Proposition 2.5 guarantees that $\hat{d}-q+1 > 4$ so that the first two moments of $T^2_{q,\hat{d}}$ are finite as seen from the moment expression (20). Based on (16), the critical value of the AHT test can be specified as $\frac{q\hat{d}}{\hat{d}-q+1} F_{q,\hat{d}-q+1}(1-\alpha)$ for the nominal significance level α. We reject the null hypothesis in (1) when this critical value is exceeded by the observed test statistic T. The AHT test can also be conducted via computing the p-value based on the approximate distribution specified in (16).

We conclude this subsection by three more remarks as mentioned briefly in the introduction.

Remark 2.10 *When $q = 1$, the matrix $\mathbf{C}$ reduces to a $1 \times k$ row vector, namely, $\boldsymbol{\lambda}^T$ and the constant vector $\mathbf{c}$ reduces to a scalar, namely, c. In this case, equivalently, one can conduct an approximate t-test for (1), i.e., $H_0 : \boldsymbol{\lambda}^T\boldsymbol{\mu} = c$ vs $H_1 : \boldsymbol{\lambda}^T\boldsymbol{\mu} \neq c$, using*

$$\frac{\boldsymbol{\lambda}^T\hat{\boldsymbol{\mu}} - c}{\sqrt{\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}}} \sim t_{\hat{d}} \quad \textit{approximately,}$$

where $\hat{d}$ is still computed using (15) except replacing $\mathbf{C}$ and $\mathbf{c}$ with $\boldsymbol{\lambda}^T$ and c respectively. It is not a surprise since the Hotelling T^2 test is a multivariate generalization of the usual univariate t-test. Alternatively, one can construct the $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\lambda}^T\boldsymbol{\mu} - c$ as

$$(\boldsymbol{\lambda}^T\hat{\boldsymbol{\mu}} - c) \pm t_{\hat{d}}(\alpha/2)\sqrt{\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}},$$

where $t_{\hat{d}}(\alpha/2)$ denotes the upper (50α)-th percentile of the t-distribution with $\hat{d}$ degrees of freedom.

Remark 2.11 *For the two-sample BF problem, the AHT test coincides with Welch's [27] approximate degrees of freedom (ADF) test but the two tests are different for $k \geq 3$. In fact, when $k = 2$, the k-sample BF problem (2) reduces to the well-known two-sample BF problem. In this case, by setting $\mathbf{C} = [1, -1]$ and $\mathbf{c} = 0$, the test statistic T (3) reduces to Welch's [27] test statistic and the parameter $\hat{d}$ (15) reduces to the ADF of Welch [27]. Therefore, the AHT test in this case reduces to Welch's [27] ADF test. However, for the general k-sample BF problem (2) with $k \geq 3$, the AHT test is different from Welch's [28] ADF test which was proposed for the k-sample BF problem (2) only.*

Remark 2.12 *The AHT test can be regarded as an extension of the MNV test of Krishnamoorthy and Yu [17] to the context of the heteroscedastic GLHT problem (1). From the previous parts of this section, it is seen that the AHT test was developed for the heteroscedastic GLHT problem (1) in a similar manner as the MNV test was developed for the two-sample multivariate BF problem. The development of the MNV test can be briefly described as follows. Nel and Van der Merwe [22] first investigated how to approximate a Wishart mixture by a single Wishart random matrix. Using this technique, they proposed an AHT test for the two-sample multivariate BF problem. Unfortunately, the AHT test they proposed is not affine-invariant, as pointed out by Krishnamoorthy and Yu [17]. The latter two authors then modified Nel and Van der Merwe's [22] AHT test via expressing the associated test statistic in the form (5), resulting in the so-called MNV test. Intensive simulations in the literature ([17], [31], [16], and among others) show that the MNV test performs quite well for various sample sizes and parameter configurations. It is then natural to expect that the AHT test will have a similar nice performance for the heteroscedastic GLHT problem (1).*

2.3 Invariance properties of the AHT test

The AHT test has several desirable invariance properties. First of all, the AHT test is affine-invariant. That is, it is invariant under the following affine-transformation:

$$\tilde{x}_{lj} = ax_{lj} + b,\ j = 1, 2, \cdots, n_l;\ l = 1, 2, \cdots, k, \tag{17}$$

where a and b are two constants but $a \neq 0$.

Proposition 2.13 *The AHT test is affine-invariant in the sense that both the test statistic T (3) and the estimated parameter $\hat{d}$ (15) are invariant under the affine-transformation (17).*

Notice that in the heteroscedastic GLHT problem (1), the matrix $\mathbf{C}$ can be a contrast matrix with rows being contrasts such that the row totals of the matrix equal to 0. It is well known that for the same hypothesis in (2), the contrast matrix $\mathbf{C}$ is not unique. For example, both $\mathbf{C} = \Big(\mathbf{I}_{k-1}, -\mathbf{1}_{k-1}\Big)$ and $\tilde{\mathbf{C}} = \Big(-\mathbf{1}_{k-1}, \mathbf{I}_{k-1}\Big)$ are the contrast matrices for the hypothesis in (2). It is known from Kshirsagar ([19], Ch. 5, Sec. 4) that for any two contrast matrices $\mathbf{C}$ and $\tilde{\mathbf{C}}$ for the same hypothesis, there is a nonsingular matrix $\mathbf{P}$ such that

$$\tilde{\mathbf{C}} = \mathbf{PC}. \tag{18}$$

The AHT test is invariant to the choice of the contrast matrix $\mathbf{C}$ used in (1) for the same hypothesis. That is, it is invariant under the transformation (18). More generally, we have the following result.

Proposition 2.14 *The AHT test is invariant when the coefficient matrix $\mathbf{C}$ and the constant vector $\mathbf{c}$ in (1) are replaced with*

$$\tilde{\mathbf{C}} = \mathbf{PC} \ \ \textit{and} \ \ \tilde{\mathbf{c}} = \mathbf{Pc}, \tag{19}$$

respectively where $\mathbf{P}$ is any nonsingular matrix.

Finally, we have the following result.

