# Testing monotonic trends among multiple group means against a composite null

Chenxiao Hu[1], Thomas Sharpton[1,2], and Duo Jiang [1]

[1]Department of Statistics, Oregon State University

[2]Department of Microbiology, Oregon State University

**Abstract**

In biomedical applications, it is often of interest to test the alternative hypothesis that the means of three or more groups follow a strictly monotonic trend such as $\mu_1 > \mu_2 > \mu_3$ against the null hypothesis that the group means are either equal or unequal but are not monotonic. This is useful, for example, for detecting biomarkers whose level in healthy, low-risk cancer and aggressive cancer subjects increases or decreases throughout the three groups. Various trend tests are available for testing monotonic alternatives. However, existing methods are designed for a highly restrictive null hypothesis where all group means are equal, which represents a special case of the null space in our problem. We demonstrate that these methods fail to control type I error when the group means may be unequal under the null. To test this broader null hypothesis, develop a greedy testing method which has an intuitive interpretation related to two-sample t tests. We show both theoretically and through simulations that the proposed method effectively controls type 1 error throughout the entire null space and achieves higher power than a naive implementation of multiple t-tests. We illustrate the greedy trend test method in real data to study microbial associations with parasite-related pathology in zebrafish.

## 1 Introduction

Statistical testing of equality of population means among three or more groups has been studied for decades. Familiar techniques including analysis of variance and non-parametric methods such as the Kruskal-Wallis test are widely used. In many applications [1, 2], it is of interest to focus on a more specific alternative

hypothesis where the group means follow a monotonic order. Various approaches are available to test for monotonic trends among the group means. For binary data, the Cochran-Armitage test[3, 4] is a modification of the Pearson Chi-square test to detect association among multiple groups. For continuous data, under some assumptions, linear regression may be performed by including the group label $i$ as an explanatory variable in order to detect a linear trend between the group means. There are also nonparametric methods such as rank-based tests for the monotonic ordering problem. Examples include the Jonckheere-Terpstra test (JT)[5, 6], in which the test statistic is constructed based on pairwise comparisons between groups, the Cuzick test (CU) [7], which extends the Mann Whitney test (MW) to compare more than two groups, the Terpstra-Magel test [8], which simultaneously compares one observation from each group, and a method more recently proposed by Shan et al. [9], which incorporates the pairwise differences in ranks between observations into the test statistic to capture extra information from the data.

However, existing trend testing methods including rank-based tests are designed to test the null hypothesis that all group means are equal. They do not account for the situation that the group means are unequal but do not follow a monotonic trend. To illustrate, suppose we have data collected on independent samples from three populations. Let the data from the $i^{th}$ population be denoted by $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{im_i})$, where $m_i$ is the sample size for group $i$. Let $\mu_1, \mu_2$ and $\mu_3$ denote the population means of the three groups. The null hypothesis tested by existing methods is $H_0 : \mu_1 = \mu_2 = \mu_3$, which we refer to as "the narrow null" in which all groups must have the same mean. The alternative hypothesis can be 1-sided such as $H_a : \mu_1 \leq \mu_2 \leq \mu_3, \mu_1 < \mu_3$ or its 2-sided analogy. Neither the null nor the alternative hypothesis covers what we refer to as a *trendless pattern*, in which the means are unequal but do not follow a monotonic order, for example, $\mu_3 < \mu_1 < \mu_2$. As a result, while the rejection of the null hypothesis by existing methods indicates that the three-group means are unlikely to be equal, it does not necessarily differentiate a monotonic trend from the trendless pattern. In practice, a trendless pattern is often plausible, in which case it can be useful and important to disentangle a trendless pattern from the truly interesting scenario where the group means do follow a monotonic trend. For example, when a biomarker whose levels in healthy, low-risk cancer and aggressive cancer subjects increase or decrease throughout the three groups, one can use the biomarker to study the prognosis of cancer. However, we argue that existing methods designed for the narrow null are not suited to study such signals because the rejection of the null hypothesis by such methods cannot be interpreted as the presence of a monotonic trend. To the best of our knowledge, no trend testing methods have been published that target the detection and delineation of a monotonic trend from a trendless pattern.

To address this challenge, we propose to formulate the problem using a framework with a less restrictive

null hypothesis. We first demonstrate how the application of existing trend testing methods to this problem result in inadequate type 1 error control. We then propose a greedy testing method tailored for the less restrictive null. The greedy test motivated by the likelihood ratio test. Unlike rank-based methods, it does not rely on the assumptions that the shapes and variances of the distributions are identical across groups. The greedy test incurs minimal computational cost and has a simple interpretation related to two-sample t-tests. In Section 4, we use simulation studies to validate the type 1 error control of the method and to evaluate its power, and describe a real data application to illustrate the use of the greedy test. In Section 5, we discuss a more general problem beyond trend testing, in which the proposed test can be used.

## 2 The narrow null versus the broad null

In this paper, we are interested in disentangling a monotonic trend among population means from a trendless pattern where the population means may or may not be equal. Because existing methods ignore the possibility of a trendless pattern in the parameter space, when the $\mu_i$'s are unequal but do not follow a monotonic trend, these methods are not guaranteed (and as we will show in the simulation studies in Section 4.1, they often fail) to adequately control type 1 error. To solve this problem, we propose to expand the narrow null hypothesis by adding the trendless pattern to it, thus yielding what we refer to as *the broad null*. In this paper, we consider the case when there are three groups, for which the alternative and the broad null hypotheses can be expressed as:

$$H_a: \ \mu_1 > \mu_2 > \mu_2 \text{ or } \mu_1 < \mu_2 < \mu_3 \tag{1}$$

$$H_0: \ \text{Otherwise} \tag{2}$$

Under this framework, we aim to develop a method that effectively controls type 1 error in the entire null space including the trendless pattern. We denote the parameter space by $\Theta = \{(\mu_1, \mu_2, \mu_3) : \mu_i \in \mathbb{R}, i = 1, 2, 3\}$, and the subspace in $\Theta$ corresponding to the alternative hypothesis $H_a$ by $\Theta_a = \{\boldsymbol{\mu} : \mu_1 < \mu_2 < \mu_3 \text{ or } \mu_1 > \mu_2 > \mu_3\}$ and the subspace corresponding to the broad null $H_0$ by $\Theta_0 = \Theta \setminus \Theta_a$ . Figure 1 shows the contrasts between the narrow and broad null hypotheses and between their corresponding alternative hypotheses.

Although few studies have identified this crucial problem and defined a suitable method to solve it, some naive procedures based two-sample t-tests (TSTTs) may be attempted to tackle it. We will describe one

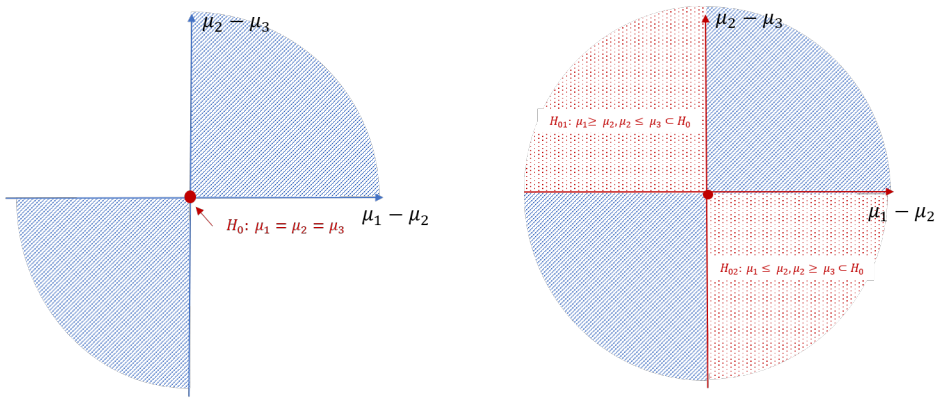


Figure 1: The narrow (Panel a) and broad (Panel b) null hypotheses and the corresponding alternative hypotheses. In each panel, the red region (including areas dotted in red and red solid lines when applicable) represents the null space, and the blue region (including blue solid lines) represents the alternative space.

such procedure using 2-sided TSTTs before introducing the greedy test in Section 3.

To utilize TSTTs to test for a strictly monotonic trend in Equation (1), one can first test the hypothesis that $\mu_1 \neq \mu_2$ and $\mu_2 \neq \mu_3$ and then verify that the sample means follow a monotonic trend. For the former, 2-sided TSTTs can be performed comparing groups 1 and 2 and comparing groups 2 and 3. We denote the $p1$ and $p2$ as the p-values of two TSTTs respectively. Because two TSTTs are conducted, the significance level for each test needs to be adjusted. Using Bonferronni correction, for example, one can conduct each TSTT at level $\alpha/2$ for a desired overall type 1 error rate controlled below $\alpha$. If both tests yield significant results at level $\alpha/2$, then both of the two pairs of means are considered significantly different. To conclude that the $\mu_i$'s follow a monotonic trend, one can then examine if the $\bar{x}_i$'s are monotonic. Namely, if the p-value of this test procedure $p = 2 * max(p1, p2)$ is less than $\alpha$ and $\bar{x}_i$'s follow a monotonic trend, we can reject the null hypothesis. In table 1, we detail the specific steps in this procedure, which we will refer to as the naive procedure based on 2-sided TSTTs.

Two other naive procedures are described in Section 5. We additionally note that the Delta method is not applicable for this problem. To attempt the Delta method, one would consider $\eta = \lambda_1 \cdot \lambda_2$ to be the parameter of interest, where $\lambda_1 = \mu_1 - \mu_2$ and $\lambda_2 = \mu_2 - \mu_3$, and construct a test based on $\hat{\eta} = (\bar{x}_1 - \bar{x}_2)(\bar{x}_2 - \bar{x}_3)$. However, the conditions for the multivariate Delta method do not hold because the gradient of $\hat{\eta}$ is zero when $\lambda 1 = \lambda_2 = 0$[10].

| Testing Steps | Test hypothesis |
|---|---|
| Step 1.<br>Compare $\mu_1$ and $\mu_2$ | Test 1 (level $\alpha/2$):<br>$H_{01} : \mu_1 = \mu_2$<br>$H_{a1} : \mu_1 \neq \mu_2$ |
| Step 2.<br>Compare $\mu_2$ and $\mu_3$ | Test 2 (level $\alpha/2$):<br>$H_{02} : \mu_2 = \mu_3$<br>$H_{a2} : \mu_2 \neq \mu_3$ |
| Step 3.<br>Check the order of the sample means | $\bar{x}_1 > \bar{x}_2 > \bar{x}_3$<br>or<br>$\bar{x}_1 < \bar{x}_2 < \bar{x}_3$ |
| Reject the overall $H_0$ if both of the following hold:<br>(i) both tests 1 and 2 yield a rejection, and (ii)the sample means have a monotonic order | |

Table 1: Naive testing procedure based on 2-sided TSTTs.

# 3 A trend test for the broad null

We focus on the problem of testing the hypothesis in Equation (1). We allow the sample sizes and the population variances to be different between the three groups. We assume a reasonably large sample size for all three groups so that the sample mean of each group is well approximated by a normal distribution based on the central limit theorem. The asymptotic distributions of the sample means are given by $\bar{x}_i \sim N(\mu_i, \sigma_i^2)$, where $\mu_i$ is the population mean of group $i$ and $\sigma_i^2 = \mathrm{Var}(\bar{x}_i)$ is the variance of the $i^{th}$ sample mean for $i = 1, 2, 3$.

## 3.1 A test statistic motivated by the likelihood-ratio test

We will first tackle the case if $\sigma_i$'s are known, and the general case with unknown $\sigma_i$'s will be discussed later (Section 3.2). Ignoring constants that do not depend on the parameters $\mu_i$'s, the log-likelihood function can be written as $\ell(\mu_1, \mu_2, \mu_3) = -\frac{1}{2}\left(\frac{(\bar{x_1}-\mu_1)^2}{\sigma_1^2} + \frac{(\bar{x_2}-\mu_2)^2}{\sigma_2^2} + \frac{(\bar{x_3}-\mu_3)^2}{\sigma_3^2}\right)$. To derive the likelihood ratio test, we start by finding the restricted and unrestricted maximum likelihood estimators (MLEs) for the unknown parameters $\mu_1, \mu_2$ and $\mu_3$. It is easily seen that the unrestricted MLE is given by the sample means $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$. The restricted MLE which maximizes the likelihood function in the null space, we need to consider two situations, depending on whether $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ is consistent with $\Theta_0$. We therefore divide the "data space" $\Theta^x := \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_i \in \mathbb{R},\ i = 1, 2, 3\}$ into two parts (Figure 2), one consistent with $\Theta_a$ and one with $\Theta_0$:

$$\Theta_a^x = \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) | \bar{x}_1 < \bar{x}_2 < \bar{x}_3 \ or\ \bar{x}_1 > \bar{x}_2 > \bar{x}_3\} \tag{3}$$
$$\Theta_0^x = \Theta^x \setminus \Theta_a^x$$

The following lemma concerns the restricted MLE.

**Lemma 1.** *1. When* $(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) \in \Theta_0^x$*, the restricted MLE which maximizes the log-likelihood function over* $\Theta_0$ *is* $(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (\bar{x}_1, \bar{x}_2, \bar{x}_3)$.

*2. When* $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Theta_a^x$*, the restricted MLE which maximizes the log-likelihood function over* $\Theta_0$ *is*

$$(\hat{\mu_1}, \hat{\mu_2}, \hat{\mu_3}) = \begin{cases} (\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}), & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \geq \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}} \\ (\tilde{\tilde{\mu_1}}, \tilde{\tilde{\mu_2}}, \tilde{\tilde{\mu_3}}), & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} < \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}}, \end{cases}$$

*where* $(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}) = \left(\bar{x}_1, \frac{\bar{x}_2\sigma_3^2 + \bar{x}_3\sigma_2^2}{\sigma_3^2 + \sigma_2^2}, \frac{\bar{x}_2\sigma_3^2 + \bar{x}_3\sigma_2^2}{\sigma_3^2 + \sigma_2^2}\right)$, $(\tilde{\tilde{\mu_1}}, \tilde{\tilde{\mu_2}}, \tilde{\tilde{\mu_3}}) = \left(\frac{\bar{x}_1\sigma_2^2 + \bar{x}_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\bar{x}_1\sigma_2^2 + \bar{x}_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \bar{x}_3\right)$.

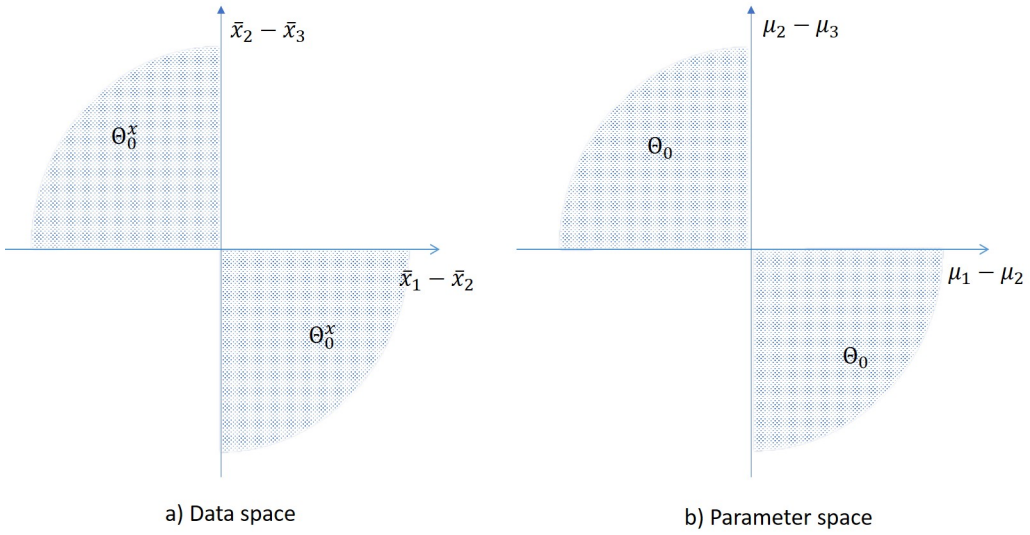


Figure 2: Partitions of the data space $\Theta^x$ and the parameter space $\Theta$.

To intuitively explain the Lemma, when $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Theta_0^x$, the data are consistent with the pattern in the null hypothesis and therefore the restricted MLE coincides with the unrestricted MLE.

When $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Theta_a^x$, under the null, the MLEs are attained when two populations share the same mean. In other words, the MLEs are taken on one of the two boundaries of $\Theta_0$, which is not surprising given that the trend in the data does not align with the ordering of the population means. The smaller of the two standardized distances (i.e. $\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ and $\frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}}$) reveals which of the two boundaries of $\Theta_0$ yields a greater log likelihood.

Given the restricted and unrestricted MLEs, we can derive the maximum value of the log-likelihood function under restricted and unrestricted cases and hence the likelihood ratio. First, we derive the maximum of the log-likelihood under $\Theta_0$. When$(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Theta_a^x$, plugging in the MLEs for $(\mu_1, \mu_2, \mu_3)$ given in Lemma 1 into $\ell(\mu_1, \mu_2, \mu_3)$ yields a log-likelihood value that depends on the standardizes distances. When the sample means do not follow a monotonic trend, the restricted MLEs coincide with the unrestricted MLEs, and thus the likelihood ratio statistic is zero. These results are summarized in Theorems 2 and 3.

**Theorem 2.** *When* $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Theta_a^x$, $\sup_{\boldsymbol{\mu} \in \Theta_0} \ell(\mu_1, \mu_2, \mu_3) = -\frac{1}{2} \min\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}}\right)$

**Theorem 3.** *The likelihood ratio statistics is given by,*

$$\lambda = \begin{cases} -2 \min\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}}\right), & \text{if } \bar{x}_1 < \bar{x}_2 < \bar{x}_3 \text{ or } \bar{x}_1 > \bar{x}_2 > \bar{x}_3 \\ 0, & \text{Otherwise} \end{cases}$$

This statistic makes intuitive sense as a way of measuring how inconsistent the data is with $H_0$, because when sample means have a monotonic trend and the minimal of the distances between sample means is large, the population means are likely to have a strictly monotonic trend.