Proposition 2.15 *The AHT test is invariant under different labeling schemes of the group means $\mu_l, l = 1, 2, \cdots, k$.*

2.4 Minimum sample size determination

Let $[a]$ denote the integer part of a. When $X \sim F_{q,v}$, it is easy to show that X has $([v/2] - 1)$ finite moments:

$$\mathrm{E}(X^r) = \frac{v^r q(q+2)\cdots\{q+2(r-1)\}}{q^r(v-2)(v-4)\cdots(v-2r)},\ \ r = 1, 2, \cdots, [v/2] - 1. \tag{20}$$

This moment expression can be used to determine the minimum sample size required to guarantee that the AHT test is validly constructed. In fact, from (20), we see that the condition

$$n_{\min} > 3 + \frac{4(r-1)}{q+1}, \tag{21}$$

which is obtained via using the lower bound of d (and $\hat{d}$ as well) given in (13), guarantees that $T^2_{q,\hat{d}}$ has r finite moments for all possible situations.

Remark 2.16 *When $r = 2$, the condition (21) reduces to $n_{\min} > 3 + 4/(q+1)$, guaranteeing that $T^2_{q,\hat{d}}$ has finite first two moments for all possible situations so that the AHT test will be well defined. When this condition is not satisfied, i.e., $n_{\min} \leq 3 + 4/(q+1)$, the first two moments of $T^2_{q,\hat{d}}$ may not be finite. In this case, the AHT test may not work well when the underlying null distribution of T actually has finite first two moments.*

Remark 2.17 *For the heteroscedastic one-way ANOVA problem (2), $q = k - 1$ so that the sufficient condition given in Remark 2.16 can be further expressed as*

$$n_{\min} \geq \begin{cases} 6, & \text{when } k = 2, \\ 5, & \text{when } k = 3, 4, \\ 4, & \text{when } k \geq 5. \end{cases} \tag{22}$$

Therefore, "$n_{\min} \geq 6$" guarantees that the AHT test is well defined for any given k.

3 Simulation studies

In this section, intensive simulations are conducted to evaluate the empirical Type I error rates and powers of the AHT test, together with Welch's [28] ADF test and the PB test ([15]). We choose the Welch and PB tests as the competitors for the AHT test since the Welch test is the most popular test and the PB test is the most accurate test for heteroscedastic one-way ANOVA ([15]).

For a given sample size vector $\mathbf{n} = [n_1, n_2, \cdots, n_k]$, the mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_k]$ and the variance vector $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2]$, we first generate a sample mean vector $\bar{\mathbf{x}} = [\bar{x}_1, \cdots, \bar{x}_k]$ and a sample variance vector $\hat{\boldsymbol{\sigma}}^2 = [\hat{\sigma}_1^2, \cdots, \hat{\sigma}_k^2]$ by $\bar{x}_l \sim N(\mu_l, \sigma_l^2/n_l)$ and $\hat{\sigma}_l^2 \sim \frac{\sigma_l^2}{n_l - 1}\chi^2_{n_l - 1}, l = 1, 2, \cdots, k$, then apply the AHT, Welch and PB tests to the generated $\bar{\mathbf{x}}$ and $\hat{\boldsymbol{\sigma}}^2$ and record their p-values. Notice that in this section only, row vectors instead of column vectors are used for easy presentation of the simulation results. For the PB test, 10000 inner bootstrap replicates are conducted. This process is repeated $N = 10000$ times. The empirical Type I error rates of the tests are the proportions of rejecting the null hypothesis, i.e., when their p-values are less than the nominal significance level α. In all the simulations conducted, we use $\alpha = 5\%$ for simplicity.

For simplicity, the tuning parameters and sample sizes are specified exactly the same as those in Krishnamoorthy, Lu and Mathew [15]. This allows to compare our simulation results with theirs, especially

Table 1: Empirical Type I error rates for $k = 2, \sigma_1^2 = 1$.

n		$(3,3)$			$(5,5)$			$(8,8)$			$(4,8)$	
σ_2^2	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT
0.01	.0576	.0670	.0576	.0507	.0517	.0507	.0468	.0474	.0468	.0514	.0549	.0514
0.05	.0565	.0647	.0565	.0478	.0506	.0478	.0532	.0536	.0532	.0587	.0646	.0587
0.10	.0505	.0577	.0505	.0563	.0596	.0563	.0526	.0532	.0526	.0569	.0633	.0569
0.20	.0432	.0500	.0432	.0530	.0553	.0530	.0494	.0504	.0494	.0572	.0629	.0572
0.30	.0398	.0452	.0398	.0495	.0515	.0495	.0480	.0488	.0480	.0577	.0639	.0577
0.40	.0362	.0408	.0362	.0464	.0480	.0464	.0470	.0475	.0470	.0546	.0590	.0546
0.50	.0378	.0426	.0378	.0456	.0474	.0456	.0480	.0493	.0480	.0569	.0611	.0569
0.60	.0393	.0434	.0393	.0424	.0438	.0424	.0470	.0477	.0470	.0565	.0606	.0565
0.70	.0323	.0371	.0323	.0426	.0438	.0426	.0485	.0495	.0485	.0509	.0547	.0509
0.80	.0359	.0417	.0359	.0426	.0446	.0426	.0487	.0489	.0487	.0546	.0576	.0546
0.90	.0356	.0392	.0356	.0446	.0461	.0446	.0455	.0458	.0455	.0486	.0515	.0486
1.00	.0387	.0423	.0387	.0450	.0456	.0450	.0493	.0496	.0493	.0496	.0516	.0496
ARE	21	**17.8**	21	9	**8.2**	9	4.6	**3.8**	4.6	**9.6**	17.6	**9.6**

Table 2: Empirical Type I error rates for $k = 3, \sigma_1^2 = 1$.