Motivated by the form of the likelihood ratio test, we define our test statistic as follows

$$T(\boldsymbol{x}) = \begin{cases} \min\left(\frac{|\bar{x}_1-\bar{x}_2|}{\sqrt{\sigma_1^2+\sigma_2^2}}, \frac{|\bar{x}_2-\bar{x}_3|}{\sqrt{\sigma_2^2+\sigma_3^2}}\right), & \text{if } \bar{x}_1 \leq \bar{x}_2 \leq \bar{x}_3 \text{ or } \bar{x}_1 \geq \bar{x}_2 \geq \bar{x}_3 \\ 0, & \text{Otherwise} \end{cases} \tag{4}$$

## 3.2 Assessing Statistical Significance

Let $t$ be the realized value of the test statistic defined in Equation (2). For fixed $\boldsymbol{\mu}$, $P_{\boldsymbol{\mu}}(T(\boldsymbol{x}) > t)$ is a measure of compatibility of the observed data with null hypothesis. To effectively control type 1 error in the entire null space, we obtain the p-value as $p = \sup_{\boldsymbol{\mu}\in\Theta_0} g(\boldsymbol{\mu})$. With some reparameterization, we will show that the p-value has a simple and intuitive analytical form.

Recall that $\lambda_1 = \mu_1 - \mu_2$ and $\lambda_2 = \mu_2 - \mu_3$. Note that $H_0$ can be equivalently expressed in terms of $\lambda_1$ and $\lambda_2$ as $\lambda_1\lambda_2 > 0$. It is not hard to show that $P_{\boldsymbol{\mu}}(T(\boldsymbol{x}) > t)$ depends on $\boldsymbol{\mu}$ only through $\lambda_1$ and $\lambda_2$. We hence let $g(\lambda_1, \lambda_2) = P_{\lambda_1,\lambda_2}(T(\boldsymbol{x}) > t)$, and it follows that the p-value is given by $p = \sup_{\boldsymbol{\mu}\in\Theta_0} g(\lambda_1, \lambda_2)$. Under the null, the population means have higher chance to yield strictly monotonically ordered sample means when two of the sample means are equal. Therefore, the measure of compatibility is higher when one of the $\lambda$s is zero and lower when otherwise. In fact, the measure of compatibility is maximized when one of the $\lambda$'s is zero and the other approaches infinity. This is formalized in the following Lemma.

**Lemma 4.** $\sup_{\boldsymbol{\mu}\in\Theta_0} g(\lambda_1, \lambda_2) = \max(g(0,\infty), g(\infty,0), g(0,-\infty), g(-\infty,0))$

By evaluating the right hand side of the equation above, the following Theorem for the p-value follows.

**Theorem 5.** *The p-value is given by* $1 - \Phi(t)$*, where* $\Phi(\cdot)$ *is the cumulative distribution function of* $N(0,1)$ *and* $t$ *is the observed value of the test statistic* $T$*.*

With the great simplicity of the form for the p-value, the greedy test can be implemented using a two-step procedure: 1) find the test statistic by taking the minimum of the two standardized distances. 2) derive the p-value by obtaining the upper tail cumulative probability for the standard normal distribution. It is noteworthy that the test only depends on the standardized distances between the group means and is unaffected by the amount of correlation between the standardized distances.

In practice, when $\sigma_i^2$'s are unknown, we replace them with a consistent estimator such as the sample variance estimator given by $\hat{\sigma}_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_i)^2$ in the test statistic defined in Equation (4). The test statistic hence becomes

$$T(\boldsymbol{x}) = \begin{cases} \min\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\hat{\sigma}_2^2 + \hat{\sigma}_3^2}}\right), & \text{if } \bar{x}_1 \leq \bar{x}_2 \leq \bar{x}_3 \text{ or } \bar{x}_1 \geq \bar{x}_2 \geq \bar{x}_3 \\ 0, & \text{Otherwise} \end{cases} \tag{5}$$

### 3.3 Rejection region and connection with two-sample t-tests

Given a desired significance level $\alpha$, it is easily seen that the rejection region of the greedy test is given by $\{\boldsymbol{x} : p < \alpha\} = \{\boldsymbol{x} : T_1(\boldsymbol{x}) \cdot T_2(\boldsymbol{x}) > 0, \quad |T_1(\boldsymbol{x})| > z_{1-\alpha}, \quad |T_2(\boldsymbol{x})| > z_{1-\alpha}\}$, where $T_1(\boldsymbol{x}) = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$ and $T_2(\boldsymbol{x}) = \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\hat{\sigma}_2^2 + \hat{\sigma}_3^2}}$ are 2-sided TSTT statistics, and $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. This brings about a natural interpretation of the greedy test method based on TSTTs. It can be equivalently implemented using the following simple procedure:

1. Are the sample means located in $\Theta_a^x$ (i.e. do the sample means follow a monotonic trend)? If so, go to Step 2; If not, fail to reject the null (p-value $> 0.5$).

2. Conduct 2-sided TSTTs at level $2\alpha$ between groups 1 and 2 and between groups 2 and 3.

3. Follow either one of the follow two procedures, which are equivalent:

   - Reject the null if both TSTTs reject their corresponding nulls at level $2\alpha$.
   - Equivalently, let the p-values of the TSTT's be $p_1$ and $p_2$. The final p-value of the greedy test can be obtained by taking $p = \max(p_1, p_2)/2$. Reject the null if $p < \alpha$.

It is worth noting that the procedure described above is closely connected to the naive procedure based on 2-sided TSTTs (see Section 2 and Table 1). The key difference between the two is that the greedy test requires both 2-sided TSTTs to be performed at level $2\alpha$, whereas the naive procedure performs both TSTTs at level $\alpha/2$.

In other words, the p-values for 2-sided TSTT is a linear transformation from the greedy test. Recall the p-value for 2-sided TSTT is $p_{2-sidedTSTT} = 2\max(p_1, p_2)$, so $p_{2-sidedTSTT} = 4p_{greedy}$. Therefore, the greedy test is expected to achieve substantial power gain over the 2-sided TSTT procedure and the listed naive procedures in Appendix C.

# 4 Results

## 4.1 Simulation studies

First, we verify the effective type 1 error control of the greedy test and three naive procedures: one based on 2-sided TSTTs as described in Section 2, one based on 1-sided TSTTs and one based on confidence intervals (see Appendix C). In addition, we show that existing methods proposed for the narrow null do not control type 1 error effectively. As examples of rank-based methods, we focus on JT and CU, which, like other rank-based tests, are proposed to test the narrow null and further assume that the populations follow the same distribution under the null.

We consider the following simulation settings from the broad null.

I. Boundary of the null: Two of the groups have the same mean, which differs from the mean of the third group. The three groups are assumed to follow N(0,1), N(0,1), and N($\mu_3$, 1), respectively, where $\mu_3$ is varied from 0 to 1.

II. Interior of the null: All three group means are different but do not follow a monotonic trend. The three groups are assumed to follow N(0,1), N(1,1), and N($\mu_3$, 1), respectively, where $\mu_3$ is varied from 0 to 1.

III. Unequal variances and different distribution families: The three groups do not come from the same distribution family and do not have the same variance. The three groups are assumed to follow N(1,10), N(1,1), and Poisson($\mu_3$), respectively, where $\mu_3$ is varied from 1 to 2.

For each simulation setting, we simulate $m_i = 30$ samples independently from each of the three populations. Then, the type error 1 rate of a method is estimated based on 10,000 replicates as the frequency at which the p-value is below a nominal level of $\alpha = 0.05$.

Across the simulation settings, the greedy test as well as the naive methods effectively control type 1 error at level 0.05. However, both CU and JT suffer from severely inflated type 1 error for a vast majority of the settings. Settings I and II assume normally distributed data with equal variances across groups. In setting I (Figure 1a), the first two groups have the same mean and hence the parameters are located on the boundary of the broad null which we aim to test. In this setting, the parameters are consistent with the narrow null if and only if $\mu_3 = 0$, in which case type 1 error is also well controlled by the rank-based tests. However, when $\mu_3 > 0$, the parameters are treated as being part of the alternative space by rank-based methods. As such, the rejection probability tends to fall well above the nominal level 0.05 for JT and CU.