n		$(5,5,5)$			$(10,10,10)$			$(4,6,20)$			$(2,3,2)$	
(σ_2^2,σ_3^2)	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT
$(1,1)$	.0471	.0467	.0427	.0525	.0525	.0521	.0560	.0565	.0521	.0380	.0309	.0096
$(1,0.5)$	.0471	.0459	.0418	.0476	.0478	.0474	.0613	.0627	.0566	.0347	.0293	.0096
$(1,0.1)$	.0538	.0527	.0488	.0450	.0448	.0444	.0647	.0620	.0557	.0507	.0419	.0135
$(0.5,0.5)$	.0460	.0448	.0413	.0506	.0511	.0504	.0562	.0569	.0529	.0400	.0323	.0110
$(0.5,0.7)$	.0449	.0441	.0402	.0451	.0451	.0448	.0602	.0610	.0562	.0408	.0334	.0077
$(0.1,0.1)$	.0498	.0477	.0453	.0506	.0494	.0502	.0631	.0587	.0569	.0596	.0520	.0159
$(0.1,0.9)$	.0551	.0536	.0489	.0495	.0494	.0487	.0593	.0552	.0554	.0600	.0461	.0134
$(0.5,0.9)$	.0444	.0437	.0393	.0522	.0519	.0519	.0527	.0538	.0501	.0377	.0309	.0096
$(0.3,0.9)$	.0476	.0461	.0422	.0486	.0490	.0481	.0593	.0579	.0556	.0430	.0341	.0096
$(0.3,0.6)$	.0492	.0475	.0442	.0513	.0520	.0507	.0618	.0604	.0584	.0452	.0359	.0106
$(0.1,0.3)$	.0517	.0491	.0458	.0518	.0512	.0508	.0615	.0581	.0581	.0521	.0419	.0124
$(0.05,0.05)$	.0495	.0464	.0443	.0497	.0484	.0488	.0615	.0552	.0546	.0671	.0609	.0195
ARE	**5.8**	7.4	12.6	**4**	4.2	**4**	19.6	16.4	**10.4**	**18.4**	26	76.2

Table 3: Empirical Type I error rates for $k = 6, \sigma_1^2 = 1, a = (\sigma_2^2, ..., \sigma_6^2)$.

n		(5_5)			(10_5)		$(3,3,4,5,6,6)$			$(4,8,12,24,30,40)$		
$a \times 10$	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT
(10_5)	.0576	.0438	.0404	.0502	.0477	.0487	.0708	.0429	.0374	.0705	.0586	.0655
$(1_2,5_3)$	.0668	.0474	.0435	.0542	0507	.0522	.0806	.0491	.0446	.0735	.0524	.0684
$(1:1:5)$	.0661	.0461	.0432	.0505	.0460	.0484	.0749	.0441	.0399	.0711	.0576	.0682
$(1,10_4)$	.0688	.0492	.0463	.0556	.0516	.0530	.0755	.0463	.0427	.0689	.0561	.0662
$(2,4_2,2,1)$	.0629	.0450	.0415	.0537	.0500	.0512	.0835	.0478	.0435	.0726	.0534	.0671
$(5_4,10)$	.065	.0463	.0430	.0540	.0496	.0517	.0895	.0550	.0494	.0700	.0540	.0664
$(3,9,4,7,1)$	.0674	.0498	.0464	.0571	.0531	.0547	.0902	.0554	.0505	.0734	.0539	.0682
$(.1_2,.6,1_2)$	.0677	.0485	.0466	.0560	.0503	.0531	.0845	.0537	.0532	.0725	.0508	.0684
ARE	30.6	**6**	12.2	7.8	**3.2**	4.6	62.4	**8.4**	11.6	43.2	**9.2**	34.6

when one wants to compare the AHT test with the generalized F-test considered in Krishnamoorthy, Lu and Mathew [15]. In addition, some shorthand notations are used in the sample size vector $\mathbf{n}$ and the variance vector $\boldsymbol{\sigma}^2$, e.g. (a_r) denotes a vector meaning "a repeats r times" and $(a:b:c)$ denotes "a sequence from a to c with increment of b units". For example, $(1_2:2_2:5_2)$ represents $(1,1,3,3,5,5)$. Tables 1-5 show the empirical Type I error rates for various sample sizes, ranging from very small to moderate with $k = 2, 3, 6, 10$ and 20. In each table, the last row lists the values of ARE associated with

Table 4: Empirical Type I error rates for $k = 10, \sigma_1^2 = 1, a = (\sigma_2^2, \cdots, \sigma_{10}^2)$.

n		(5_{10})			(15_{10})		$(3_3,4_3,5_4)$			$(4_3,12_3,15_4)$		
$a \times 10$	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT	Welch	PB	AHT
(10_9)	.0801	.0447	.0444	.0517	.0476	.0511	.1155	.0409	.0349	.0921	.0609	.0793
$(1:1:9)$	.0892	.0471	.0481	.0506	.0473	.0498	.1080	.0382	.0331	.0868	.0561	.0754
$(1_2:1_2:4_2,5)$	.0860	.0471	.0467	.0537	.0496	.0532	.1076	.0402	.0365	.0839	.0534	.0717
$(1_5,2_4)$	.0836	.0449	.0450	.0544	.0497	.0536	.1154	.0439	.0403	.0922	.0588	.0793
$((1,10)_4,1)$	.0883	.0484	.0501	.0565	.0519	.0558	.1317	.0508	.0485	.0939	.0584	.0802
$(3_3,6_3,9_3)$	.0836	.0431	.0427	.0530	.0494	.0519	.1043	.0374	.0329	.0893	.0576	.0760
(1_9)	.0782	.0432	.0422	.0531	.0490	.0525	.1210	.0456	.0410	.0908	.0552	.0769
ARE	63.2	9	**8.8**	6.6	**2.6**	5.2	129.6	**15.6**	23.6	79.8	**14.4**	54

Table 5: Empirical Type I error rates for $k = 20, \sigma_1^2 = 1$.

$\mathbf{n}$		(5_{20})	
$(\sigma_2^2, \cdots \sigma_{20}^2)$	Welch	PB	AHT
(1_{19})	.1334	.0437	.0524
$(.1_2 : .1_2 : .9_2, 1)$	.1262	.0448	.0521
$((.1 : .1 : .5)_3, .1 : .1 : .4)$	.1263	.0477	.0554
$(.1_{19})$	.1252	.0443	.0521
$((.2_4 : .2_4 : .8_4), 1_3)$	.1253	.0426	.0504
$((.9 : .1 : .1)_2, 1)$	.1356	.0472	.0558
$(.01_3, .05_3, .1_3, .5_3, .6_3, .8_4)$	.1402	.0516	.0637
ARE	60.6	**9**	9.2

Table 6: Powers of the tests for $k = 3, \sigma_1^2 = 1$ and $\mu_1 = 0$.

		(μ_2, μ_3)						
(σ_2^2, σ_3^2)	Tests	$(0, 0)$	$(0, 0.2)$	$(0, 0.5)$	$(0, 0.7)$	$(0.5, 1)$	$(0, 1)$	$(1.5, 1)$
$\mathbf{n} = (10, 10, 10)$								
	Welch	.0510	.0729	.2184	.3722	.4582	.6755	.9333
$(0.3, 0.9)$	PB	.0511	.0727	.2171	.3708	.4572	.6759	.9326
	AHT	.0509	.0718	.2170	.3704	.4562	.6736	.9324
	Welch	.0498	.1011	.3510	.6331	.5877	.9121	.9752
$(0.1, 0.5)$	PB	.0491	.1012	.3509	.6314	.5851	.9108	.9745
	AHT	.0492	.0998	.3486	.6305	.5841	.9099	.9747
$\mathbf{n}$=(10,5,15)								
	Welch	.0454	.0676	.2248	.4224	.5076	.7376	.8633
$(0.3, 0.9)$	PB	.0466	.0685	.2279	.4251	.5091	.7396	.8654
	AHT	.0451	.0671	.2221	.4182	.5032	.7344	.8594
	Welch	.0444	.0948	.4220	.7198	.6821	.9619	.9556
$(0.1, 0.5)$	PB	.0446	.0956	.4237	.7222	.6821	.9626	.9561
	AHT	.0440	.0938	.4201	.7174	.6805	.9613	.9550

the Welch, PB or AHT tests. The quantity ARE is referred to as the average relative error (in percentage) and is defined by $\text{ARE} = 100M^{-1} \sum_{j=1}^{M} |\hat{\alpha}_j - \alpha|/\alpha$ where $\hat{\alpha}_j$ denotes the j-th Type I error rate and M

Table 7: Powers of the tests for $k = 10, (\mu_1, \cdots, \mu_8) = 0$.

		(μ_9, μ_{10})						
$\boldsymbol{\sigma}^2$	Tests	$(0, 0)$	$(0, 0.2)$	$(0, 0.5)$	$(0, 0.7)$	$(0.5, 1)$	$(0, 1)$	$(1.5, 1)$
$\mathbf{n}$=$(15_3, 20_3, 25_4)$								
	Welch	.0496	.0832	.3272	.6217	.9794	.9387	1
$(1, .1 : .1 : .9)$	PB	.0466	.0794	.3183	.6113	.9779	.9354	1
	AHT	.0497	.0833	.3278	.6223	.9794	.9393	1
	Welch	.0527	.0954	.4285	.7555	.9968	.9821	1
$(1, .1_3, .3_3, .7_3)$	PB	.0508	.0922	.4197	.7472	.9965	.9810	1
	AHT	.0530	.0956	.4291	.7558	.9968	.9823	1
$\mathbf{n}$=$(15 : 2 : 33)$								
	Welch	.0466	.0982	.4395	.7800	.9973	.9858	1
$(1, .1 : .1 : .9)$	PB	.0450	.0933	.4329	.7733	.9971	.9846	1
	AHT	.0472	.0986	.4409	.7808	.9974	.9858	1
	Welch	.0504	.1104	.5724	.8871	.9997	.9981	1
$(1, .1_3, .3_3, .7_3)$	PB	.0489	.1072	.5658	.8837	.9997	.9980	1
	AHT	.0510	.1111	.5738	.8877	.9997	.9981	1

is the number of the empirical Type I error rates in the associated column. The smallest ARE value (in boldface) indicates the best overall performance of the associated test among the three tests in terms of maintaining the nominal significance level. Tables 6 and 7 give the empirical powers of the three tests with $k = 3$ and 10.

First of all, from Tables 1-5, it is seen that in terms of ARE, the PB test generally outperforms the other two tests. That is, it is the best among the three tests in the sense of maintaining the nominal significance level most closely. This conclusion is consistent with the one drawn by Krishnamoorthy, Lu and Mathew [15]. However, carefully comparing the empirical Type I error rates of the AHT test and the PB test for various cases allows us to conclude that the AHT test is generally comparable to the PB test in the sense that for most cases, their empirical Type I error rates are about the same, and in some cases, e.g., the last case (associated with $\mathbf{n}$) of Table 1 and the central two cases of Table 2, the AHT test slightly outperforms the PB test while in some other cases, e.g., the first case of Table 1 and the last cases of Tables 2-4, the PB test outperforms the AHT test. In these latter cases, either $n_{\min}$ is too small, e.g., $n_{\min} < 3 + 4/k$, or the ratio $n_{\max}/n_{\min}$ is too large, or both, so that the AHT test may not perform well as pointed out in Remarks 2.8, 2.9, 2.16, and 2.17.

From Tables 1-5, it is seen that the AHT test generally outperforms the Welch test. When $k = 2$ (see Table 1), the two tests produce identical empirical Type I error rates as pointed out in Remark 2.11 and when $k = 3$ (see Table 2), the Welch test slightly outperforms the AHT test in the first case, and substantially outperforms the AHT test in the last case in which $n_{\min} = 2$ which is much smaller than 5 required by the AHT test as given Remark 2.17. However, in other two cases, the AHT test performs similarly or outperforms the Welch test. When $k = 6, 10, 20$ (see Tables 3-5), the AHT test outperforms the Welch test substantially in various cases. Based on the above results, it is expected that when k is large, the AHT test will generally outperform the Welch test.

Tables 6 and 7 show the empirical powers of the three tests for $k = 3$ and 10. It is seen that with large sample sizes, all the tests control the Type I error rates satisfactorily, and exhibit similar power properties for various sample sizes and parameter configurations although when $k = 3$ (resp. when $k = 10$), the Welch test has slightly higher (resp. lower) powers than the AHT test.

In summary, the AHT test generally outperforms the Welch test and is comparable to the PB test most of time. One may also conclude that the AHT test is also comparable with the James second order test when one compares the empirical Type I error rates of the AHT test with those of the James second order test presented in Krishnamoorthy, Lu and Mathew [15]; see also Zhang [32] for some simulation results about comparisons of the AHT test with several well-known approximate tests, including the James second order test, for heteroscedastic one-way ANOVA. Although the PB test generally controls the Type I error rates better than the AHT test, it requires substantially more computational efforts than the AHT test. For a single case as in real data application, given the fast and affordable computational resources available nowadays, this may not be a concern. However, for mass computation as in simulation studies, this is a big burden. In fact, the time spent by the PB test in the simulation studies presented in this paper is about 10000 times of the time spent by the AHT test. Therefore, we generally prefer the AHT test especially when many normal population means are compared and when a quick test result is desired.