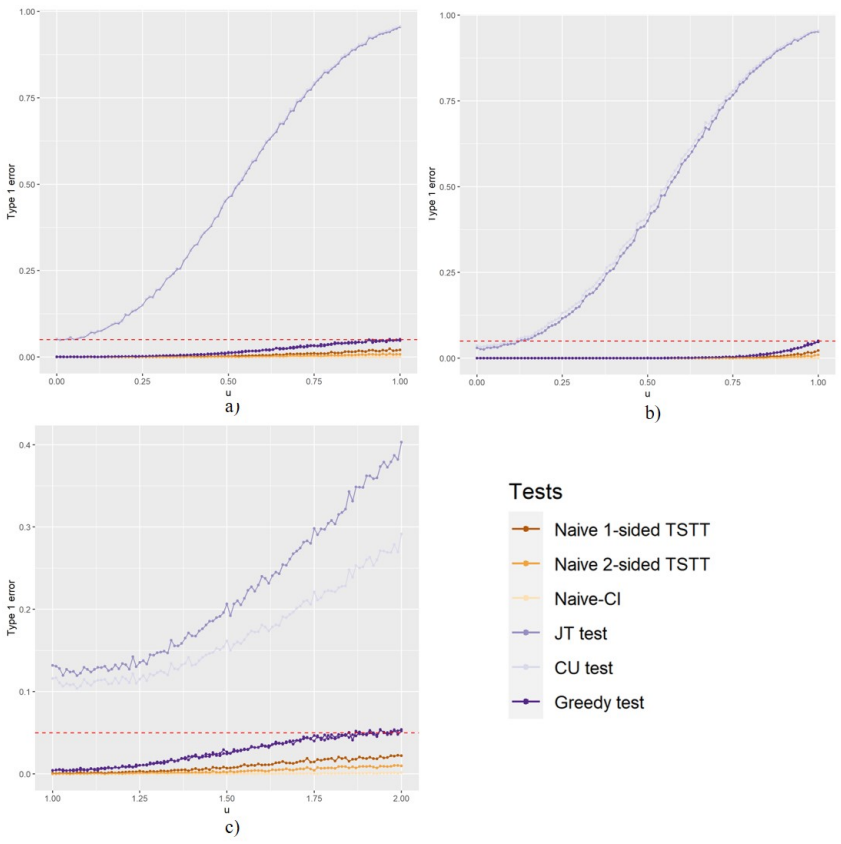


Figure 3: Empirical type 1 error results. The three panels compare the empirical type 1 error rates of the greedy test, CU, JT and three naive procedures under settings I) boundary of null, II) interior of null and III) unequal variances and different distribution families, respectively. Each curve shows the empirical type 1 error rate (vertical axis) as a function of $mu_3$ (horizontal axis). The curves are color coded to represent the methods in comparison. The horizontal dashed line display the nominal level at 0.05. The greedy test and the naive procedures consistently controls type 1 error effectively, whereas both CU (grey curve) and JT (light purple curve) fail to control type 1 error in the majority of the settings.

In setting II, $\mu_1 \leq \mu_3 \leq \mu_2$. When $\mu_3$ is strictly between $\mu_1$ and $\mu_2$, the group means are all unequal but not monotonic. This setting is in fact treated by rank-based methods as implausible because it belongs neither to the narrow null nor to the alternative hypothesis targeted by rank-based methods. In this case, it is not surprising that rank-based methods do not guarantee adequate control of type 1 error. In fact, JT and CU can result in excessive false positives with a rate of up to 100%. In summary, rank-based methods are severely mis-calibrated both on the boundary and in the interior of the broad null hypothesis, whereas the greedy test and the naive procedures effectively control type 1 error globally.

For setting III, the three populations have distributions from different families. When $\mu_3 > 1$, CU and JT result in excessive type 1 error for the same reason that they do in Setting I. In contrast to Setting II,

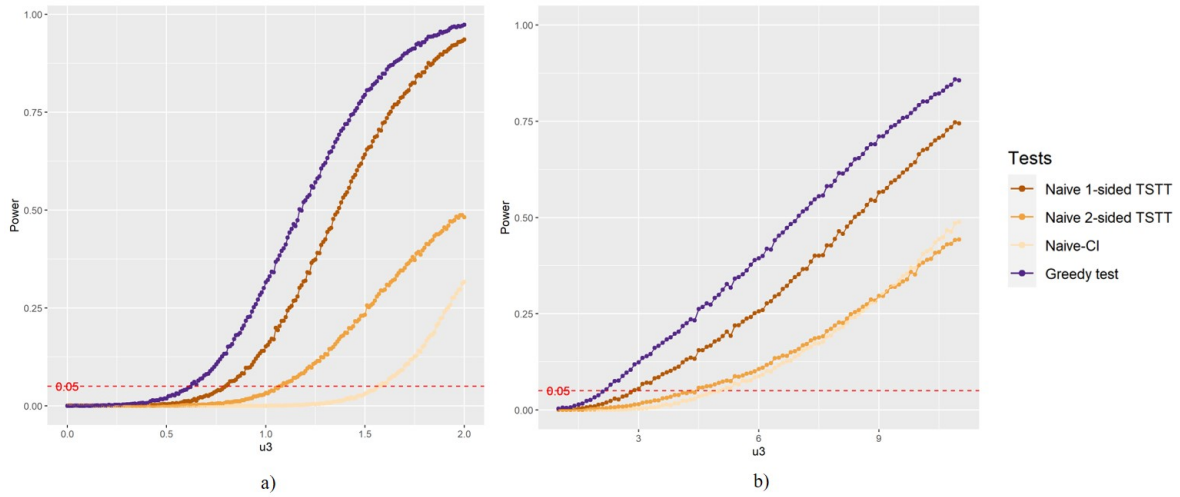


Figure 4: Empirical power results of the greedy test and three naive procedures for settings IV and V. Each curve shows the empirical power (vertical axis) of a method as a function of $\mu_3$ (horizontal axis). The curves are color coded to represent the four methods. Across the settings, the greedy test (purple curve) is the most powerful, followed by the naive procedure based on 1-sided TSTTs (Brown curve), the naive procedure based on 2-sided TSTTs (dark yellow curve), and the naive procedured based on confidence intervals (light yellow curve).

the three groups follow distributions from different families and have unequal variances, which is a violation of the assumption of rank-based methods that the groups have identical distributions under the null. This explains why CU and JT have inflated type 1 error rate even when $\mu_1 = \mu_2 = \mu_3 = 1$. Effective type 1 error control by the greedy test and the naive procedures, however, do not rely on the identical distribution assumption.

We then evaluate the power of our approach and compare it with the naive methods. Rank-based methods are not included in the power simulations, because they do not control type 1 error adequately. To this end, we consider two settings under the alternative hypothesis.

IV. Increasing trend: The three groups follow normal distributions with monotonically increasing means and the same variance. The three groups are assumed to follow $N(0,1)$, $N(\frac{1}{2}\mu_3, 1)$, and $N(\mu_3, 1)$, respectively, where $\mu_3$ is varied from 0 to 1.

V. Unequal variances and different distributions families: The three groups do not come from the same distribution family and do have the same variance. The group means follow an increasing trend. The three groups are assumed to follow $N(1, 10)$, $N(1 + \frac{1}{2}\mu_3, 1)$, and Poisson$(1 + \mu_3)$, respectively, where $\mu_3$ is varied from 0 to 10.

In both settings, power grows for all methods as $\mu_3$ increases. Throughout the simulations, the greedy

test achieves the highest power. This demonstrates the power advantage of our method over the naive procedures.

## 4.2 Real data application

We illustrate the greedy test in an application to the gut microbiome data in a zebrafish study from Gaulke et al. 2019 [11]. The study aimed to investigate the link between the gut microbiome and parasite infection and pathology. As part of the study, a sample of 105 zebrafish were exposed to Pseudocapillaria tomentosa, a parasite that infects the gut of the fish to induce pathological changes including intestinal inflammation, tissue damage, and hyperplasia. The overarching goal of our analysis is to detect monotonic trends in the gut microbiome that associate with host pathology.

The gut microbiome of the fish was measured over 12 weeks of infection via cross-sectional 16S sequencing, which generated the relative abundances of 565 genera for each fish. Two types of pathological changes on the parasite-exposed fish, inflammation and hyperplasia, were assessed by a pathologist based on examination of intestine tissues. For either type of changes, severity was scored between 0 and 3 for each fish. To detect microbiome features that vary monotonically with the severity of pathological changes, we group the samples into three categories based on the inflammation score: no inflammation (score = 0, n = 47), mild to moderate inflammation (score = 1 or 2, n = 45), and severe inflammation (score = 3, n = 10). We also group the samples based on the hyperplasia score: no hyperplasia (score =0, n = 32), mild and moderate hyperplasia (score = 1 or 2, n = 53), and severe hyperplasia (score = 3, n = 17).