4 The PTSD data

The study by Foa, Rothbaum, Riggs, and Murdock [9] involved 45 subjects (rape victims) who were randomly assigned to one of four groups treated by four different treatments: (1) Stress Inoculation Therapy (SIT) in which subjects were taught a variety of coping skills; (2) Prolonged Exposure (PE) in which subjects went over the rape in their mind repeatedly for seven sessions; (3) Supportive Counseling (SC) which was a standard therapy control group; and (4) a Waiting List (WL) control. In the actual study, pre- and post-treatment measures were taken on a number of variables. Here, however, we only look at the post traumatic stress disorder (PTSD) data (the total number of symptoms endorsed by the subject). The data

are available at `http://www.uvm.edu/~dhowell/StatPages/FoaFolder/Foa_Anova.html` where David C. Howell presented a standard ANOVA analysis, assuming homogeneity of the group variances. However, the sample group variances of the PTSD data are quite different, ranging from 15.61 to 123.6, as seen from Table 8. In fact, Chang, Pal, Lim and Lin [6] analyzed this dataset using a heteroscedastic t-model. They showed that for this dataset, the normality assumption is satisfied but the group variance homogeneity assumption is violated.

Table 8: *Summary statistics for the PTSD Data.*

Group	Size	Group Mean	Group Variance
1 (SIT)	14	11.07	15.61
2 (PE)	10	15.40	123.60
3 (SC)	11	18.09	50.89
4 (WL)	10	19.50	50.50

We made use of this PTSD data only to illustrate the AHT test and to compare it with the Welch and PB tests, and the standard ANOVA as well. This is not a formal analysis of the PTSD data. For this purpose, we considered all the four and three-group mean comparisons (first 5 cases) and a contrast and a non-contrast linear hypothesis tests (last 2 cases) as listed in Column 1 of Table 9. For the first 5 cases, the tests were conducted using only the related 3 or 4 samples so that the Welch and PB tests can be applied properly. For the last two cases, all the 4 samples were involved and the results of the Welch and PB tests are not directly available unless some further development for these two tests is done. From Table 8, it is seen that the homogeneity assumption is violated in various degrees in these cases with Case (2) most seriously and Case (4) most lightly. Overall speaking, the degrees of heteroscedasticity for Cases (1)-(3) are much higher than those for Cases (4)-(5). For the first 5 cases, the test results of the standard ANOVA, Welch, PB, and AHT tests are presented in Columns 2-5 of Table 9 respectively. For good accuracy, the p-values of the PB test were obtained based on 100000 inner bootstrap replicates.

We first compare the results by the AHT test with those by the Welch and PB tests. Since the group sizes are relatively large and are close to each other, it is seen that the p-values of the AHT test are nearly the same as those of the Welch and PB tests for the first five cases.

Secondly, we compare the results by the AHT test with those by the standard ANOVA. This allows us to check the impact of the heteroscedasticity on the test results. We did this via looking at the p-value discrepancies between the standard ANOVA and the AHT test for all the seven cases. It is seen that for Cases (4), (5) and (7) where the degrees of heteroscedasticity are not high, the p-value discrepancies between the standard ANONA and the AHT test are not large enough to yield inconsistent conclusions. For Cases (1)-(3) and (6), where the degrees of heteroscedasticity is high, however, the p-values of the standard ANOVA are about 2 to 5 times larger than those of the AHT test. Depending on the specified nominal significance level α, the conclusions made by the standard ANOVA may be opposite to those made by the AHT test. For example, when $\alpha = 1\%$, the standard ANOVA accepts (while the AHT test rejects) the null hypotheses of Cases (1), (3) and (6); when $\alpha = 5\%$, the standard ANOVA accepts (while the AHT test rejects) the null hypothesis of Case (2). Since the AHT test assumes much weaker conditions

Table 9: Group mean comparisons and tests of linear hypotheses for the PSTD data.

	p-values			
Case (Null Hypothesis)	ANOVA	Welch	PB	AHT
(1) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$	.0394	.0075	.0080	.0074
(2) $H_0 : \mu_1 = \mu_2 = \mu_3$	.0785	.0295	.0299	.0298
(3) $H_0 : \mu_1 = \mu_2 = \mu_4$	.0371	.0136	.0137	.0136
(4) $H_0 : \mu_1 = \mu_3 = \mu_4$	.0031	.0033	.0034	.0032
(5) $H_0 : \mu_2 = \mu_3 = \mu_4$	.5629	.6336	.6329	.6372
(6) $H_0 : 3\mu_1 - \mu_2 - 2\mu_3 = 0$	.0241	-	-	.0076
(7) $H_0 : \mu_1 - \mu_2 - 3\mu_4 = 0$	.0000	-	-	.0000

than the standard ANOVA does, the conclusions made by the AHT test will be more reliable than those made by the standard ANOVA. Thus, when the degree of heteroscedasticity is high, application of the standard ANOVA may yield misleading conclusions and the Welch, PB or AHT test should be used.

5 Concluding remarks

In this paper, we proposed the AHT test for the heteroscedastic GLHT problem (1) which includes the well-known k-sample BF problem (2) as a special case. Simulation studies and real data applications indicate that the AHT test is comparable with Krishnamoorthy, Lu and Mathew's [15] parametric bootstrap test and generally outperforms Welch's [28] approximate degrees of freedom test to the k-sample BF problem (2) with small samples and large k. We also discussed the effect of the sample sizes to the AHT test and identified the minimum sample sizes which guarantee that the AHT test will work. These two issues are often overlooked in the development of approximate solutions to the BF problems by other authors.

MATLAB codes for carrying out the computations in the paper and some further simulation results comparing the AHT test against some other well-known tests can be requested from the author (*stazjt@nus.edu.sg*).

Acknowledgements

The work was supported by the National University of Singapore academic research grant R-155-000-108-112.

Appendix: Technical Proofs

Proof of Proposition 2.4 First of all, notice that when $\mathbf{U} \sim W_p(m, \mathbf{V})$, by the property of Wishart random matrices and a lemma of Johansen (1980, p. 89), we have

$$\mathrm{E}(\mathbf{U}) = m\mathbf{V} \text{ and } \mathrm{Etr}(\mathbf{U} - \mathrm{E}\mathbf{U})^2 = m[\mathrm{tr}(\mathbf{V}^2) + \mathrm{tr}^2(\mathbf{V})]. \tag{A.1}$$

Let $X \overset{d}{=} Y$ denote "X and Y have the same distribution". Since $\hat{\sigma}_l^2 \sim \frac{\chi^2_{n_l-1}}{n_l-1} \overset{d}{=} W_1(n_l - 1, \frac{\sigma_l^2}{n_l-1})$, we have $\mathbf{W}_l = \frac{\hat{\sigma}_l^2}{n_l}\mathbf{h}_l\mathbf{h}_l^T \sim W_q(n_l - 1, \frac{\mathbf{\Omega}_l}{n_l-1})$ where $\mathbf{\Omega}_l = \frac{\sigma_l^2}{n_l}\mathbf{h}_l\mathbf{h}_l^T$. Applying (A.1) to $\mathbf{W}_l$, we have $\mathrm{E}(\mathbf{W}_l) = \mathbf{\Omega}_l$ and $\mathrm{Etr}(\mathbf{W}_l - \mathrm{E}\mathbf{W}_l)^2 = (n_l - 1)^{-1}[\mathrm{tr}(\mathbf{\Omega}_l^2) + \mathrm{tr}^2(\mathbf{\Omega}_l)] = 2(n_l - 1)^{-1}\delta_l^2$, where we use the fact that $\mathrm{tr}(\mathbf{\Omega}_l^2) = \mathrm{tr}^2(\mathbf{\Omega}_l) = \left[\frac{\sigma_l^2}{n_l}\mathbf{h}_l^T\mathbf{h}_l\right]^2 = \delta_l^2$. Since $\mathbf{W}_l$ are independent, we have $\mathrm{E}(\mathbf{W}) = \sum_{l=1}^k \mathbf{\Omega}_l = \sum_{l=1}^k \frac{\sigma_l^2}{n_l}\mathbf{h}_l\mathbf{h}_l^T = \mathbf{H\Sigma H} = \mathbf{I}_q$ and $\mathrm{Etr}(\mathbf{W} - \mathrm{E}\mathbf{W})^2 = \sum_{l=1}^k \mathrm{Etr}(\mathbf{W}_l - \mathrm{E}\mathbf{W}_l)^2 = 2\sum_{l=1}^k (n_l - 1)^{-1}\delta_l^2$. The proposition is proved.

Proof of Proposition 2.5 By Proposition 2.4, we have $\mathrm{E}(\mathbf{W}) = \mathbf{I}_q$ and $\mathrm{Etr}(\mathbf{W} - \mathrm{E}\mathbf{W})^2 = 2\sum_{l=1}^k \frac{\delta_l^2}{n_l-1}$. Applying (A.1) to $\mathbf{R}$, we have $\mathrm{E}(\mathbf{R}) = d\mathbf{\Omega}$ and $\mathrm{Etr}(\mathbf{R} - \mathrm{E}\mathbf{R})^2 = d[\mathrm{tr}(\mathbf{\Omega}^2) + \mathrm{tr}^2(\mathbf{\Omega})]$. Equating $\mathrm{E}(\mathbf{R})$ and $\mathrm{E}(\mathbf{W})$ leads to $\mathbf{\Omega} = \mathbf{I}_q/d$. We then have $\mathrm{Etr}(\mathbf{R} - \mathrm{E}\mathbf{R})^2 = q(q + 1)/d$. Equating $\mathrm{Etr}(\mathbf{W} - \mathrm{E}\mathbf{W})^2$ and $\mathrm{Etr}(\mathbf{R} - \mathrm{E}\mathbf{R})^2$ then leads to the expression in (12).

By (12), finding the lower and upper bounds of d is equivalent to finding the upper and lower bounds of the function $g(\delta_1, \delta_2, \cdots, \delta_k) = \sum_{l=1}^k (n_l - 1)^{-1}\delta_l^2$ where $\delta_l, l = 1, 2, \cdots, k$ are defined as in (10). By Proposition 2.4, we have $\sum_{l=1}^k \delta_l = \sum_{l=1}^k \mathrm{tr}(\mathbf{\Omega}_l) = \mathrm{tr}(\mathbf{I}_q) = q$. Then $\delta_k = q - \sum_{l=1}^{k-1} \delta_l$. Taking the partial derivatives of g with respect to $\delta_l, l = 1, 2, \cdots, k - 1$ and setting them to 0 leads to the following normal equation system:

$$\frac{\partial g}{\partial \delta_l} = \frac{2\delta_l}{n_l - 1} - \frac{2(q - \sum_{r=1}^{k-1} \delta_r)}{n_k - 1} = 0,\ l = 1, 2, \cdots, k - 1.$$

Solving the above equation system with respect to $\delta_l, l = 1, 2, \cdots, k-1$ leads to

$$\delta_l = q\frac{n_l - 1}{N - k}, l = 1, 2, \cdots, k, \tag{A.2}$$

where $N = \sum_{l=1}^{k} n_l$ as defined before. Since

$$\frac{\partial^2 g}{\partial\delta_l\partial\delta_r} = \begin{cases} 2(n_l-1)^{-1} + 2(n_k-1)^{-1}, & \text{if } r = l < k, \\ 2(n_k-1)^{-1}, & \text{if } r \neq l, r, l < k, \end{cases}$$

the associated Hessian matrix is positive definite. Thus, the function $g(\delta_1, \cdots, \delta_k)$ has minimum value $\frac{q^2}{N-k}$ when $\delta_l, l = 1, 2, \cdots, k$ take the values in (A.2). It follows that the upper bound of d is $\frac{q+1}{2q}(N-k)$.