We first aim to detect overall patterns of the gut microbiome that associate with inflammation severity. To this end, we perform non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarity and principal component analysis (PCA), on the microbiome data. For either NMDS or PCA, we retain the top three components, which are then tested individually for monotonic trends with inflammation severity. Bonferroni-corrected p-values are reported in Table 2 based on the proposed greedy test, the naïve procedure based on 1-sided TSTT, and the naïve procedure based on 2-sided TSTT. The greedy test finds strong evidence that the second NMDS component varies strictly monotonically across the severity groups of inflammation (MDS2, Bonferroni corrected p = 0.0030). The monotonic trend is also identified by the two naïve procedures, although with less significant p-values (Bonferroni-corrected p = 0.0059 and 0.0012, respectively). Based upon the PCA results, the greedy test identifies a statistically significant monotonic trend associated with the second principal component (PC2, Bonferroni-corrected p = 0.037). The naïve procedures (Bonferroni-corrected p = 0.074 and 0.012, respectively), however, fail to capture this signal at

| NMDS | | | |
|---|---|---|---|
| | Greedy test | Naive 1-sided | Naive 2-sided |
| MDS2 | **0.0030** | **0.0059** | **0.012** |
| PCA | | | |
| | Greedy test | Naive 1-sided | Naive 2-sided |
| PC2 | **0.037** | 0.074 | 0.15 |
| Genus-level Tests | | | |
| | Greedy test | Naive 1-sided | Naive 2-sided |
| *Cetobacterium* | 0.11 | 0.22 | 0.44 |

Table 2: Bonferroni-corrected p-values in zebrafish gut microbiome analysis, based on the greedy test, the naïve procedure based on 1-sided TSTT (Naïve 1-sided) and the naïve procedure based on 2-sided TSTT (Naïve 2-sided). P-values below 0.05 are in bold type, and p-values between 0.05 and 0.2 are italicized. Microbial features for which none of the methods yields a raw p-value below 0.05 are omitted.

level 0.05 due to their lower power compared with the greedy test. These results indicate that there exist overall features of the gut microbiome that vary monotonically with increasing severity level of inflammation in the intestine among parasite-exposed zebrafish.

In addition to overall microbiome patterns, we are also interested in identifying specific microbial taxa that are monotonically associated with inflammation severity. We focus the analysis on the common genera which are present in at least 20% of the samples. A total of 49 such genera are identified in the data. The relative abundance of each genus is tested for monotonic trends across inflammation severity groups using the greedy test and two naïve procedures (Table 2). Based on the greedy test, there is suggestive evidence that *Cetobacterium* decreases monotonically as inflammation severity elevates from no inflammation to severe inflammation (Bonferroni-corrected p = 0.11, $\bar{x} = (0.3287, 0.2092, 0.0385)$). This finding reinforces the discovery of a negative association between *Cetobracterium* and parasite burden in Gaulke et al, 2019[11]. We point that, using the naïve trend testing procedures, the monotonic trend associated with *Cetabacterium* would yield substantially greater p-values (Bonferroni-corrected p = 0.22 and 0.44, respectively).

Similar analyses are performed for the groups defined based on hyperplasia scores, and no significant signals are found. Overall, our results indicate that, both the greedy test and the naïve procedures do well in capturing highly significant monotonic trends, the power gain achieved by the greedy test has the potential of enabling the discoveries of mild to moderate signals that would be missed by the use of the naïve tests.

# 5 Discussion

We have proposed a greedy testing method for a strictly monotonic trend among the means of three independent groups against a composite null hypothesis that the group means are equal or unequal but are not

strictly monotonic. We demonstrate both theoretically and empirically that existing testing methods are not valid for this problem because they target a narrow null in which the group mean are all equal. To address this issue, we have developed a greedy testing method that achieves global type 1 error control in the entire null space. Moreover, unlike most nonparametric methods based on rank, the proposed method does not require the groups to have the same distribution under the null or the same variance and therefore is more flexible. The greedy test is computationally simple and the test statistic has a simple expression because of its connection with TSTTs. However, compared to a naive procedure based on pairwise TSTTs, the greedy test improves power substantially because we have shown that, for each pairwise comparison, it only requires a significance level that is four times the significance level entailed by the naive procedure to achieve global type 1 error control.

We have shown in simulation studies that methods intended for testing the narrow null result in excessive false positives when used to test the broad null, whereas the greedy test effectively controls the type 1 error rate in a wide range of settings including those in which the groups have distributions from different families and different variances. In addition, we have demonstrated that the proposed test outperforms three naive procedures that we have described in power. Finally, we have applied the method to the data from GSEP, which is a case-control study on prostate cancer. We have found suggestive evidence that Gleason score increases consistently across the three age groups defined as (age $< 60$, $60 \leq$ age $\leq 70$,age $> 70$) and that PSA level increases consistently as the minor allele count for rs851023 goes from 0 to 2.

Although the development of the method has been primarily motivated by the trend test problem with a broad null, the method can be used to solve a more general class of problems beyond trend tests. To see this, we first write the hypotheses in (1) and (2) in the following form

$$H_0 : \lambda_1 \cdot \lambda_2 > 0$$
$$H_a : \lambda_1 \cdot \lambda_2 \leq 0.$$

and write the test statistic in Equation (4) as

$$T := \min\{|T_1|, |T_2|\} \cdot I\{T_1 \cdot T_2 > 0\} \tag{6}$$

In the context of a trend test between three groups, $\lambda_1$ and $\lambda_2$ represent the pairwise differences in means and $T_1$ and $T_2$ are the corresponding 1-sided test statistics. However, the proof of Lemma 4 does not rely on these specific definitions of $\lambda_1$ and $\lambda_2$. In a more general framework, $\lambda_1$ and $\lambda_2$ can be any parameters for

which it is of interest to test whether they are both nonzero and have the same sign. For this general setup, the test statistic defined in Equation (6) can be used and its p-value can be assessed using the expression given in Theorem 5. The proof of Lemma 4 in the Appendix is still valid to show that Theorem 5 holds in the more general setup, so long as the following conditions hold: a) $T_1$ and $T_2$ are valid 2-sided statistics for $\lambda_1 = 0$ and $\lambda_2 = 0$, respectively; b) $(T_1, T_2)$ has an (asymptotic) bivariate normal distribution, $E(T_i)$ has the same sign as $\lambda_i$, $\text{var}(T_i) = 1$, and $\text{cov}(T_1, T_2) \leq 0$. In particular, this includes the setting in which $T_1$ and $T_2$ are independent z-test statistics. This is potentially relevant to many medical applications. For example, when two separate studies both evaluate the association between a genetic factor and a disease, the test statistics are independent provided that the studies are conducted on separate cohorts. In this case, the greedy test has the potential to take advantage of combining the summary statistics of the two studies to boost power by enabling the detection of a significant association at level $\alpha$ so long as the p-values from both studies are below $2\alpha$.

## Acknowledgment

I would like to express my gratitude to Professor Duo Jiang for extraordinary support and encouragement for this work, for commenting on earlier drafts of this paper, and for advising on the presentation and positioning of this paper. I also would like to thank Professor Yuan Jiang and Professor Thomas Sharpton for their valuable feedback.

## References

[1] Gerald van Belle and James P. Hughes. Nonparametric tests for trend in water quality. *Water Resources Research*, 20(1):127–136, 1984.

[2] Michael Budde and Peter Bauer. Multiple test procedures in clinical dose finding studies. *Journal of the American Statistical Association*, 84(407):792–796, 1989.

[3] William G. Cochran. Some methods for strengthening the common chi-square tests. *Biometrics*, 10(4):417–451, 1954.

[4] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[5] A. R. Jonckheere. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145, 1954.

[6] T. J. TERPSTRA. The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. *Indagations Math.*, 14:327–333, 1952.

[7] Jack Cuzick. A wilcoxon-type test for trend. *Statistics in Medicine*, 4(1):87–90, 1985.

[8] Jeffrey Terpstra and Rhonda Magel. A new nonparametric test for the ordered alternative problem. *Journal of Nonparametric Statistics*, 15(3):289–301, 2003.

[9] Guogen Shan, Daniel Young, and Le Kang. A new powerful nonparametric rank test for ordered alternative problem.(research article). *PLoS ONE*, 9(11), 2014.

[10] Richard Barfield, Jincheng Shen, Allan C. Just, Pantel S. Vokonas, Joel Schwartz, Andrea A. Baccarelli, Tyler J. VanderWeele, and Xihong Lin. Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology*, 41(8):824–833, 2017.

[11] Christopher A. Gaulke, Mauricio L. Martins, Virginia G. Watral, Ian R. Humphreys, Sean T. Spagnoli, Michael L. Kent, and Thomas J. Sharpton. A longitudinal assessment of host-microbe-parasite interactions resolves the zebrafish gut microbiome's link to pseudocapillaria tomentosa infection and pathology. *Microbiome*, 7(1), 2019.

[12] Milton Abramowitz. Handbook of mathematical functions, with formulas, graphs, and mathematical tables,, 1965.