We now find the lower bound of d. Notice that for any real matrix $\mathbf{A}$, the matrices $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same nonzero eigenvalues. This implies that $\mathbf{\Omega}_l = \frac{\sigma_l^2}{n_l}(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T)^{-1/2}\mathbf{c}_l\mathbf{c}_l^T(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T)^{-1/2}$ has only one nonzero eigenvalue $\delta_l = \frac{\sigma_l^2}{n_l}\mathbf{c}_l^T(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T)^{-1}\mathbf{c}_l$. Since $\sum_{l=1}^{k}\mathbf{\Omega}_l = \mathbf{I}_q$, we have $\mathbf{I}_q - \mathbf{\Omega}_l = \sum_{r=1,r\neq l}^{k}\mathbf{\Omega}_r$. This implies that $\mathbf{I}_q - \mathbf{\Omega}_l$ is nonnegative. Using the singular value decomposition of $\mathbf{\Omega}_l$, it is easy to show that the nonzero eigenvalue δ_l of $\mathbf{\Omega}_l$ is less than 1. We then have $\sum_{l=1}^{k}(n_l-1)^{-1}\delta_l^2 \leq (n_{\min}-1)^{-1}\sum_{l=1}^{k}\delta_l = (n_{\min}-1)^{-1}q$. It follows that $d \geq (n_{\min}-1)\frac{q+1}{2}$. The proposition is proved.

Proof of Proposition 2.13 Recall that μ_l and σ_l^2 are the mean and variance of the responses x_{lj} and their unbiased estimators are $\hat{\mu}$ and $\hat{\sigma}^2$ as defined previously. We now denote the mean and variance of the affine-transformed responses $\tilde{x}_{lj}$ by $\tilde{\mu}_l$ and $\tilde{\sigma}_l^2$ respectively with their unbiased estimators denoted by $\widehat{\tilde{\mu}}_l$ and $\widehat{\tilde{\sigma}}_l^2$ respectively.

First of all, under the affine-transformation (17), we have $\tilde{\mu}_l = a\mu_l + b$ and $\tilde{\sigma}_l^2 = a^2\sigma_l^2,\ \ l = 1, 2, \cdots, k$. It follows that $\mu_l = (\tilde{\mu}_l - b)/a$. As we defined the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{\Sigma}$ in Sections 1 and 2, we now define $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{\Sigma}}$ similarly. Then we have $\boldsymbol{\mu} = (\tilde{\boldsymbol{\mu}} - b\mathbf{1}_k)/a$ and $\tilde{\mathbf{\Sigma}} = a^2\mathbf{\Sigma}$. It follows that the GLHT problem (1) can be equivalently expressed as $\tilde{H}_0 : \tilde{\mathbf{C}}\tilde{\boldsymbol{\mu}} = \tilde{\mathbf{c}}$ versus $\tilde{H}_1 : \tilde{\mathbf{C}}\tilde{\boldsymbol{\mu}} \neq \tilde{\mathbf{c}}$, where $\tilde{\mathbf{C}} = \mathbf{C}/a$ and $\tilde{\mathbf{c}} = b\mathbf{C}/a + \mathbf{c}$.

Similarly, under the affine-transformation (17), we have $\widehat{\tilde{\mu}}_l = a\hat{\mu}_l + b$, and $\widehat{\tilde{\sigma}}_l^2 = a^2\hat{\sigma}_l^2$. It follows that $\widehat{\tilde{\boldsymbol{\mu}}} = a\hat{\boldsymbol{\mu}} + b\mathbf{1}_k$ and $\widehat{\tilde{\mathbf{\Sigma}}} = a^2\hat{\mathbf{\Sigma}}$. Therefore, $\tilde{\mathbf{C}}\widehat{\tilde{\boldsymbol{\mu}}} - \tilde{\mathbf{c}} = \mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c}$ and $\tilde{\mathbf{C}}\widehat{\tilde{\mathbf{\Sigma}}}\tilde{\mathbf{C}}^T = \mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T$. That is, both $\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c}$ and $\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T$ are affine-invariant. It follows that the test statistic T (3) is affine-invariant.

To show that $\hat{d}$ is affine-invariant, by (15), it is sufficient to show that $\hat{\delta}_l^2$ are affine-invariant. Notice that $\tilde{\mathbf{C}} = \mathbf{C}/a$ implies $\tilde{\mathbf{c}}_l = \mathbf{c}_l/a,\ \ l = 1, 2, \cdots, k$ and $\widehat{\tilde{\mathbf{\Sigma}}} = a^2\hat{\mathbf{\Sigma}}$ implies $\widehat{\tilde{\sigma}}_l^2 = a^2\hat{\sigma}_l^2\ \ l = 1, 2, \cdots, k$. Furthermore, we have already showed that $\tilde{\mathbf{C}}\widehat{\tilde{\mathbf{\Sigma}}}\tilde{\mathbf{C}}^T = \mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T$. By (14), we have showed that $\hat{\delta}_l^2$ are affine-invariant. The proposition is then proved.

Proof of Proposition 2.14 First of all, under (19), we have $\tilde{\mathbf{C}}\hat{\boldsymbol{\mu}} - \tilde{\mathbf{c}} = \mathbf{P}(\mathbf{C}\hat{\boldsymbol{\mu}} - \mathbf{c})$ and $(\tilde{\mathbf{C}}\hat{\mathbf{\Sigma}}\tilde{\mathbf{C}}^T)^{-1} = (\mathbf{P}^{-1})^T(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)^{-1}\mathbf{P}^{-1}$. The invariance of T under (19) then follows.

To show that $\hat{d}$ is invariant under (19), by (15), it is sufficient to show that $\hat{\delta}_l^2$ are invariant under (19). Notice that $\tilde{\mathbf{C}} = \mathbf{P}\mathbf{C}$ implies that $\tilde{\mathbf{c}}_l = \mathbf{P}\mathbf{c}_l,\ \ l = 1, 2, \cdots, k$ and again $(\tilde{\mathbf{C}}\hat{\mathbf{\Sigma}}\tilde{\mathbf{C}}^T)^{-1} = (\mathbf{P}^{-1})^T(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)^{-1}\mathbf{P}^{-1}$. By (14), the invariance of $\hat{\delta}_l^2, l = 1, 2, \cdots, k$ under (19) then follows. The proposition is then proved.