# Appendix A

In this Appendix, we show the results in Section 3.1

## A1. Proof of Lemma 1

We start by introducing some notation that we will use to partition the parameter space $\Theta$ and the data space $\Theta^x := \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_i \in \mathbb{R},\ i = 1, 2, 3\}$. First, we define the four disjoint regions of the parameter space illustrated in Figure 1b as follows (see Figure 5a)

$$A = \{(\mu_1, \mu_2, \mu_3) : \mu_3 \leq \mu_2, \mu_1 \leq \mu_2\} \subset \Theta_0$$

$$B = \{(\mu_1, \mu_2, \mu_3) = \mu_3 \geq \mu_2, \mu_1 \geq \mu_2\} \subset \Theta_0$$

$$C = \{(\mu_1, \mu_2, \mu_3) = \mu_1 > \mu_2 > \mu_3\} \subset \Theta_0^c$$

$$D = \{(\mu_1, \mu_2, \mu_3) = \mu_1 < \mu_2 < \mu_3\} \subset \Theta_0^c$$

Analogous regions in the data space can also be defined (see Figure 5b):

$$A_x = \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_3 \leq \bar{x}_2, \bar{x}_1 \leq \bar{x}_2\} \subset \Theta_0^x$$

$$B_x = \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_3 \geq \bar{x}_2, \bar{x}_1 \geq \bar{x}_2\} \subset \Theta_0^x$$

$$C_x = \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_1 > \bar{x}_2 > \bar{x}_3\ \} \subset \Theta_a^x$$

$$D_x = \{(\bar{x}_1, \bar{x}_2, \bar{x}_3) : \bar{x}_1 < \bar{x}_2 < \bar{x}_3\} \subset \Theta_a^x$$

We will present the proof of Lemma 1 for the case when $\bar{\boldsymbol{x}} \in D_x$. For $\bar{x} \in C_x$, the results can be shown using a similar argument. Assuming $\bar{\boldsymbol{x}} \in D_x$, we first prove the following lemma on the maximization of the log-likelihood function $l(\mu_1, \mu_2, \mu_3)$ for $\boldsymbol{\mu} \in A$.

**Lemma 6.** *Assuming $\bar{\boldsymbol{x}} \in D_x$, we have*

$$(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}) = \underset{\boldsymbol{\mu} \in A}{\operatorname{argmin}}\, l(\mu_1, \mu_2, \mu_3) \tag{7}$$

$$\left(\tilde{\tilde{\mu_1}}, \tilde{\tilde{\mu_2}}, \tilde{\tilde{\mu_3}}\right) = \underset{\boldsymbol{\mu} \in B}{\operatorname{argmin}}\, l(\mu_1, \mu_2, \mu_3) \tag{8}$$

*where $(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3})$ and $\left(\tilde{\tilde{\mu_1}}, \tilde{\tilde{\mu_2}}, \tilde{\tilde{\mu_3}}\right)$ are as defined in Lemma 1.*

*Proof.* We first note that $\tilde{\mu_2} = \tilde{\mu_3}$ and $\tilde{\mu_2} = \frac{\bar{x}_2\sigma_3^2 + \bar{x}_3\sigma_2^2}{\sigma_3^2 + \sigma_2^2} \geq \frac{\bar{x}_1\sigma_3^2 + \bar{x}_1\sigma_2^2}{\sigma_3^2 + \sigma_2^2} = \bar{x}_1 = \tilde{\mu_1}$, where the last inequality

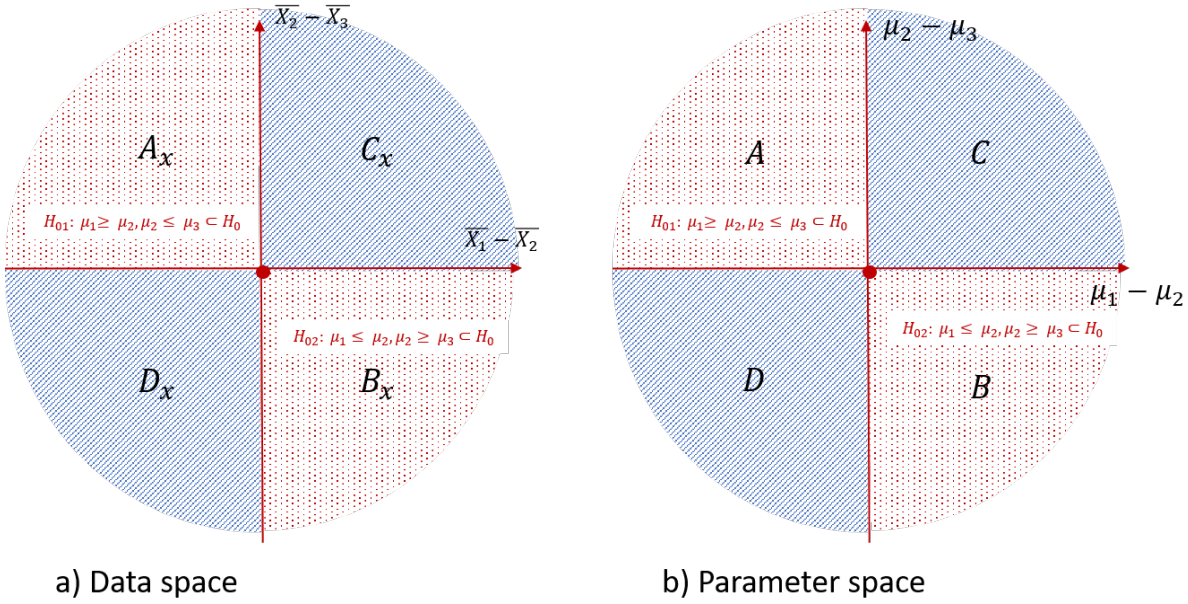


Figure 5: Further partitions of the data space $\Theta^x$ and the parameter space $\Theta$.

holds because $\bar{\boldsymbol{x}} \in D_x$. This implies that $(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}) \in A$. To show equation (7), it remains to show that

$$l\left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right) \geq l\left(\mu_1, \mu_2, \mu_3\right), \tag{9}$$

for any $(\mu_1, \mu_2, \mu_3) \in A$. To this end, we note that for such $(\mu_1, \mu_2, \mu_3) \in A$,

$$\begin{aligned} l\left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right) - l\left(\mu_1, \mu_2, \mu_3\right) &= \frac{\left(\bar{x}_2 - \mu_2\right)^2}{2\sigma_2^2} - \frac{\left(\bar{x}_2 - \frac{\bar{x}_2\sigma_3^2 + \bar{x}_3\sigma_2^2}{\sigma_3^2 + \sigma_2^2}\right)^2}{2\sigma_2^2} + \frac{\left(\bar{x}_3 - \mu_3\right)^2}{2\sigma_3^2} + \frac{\left(\bar{x}_3 - \frac{\bar{x}_2\sigma_3^2 + \bar{x}_3\sigma_2^2}{\sigma_3^2 + \sigma_2^2}\right)^2}{2\sigma_3^2} \quad (10) \\ &= \frac{1}{\sigma_2^2 + \sigma_3^2}\left(\bar{x}_3 - \bar{x}_2\right)\left(\mu_2 - \mu_3\right) + \frac{1}{\sigma_2^2}\left(p\bar{x}_2 + q\bar{x}_3 - \mu_2\right)^2 + \frac{1}{\sigma_3^2}\left(p\bar{x}_2 + q\bar{x}_3 - \mu_3\right)^2 \quad (11) \\ &\geq \frac{1}{\sigma_2^2 + \sigma_3^2}\left(\bar{x}_3 - \bar{x}_2\right)\left(\mu_2 - \mu_3\right) \geq 0, \end{aligned}$$

where $p = \frac{\sigma_3^2}{\sigma_2^2+\sigma_3^2}$, $q = \frac{\sigma_2^2}{\sigma_2^2+\sigma_3^2}$, and the last inequality follows from $\mu_2 \geq \mu_3$ and $\bar{x}_2 \leq \bar{x}_3$. Equation (7) is thereby proved. Equation (8) can be shown using a similar argument.

□

The following lemma evaluates the log likelihood function at the optima given in Lemma 6.

**Lemma 7.**

$$l\left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right) = -\frac{\left(\bar{x}_2 - \bar{x}_3\right)^2}{\sigma_2^2 + \sigma_3^2}$$
$$l\left(\tilde{\tilde{\mu}_1}, \tilde{\tilde{\mu}_2}, \tilde{\tilde{\mu}_3}\right) = -\frac{\left(\bar{x}_1 - \bar{x}_2\right)^2}{\sigma_1^2 + \sigma_2^2}$$

The proof of Lemma 7 is a straightforward substitution into the log likelihood function and is omitted here.

Combing Lemma 6, Lemma 7, and $\Theta_0 = A \cup B$,

$$\left(\hat{\mu_1}, \hat{\mu_2}, \hat{\mu_3}\right) = \begin{cases} \left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right), & \text{if } l\left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right) \geq l\left(\tilde{\tilde{\mu}_1}, \tilde{\tilde{\mu}_2}, \tilde{\tilde{\mu}_3}\right) \\ \left(\tilde{\tilde{\mu}_1}, \tilde{\tilde{\mu}_2}, \tilde{\tilde{\mu}_3}\right), & \text{if } l\left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right) < l\left(\tilde{\tilde{\mu}_1}, \tilde{\tilde{\mu}_2}, \tilde{\tilde{\mu}_3}\right) \end{cases} \tag{12}$$

$$= \begin{cases} \left(\tilde{\mu_1}, \tilde{\mu_2}, \tilde{\mu_3}\right), & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \geq \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}} \\ \left(\tilde{\tilde{\mu}_1}, \tilde{\tilde{\mu}_2}, \tilde{\tilde{\mu}_3}\right), & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} < \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}} \end{cases}. \tag{13}$$

Lemma 1 is thereby proved.