Proof of Proposition 2.15 Let $l_1, l_2, \cdots, l_k$ be any permutation of $1, 2, \cdots, k$. Then it is easy to see that $\sum_{l=1}^{k}\mathbf{c}_l\hat{\mu}_l = \sum_{u=1}^{k}\mathbf{c}_{l_u}\hat{\mu}_{l_u}$ and $\sum_{l=1}^{k}n_l^{-1}\hat{\sigma}_l^2\mathbf{c}_l\mathbf{c}_l^T = \sum_{u=1}^{k}n_{l_u}^{-1}\hat{\sigma}_{l_u}^2\mathbf{c}_{l_u}\mathbf{c}_{l_u}^T$, showing that $\mathbf{C}\hat{\mu} = \sum_{l=1}^{k}\mathbf{c}_l\hat{\mu}_l$ and $\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T = \sum_{l=1}^{k}n_l^{-1}\hat{\sigma}_l^2\mathbf{c}_l\mathbf{c}_l^T$ are invariant under different labeling schemes of the group means $\mu_l, l = 1, 2, \cdots, k$ and so is the Wald-type test statistic T (3).

To show that $\hat{d}$ is invariant under different labeling schemes of the group means, by (15), it is sufficient to show that the denominator of $\hat{d}$ has such a property. This is actually the case by noticing

that the denominator of $\hat{d} = \sum_{l=1}^{k}(n_l-1)^{-1}\hat{\delta}_l^2 = \sum_{l=1}^{k}(n_l-1)^{-1}\left[\frac{\hat{\sigma}_l^2}{n_l}\mathbf{c}_l^T(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)^{-1}\mathbf{c}_l\right]^2 = \sum_{u=1}^{k}(n_{l_u}-1)^{-1}\left[\frac{\hat{\sigma}_{l_u}^2}{n_{l_u}}\mathbf{c}_{l_u}^T(\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T)^{-1}\mathbf{c}_{l_u}\right]^2$, where we have used the fact that $\mathbf{C}\hat{\mathbf{\Sigma}}\mathbf{C}^T$ is invariant under different labeling schemes of the group means. This completes the proof of the proposition.

References

[1] A. A. Aspin, An examination and further development of a formula arising in the problem of comparing two mean values, Biometrika 35 (1948) 88–96.

[2] B. V. Behrens, Ein Beitrag zur Fehlerberechnung bei wenige Beobachtungen, Landwirtschaftliches Jahresbuch 68 (1929) 807–837.

[3] M. B. Brown, A.B. Forsytbe, The small sample behavior of some statistics which test the equality of several means, Technometrics 16 (1974) 129–132.

[4] C.-H. Chang, J.-J. Lin, N. Pal, Testing the equality of several gamma means: a parametric bootstrap method with applications, Comput. Statist. 26 (2011) 55–76.

[5] C.-H. Chang, N. Pal, A Revisit to the Behrens-Fisher Problem: Comparison of Five test methods, Comm. Statist. Theor. & Meth. 37 (2008) 1064–1085.

[6] C.-H. Chang, N. Pal, W. K. Lim, J.-J. Lin, Comparing several population means: a parametric bootstrap method, and its comparison with usual ANOVA F test as well as ANOM., Comput. Statist. 25 (2010) 71–95.

[7] Coombs, W. T., J. Algina, D.O. Oltman, Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal, Rev. Educ. Res. 66 (1996) 137–179.

[8] R. A. Fisher, The fiducial argument in statistical inference, Ann. Eugen. 11 (1935) 141–172.

[9] E. B. Foa, B. O. Rothbaum, D. S. Riggs, T. B. Murdock, Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling, J. Consult. Clin. Psych. 59 (1991) 715–723.

[10] A. K. Gupta, Y. Wang, Some tests with specified size for the Behrens-Fisher problem, Comm. Statist. Theor. Meth. 28 (1999) 511–517.

[11] G.S. James, The comparison of several groups of observations when the ratios of the population variances are unknown, Biometrika 38 (1951) 324–329.

[12] G.S. James, Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown, Biometrika 41 (1954) 19–43.

[13] S. H. Kim, A. S. Cohen, On the Behrens-Fisher problem: a review, J. Educ. Behavior. Statist. 23 (1998) 356–377.

[14] K. Krishnamoorthy, F. Lu, A parametric bootstrap solution to the MANOVA under heteroscedasticity, J. Statist. Comput. Simul. 80 (2010) 873–887.

[15] K. Krishnamoorthy, F. Lu, T. Mathew, A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models, Comput. Statist. & Data Anal. 51 (2007) 5731–5742.

[16] K. Krishnamoorthy, Y. Xia, On selecting tests for equality of two normal mean vectors, Multi. Behavior. Res. 41 (2006) 533–548.

[17] K. Krishnamoorthy, J. Yu, Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem, Statist. Prob. lett. 66 (2004) 161–169.

[18] R. G. Krutchkoff, One-way fixed effects analysis of variance when the error variances may be unequal, J. Statist. Comput. Simul. 30 (1988) 259–271.

[19] A. M. Kshirsagar, Multivariate Analysis, Marcel Decker, New York, 1972.

[20] S. Lee, C. H. Ahn, Modified ANOVA for unequal variances, Comm. Statist. Simul. Comput. 32 (2003) 987–1004.

[21] A. Lee, J. Gurland, Size and power of tests for equality of means of two normal populations with unequal variances, J. Amer. Statist. Assoc., 70 (1975) 933–947.

[22] D. G. Nel, C. A. Van der Merwe, A solution to the multivariate Behrens-Fisher problem, Comm. Statist. Theor. Meth. 15 (1986) 3719–3735.

[23] W. R. Rice, S. D. Gaines, One-way analysis of variance with unequal variances, Proc. Natl. Acad. Sci. USA 86 (1989) 8183–8184.

[24] H. Ruben, A simple conservative and robust solution of the BehrensFisher problem, Sankhya Ind. J. Statist. Ser. A 64 (2002) 139–155.

[25] S. Weerahandi, Exact Statistical Methods in Data Analysis, Springer-Verlag, New York, 1995.

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

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

[28] B.L. Welch, On the Comparison of Several Mean Values: An Alternative Approach, Biometrika 38 (1951) 330–336.

[29] R.R. Wilcox, A new alternative to the ANOVA F and new results on James's second order method, Brit. J. Math. Statist. Psych., 41 (1988) 109–117.

[30] R.R. Wilcox, Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA model, J. Educ. Statist. 14 (1989) 269–278.

[31] H. Yanagihara, K.-H. Yuan, Three Approximate Solutions to the Multivariate Behrens-Fisher Problem, Comm. Statist. Simul. Comput. 34 (2005) 975–988.

[32] J.T. Zhang, Tests of linear hypotheses in one-way ANOVA under heteroscedasticity, Manuscript (2011).