## A2. Proof of Theorem 2

Combining Lemma 1, Lemma 6, and Lemma 7 yields

$$\begin{aligned} l\left(\hat{\mu_1}, \hat{\mu_2}, \hat{\mu_3}\right) &= \begin{cases} -\frac{\left(\bar{x}_2 - \bar{x}_3\right)^2}{\sigma_2^2 + \sigma_3^2}, & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \geq \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}} \\ -\frac{\left(\bar{x}_1 - \bar{x}_2\right)^2}{\sigma_1^2 + \sigma_2^2}, & \text{if } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} < \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}} \end{cases} \\ &= -\min\left(\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \frac{|\bar{x}_2 - \bar{x}_3|}{\sqrt{\sigma_2^2 + \sigma_3^2}}\right) \end{aligned}$$

# Appendix B

In this appendix, we prove the results in Section 3.2.

## B1. Proof of Lemma 4 and Theorem 5

To show Lemma 4 and Theorem 5, we first show that $g(0,\infty) = g(\infty,0) = g(0,-\infty) = g(-\infty,0) = 1-\Phi(t)$. Recall that $T_1 = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$ and $T_2 = \frac{\bar{x}_2-\bar{x}_3}{\sqrt{\sigma_2^2+\sigma_3^2}}$. Notice that $T_1$ and $T_2$ have a bivariate normal distribution:

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} \frac{\lambda_1}{\sqrt{\sigma_1^2+\sigma_2^2}} \\ \frac{\lambda_2}{\sqrt{\sigma_2^2+\sigma_3^2}} \end{pmatrix}, \quad \begin{pmatrix} 1 & -\frac{\sigma_2^2}{\sqrt{\sigma_1^2+\sigma_2^2}\sqrt{\sigma_2^2+\sigma_3^2}} \\ -\frac{\sigma_2^2}{\sqrt{\sigma_1^2+\sigma_2^2}\sqrt{\sigma_2^2+\sigma_3^2}} & 1 \end{pmatrix} \right]$$

Without loss of generality (WLOG), we focus on the case when $\sigma_1^2+\sigma_2^2 = 1$ and $\sigma_2^2+\sigma_3^2 = 1$ for this proof, so

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}, \quad \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \right]$$

where $\rho = -\text{corr}(T_1, T_2)$ and $\rho > 0$.

We shift the center of the bivariate normal distribution to $(0,0)$, and let $Z_1 = T_1 - \lambda_1$ and $Z_2 = T_2 - \lambda_2$. Then $Z_1$ and $Z_2$ have a bivariate normal distribution

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \right]$$

We can now express $g(\lambda_1, \lambda_2)$ in terms of $Z_1$ and $Z_2$ as

$$\begin{aligned} g(\lambda_1,\lambda_2) =& P(T_1T_2 > 0,\, \min(|T_1|,|T_2|) > t) \\ =& P(T_1 > t, T_2 > t) + P(T_1 < -t, T_2 < -t) \\ =& P(Z_1 > t-\lambda_1, Z_2 > t-\lambda_2) + P(Z_1 < -t-\lambda_1, Z_2 < -t-\lambda_2) \end{aligned} \tag{14}$$

Hence, $g(0,\infty) = \lim_{\lambda_2\to\infty} P(Z_1 > t, Z_2 > t-\lambda_2) + \lim_{\lambda_2\to\infty} P(Z_1 < -t, Z_2 < -t-\lambda_2) = P(Z_1 > t) = 1-\Phi(t)$. Likewise, it can be shown that $g(\infty,0) = g(0,-\infty) = g(-\infty,0) = 1-\Phi(t)$.

To show Lemma 4 and Theorem 5, it remains to show that $g(\lambda_1,\lambda_2) \le P(Z > t)$ for $\lambda_1, \lambda_2 \in \Theta_0$, where $Z$ is a standard normal random variable. In equation (14), it is apparent that the magnitudes of $\lambda_1$ and $\lambda_2$

directly affect the value of $g(\lambda_1, \lambda_2)$. In fact, $\rho$ also plays a role in $g(\lambda_1, \lambda_2)$ as it influences the distribution of $(Z_1, Z_2)$. To highlight the dependence on $\rho$, for rest of the proof, we write $g_\rho(\lambda_1, \lambda_2)$. We introduce Lemma 8 which states that for all $\lambda_1, \lambda_2 \in \Theta_0$ and when $corr(T_1, T_2) \leq 0$ (i.e. $\rho < 0$), $g_\rho(\lambda_1, \lambda_2)$ is maximized when $\rho = 0$. With this lemma, it suffices to show that Lemma 4 holds when $\rho = 0$.

**Lemma 8.** $g_\rho(\lambda_1, \lambda_2) \leq g_{\rho=0}(\lambda_1, \lambda_2)$.

The proof of Lemma 8 will be provided in Appendix B2.

With this lemma, it suffices to show that Lemma 4 holds when $\rho = 0$. WLOG, we focus on the case when $(\lambda_1, \lambda_2) \in B$ and $|\lambda_1| > |\lambda_2|$. Let $a = -\lambda_2$. Then, $0 < a < \lambda_1$, and $g(\lambda_1, \lambda_2) = P(Z_1 > t - \lambda_1, Z_2 > t + a) + P(Z_1 > t + \lambda_1, Z_2 < -t + a)$. When $\rho = 0$, $Z_1$ and $Z_2$ are independent. Thus, $g_{\rho=0}(\lambda_1, \lambda_2) = P(Z_1 > t - \lambda_1)P(Z_2 > t + a) + P(Z_1 < -t - \lambda_1)P(Z_2 < -t + a)$. To show $g_{\rho=0}(\lambda_1, \lambda_2) \leq 1 - \Phi(t) = P(Z > t)$, we will discuss three cases depending on the magnitude of $t$: Case A. $0 \leq a \leq \frac{a+\lambda_1}{2} < t$, Case B. $0 \leq t \leq a \leq \lambda_1$, and Case C. $0 \leq a \leq t \leq \frac{a+\lambda_1}{2} \leq \lambda_1$.

*Case A* When $0 \leq a \leq \frac{a+\lambda_1}{2} < t$, we know that $t - \lambda_1 > -t + a$. Therefore, $P(Z_1 > t - \lambda_1) + P(Z_1 < -t + a) \leq 1$, and hence

$$\begin{aligned} g_{\rho=0}(\lambda_1, \lambda_2) \leq & P(Z_1 > t - \lambda_1)P(Z_2 > t) + P(Z_1 < -t)P(Z_2 < -t + a) \\ = & P(Z_1 > t - \lambda_1)P(Z > t) + P(Z > t)P(Z_2 < -t + a) \\ = & P(Z > t)\left[P(Z_1 > t - \lambda_1)P(Z > t) + P(Z > t)P(Z_2 < -t + a)\right] \leq P(Z > t) \end{aligned}$$

*Case B* When $0 \leq t \leq a \leq \lambda_1$, we will show $g_{\rho=0}(\lambda_1, \lambda_2) - [1 - \Phi(t)] = g_{\rho=0}(\lambda_1, \lambda_2) - P(Z > t) \leq 0$. We first note that

$$\begin{aligned} g_{\rho=0}(\lambda_1, \lambda_2) - P(Z > t) = & P(Z_1 > t - \lambda_1)P(Z_2 > t + a) + P(Z_1 < -t - \lambda_1)P(Z_2 < -t + a) - P(Z > t) \leq 0 \\ \iff & P(Z_1 > t - \lambda_1)P(Z_2 > t) - P(Z_1 > t - \lambda_1)P(t < Z_2 < t + a) \\ & - P(Z > t)P(Z_1 < t + \lambda_1) + P(Z_1 > t + \lambda_1)P(t - a < Z_2 < t) \leq 0 \end{aligned}$$

We write the left hand side of the last inequality as $I + II$, where

$$\begin{aligned} I =& P(Z_1 > t - \lambda_1)P(Z_2 > t) - P(Z > t)P(Z_1 < t + \lambda_1) \\ =& P(Z > t)\left[P(Z_1 > t - \lambda_1) - P(Z_1 < t + \lambda_1)\right], \\ II =& P(Z_1 > t + \lambda_1)P(t - a < Z_2 < t)) - P(Z_1 > t - \lambda_1)P(t < Z_2 < t + a) \end{aligned}$$

To show $I + II \leq 0$, it suffices to show $I \leq 0$ and $II \leq 0$. Since $t > 0$, it is obvious that $I \leq 0$. To show $II \leq 0$, it is equivalent to show that $\frac{P(t<Z_2<t+a)}{P(t-a<Z_2<t)} \frac{P(Z_1>t-\lambda_1)}{P(Z_1>t+\lambda_1)} \geq 1$. We let $III = \frac{P(t<Z_2<t+a)}{P(t-a<Z_2<t)} = \frac{P(t<Z<t+a)}{P(t-a<Z<t)}$ and $IV = \frac{P(Z_1>t-\lambda_1)}{P(Z_1>t+\lambda_1)} = \frac{P(Z>t-\lambda_1)}{P(Z>t+\lambda_1)}$. It suffices to show that $III \cdot IV \geq 1$. $IV$ obtains its minimum when $\lambda_1$ is minimized. Because $a \leq \lambda_1$, $IV \geq IV\Big|_{\lambda_1=a} = \frac{P(Z>t-a)}{P(Z>t+a)}$. Therefore,

$$\frac{P(t < Z_2 < t + a)}{P(t - a < Z_2 < t)} \frac{P(Z_1 > t - \lambda_1)}{P(Z_1 > t + \lambda_1)} \geq \frac{P(t < Z_2 < t + a)}{P(t - a < Z_2 < t)} \frac{P(Z > t - a)}{P(Z > t + a)}$$

Let $III = \frac{P(t<Z<t+a)}{P(t-a<Z<t)}$ and $V = \frac{P(Z>t-a)}{P(Z>t+a)}$. To show that $III \cdot V \geq 1$, we introduce Lemma 9 and Lemma 10.

**Lemma 9** (Abramowitz, 1965[12]). $\sqrt{\frac{\pi}{2}} e^{\frac{t^2}{2}} P(Z > t) \leq \frac{1}{t + \sqrt{t^2 + \frac{8}{\pi}}}$ *for* $t \geq 0$.

**Lemma 10.** $\frac{f(x+a)}{f(x)} \geq e^{-\frac{3a^2}{2}}$ *for* $t - a < x < t < a$.

The proof of Lemma 10 will be provided in Appendix B3.

Let $III = \frac{P(t<Z<t+a)}{P(t-a<Z<t)} = \frac{\int_t^{t+a} f(z)dz}{P(t-a<z<t)}$. With change of variable $x = z - a$ in the numerator, $III = \frac{\int_{t-a}^{t} f(x+a)dx}{P(t-a<z<t)}$. By Lemma 10, $III \geq \frac{\int_{t-a}^{t} e^{-\frac{3a^2}{2}} f(x)dx}{P(t-a<z<t)} = e^{-\frac{3a^2}{2}}$.

We know that $V = \frac{P(Z>t-a)}{P(Z>t+a)} \geq \frac{P(Z>0)}{P(Z>2a)} = \frac{1}{2P(Z>2a)}$. By Lemma 9, $V \geq \frac{1}{\sqrt{\frac{2}{\pi}} e^{-2a^2} \frac{2}{2a + \sqrt{4a^2 + \frac{8}{\pi}}}} = \sqrt{\frac{\pi}{2}} e^{2a^2} \left(a + \sqrt{a^2 + \frac{2}{\pi}}\right)$.

Therefore, $III \cdot V \geq e^{-\frac{3a^2}{2}} \sqrt{\frac{\pi}{2}} e^{2a^2} \left(a + \sqrt{a^2 + \frac{2}{\pi}}\right) = e^{\frac{a^2}{2}} \left(a + \sqrt{a^2 + \frac{2}{\pi}}\right) \sqrt{\frac{\pi}{2}} \geq 1$, when $a > 0$.

*Case C* When $0 \leq a \leq t \leq \frac{a+\lambda_1}{2} \leq \lambda_1$, we will show $I + II \leq 0$. To show $I + II \leq 0$, it's suffice to show that $III \cdot IV \geq 1$. Recall that $III = \frac{P(t<Z<t+a)}{P(t-a<Z<t)}$ and $IV = \frac{P(Z_1>t-\lambda_1)}{P(Z_1>t+\lambda_1)}$. For fixed $t$, $III$ obtains its minimum when $a$ is maximized. Given that $a \leq t$, $III \geq III\Big|_{a=t} = \frac{P(t<Z<2t)}{P(0<Z<t)}$. For fixed $a$ and $t$, $IV$ obtains its minimum when $\lambda_1$ is minimized. Given that the $\frac{a+\lambda}{2} > t$, $IV \geq IV\Big|_{\lambda_1=2t-a} = \frac{P(Z>a-t)}{P(Z>3t-a)}$. For fixed $t$, this ratio obtains its minimum when $a$ is maximized. Given that $a \leq t$, $IV \geq IV\Big|_{\lambda_1=2t-a, a=t} = \frac{P(Z>0)}{P(Z>2t)}$.

To show $III \cdot IV \geq \frac{P(t<Z<2t)}{P(0<Z<t)} \frac{P(Z>0)}{P(Z>2t)} \geq 1$, note that $\frac{P(t<Z<2t)}{P(0<Z<t)} \frac{P(Z>0)}{P(Z>2t)} = V \cdot III\Big|_{t=a}$ from the *Case B*. Hence, it follows that $\frac{P(t<Z<2t)}{P(0<Z<t)} \frac{P(Z>0)}{P(Z>2t)} \geq 1$.

To summarize, we have shown that for any $t > 0$ and $\lambda_1, \lambda_2 \in \Theta_0$, we have $g(\lambda_1, \lambda_2) < P(Z > t)$. This completes the proof of Lemma 4 and Theorem 5.

## B2. Proof of Lemma 8

To show $g(\lambda_1, \lambda_2, \rho) \leq g(\lambda_1, \lambda_2, 0)$, we will utilize Claims 1 and 2.

*Claim* 1. Suppose $W_1$ and $W_2$ are such that

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \right],$$

where $\rho > 0$. If $c > 0$, $d > 0$, then $P(W_2 > d | W_1 > c) \leq P(W_2 > d)$.

*Proof of Claim 1.* Given the joint distribution of $W_1$ and $W_2$ given above, we konw $W_2|W_1 \sim N(-\rho W_1, 1 - \rho^2)$. Thus, for $d > 0$, $P(W_2 > d) \geq P(W_2 > d | W_1 = w_1)$ if $w_1 > 0$. It follows that

$$\begin{aligned} P(W_2 > d | W_1 > c) =& \frac{\int_c^\infty P(W_2 > d | W_1 = w_1) f_{W_1}(w_1) dw_1}{P(W_1 > c)} \\ \leq& \frac{\int_c^\infty P(W_2 > d) f_{w_1}(w_1) dw_1}{P(W_1 > c)} \\ =& \frac{P(W_2 > d) P(W_1 > c)}{P(W_1 > c)} \\ =& P(W_2 > d) \end{aligned}$$

□

*Claim* 2. For $W_1$ and $W_2$ defined above, if $c > 0$ and $d > 0$, then $P(W_2 > d | W_1 < -c) \geq P(W_2 > d)$.

*Proof of Claim 2.*

$$\begin{aligned} P(W_2 > d | W_1 < -c) =& \frac{\int_{-\infty}^{-c} P(W_2 > d | W_1 = w_1) f_{W_1}(w_1) dw_1}{P(W_1 < -c)} \\ =& \frac{\int_{-\infty}^{-c} P(\sqrt{1-\rho^2} Z - \rho w_1 > d) f_{W_1}(w_1) dw_1}{P(W_1 < -c)}, \text{ where } Z \sim N(0,1) \text{ is independent of } W_1 \\ =& \frac{\int_{-\infty}^{-c} P(Z > \frac{d + \rho w_1}{\sqrt{1-\rho^2}}) f_{W_1}(w_1) dw_1}{P(W_1 < -c)} \end{aligned}$$

We first show the result for $c \geq d$. In this case, $W_1 \leq -c \implies W_1 \leq \frac{\sqrt{1-\rho^2}-1}{\rho} d \implies \frac{d+\rho W_1}{\sqrt{1-\rho^2}} \leq d \implies P\left(Z > \frac{d+\rho W_1}{\sqrt{1-\rho^2}}\right) \geq P(Z \geq d))$. So,

$$\begin{aligned} P(W_2 > d | W_1 < -c) \geq & \frac{\int_{-\infty}^{-c} P(Z > d) f_{W_1}(w_1) dw_1}{P(W_1 < -c)} \\ = & \frac{P(Z \geq d) P(W_1 \leq -c)}{P(W_1 < -c)} \\ = & P(Z \geq d) \\ = & P(W_2 \geq d) \end{aligned}$$

For $c < d$, it suffices to show,

$$\begin{aligned} & \frac{P(W_2 > d, W_1 < -c)}{P(W_1 < -c)} \geq P(W_2 > d) \\ \iff & P(W_2 > d, W_1 < -c) \geq P(W_1 < -c) P(W_2 > d) \\ \iff & \frac{P(W_2 > d, W_1 < -c)}{P(W_2 > d)} \geq P(W_1 < -c) \\ \iff & P(W_1 < -c | W_2 > d) \geq P(W_1 < -c) \end{aligned}$$

This is equivalent to the result with $c > d$. □

Now we go back to the proof of Lemma 8. We discuss three cases based on the magnitude of $t$ relative to $a$ and $\lambda_1$.

*Case 1 Assume* $0 \leq a \leq \lambda_1 \leq t$.

By Claim 1, we can show that $P(Z_1 > t - \lambda_1, Z_2 > t + a) \leq P(Z_1 > t - \lambda_1) P(Z_2 > t + a)$, and $P(Z_1 < t - \lambda_1, Z_2 < -t + a) \leq P(Z_1 > t + \lambda_1) P(Z_2 > t - a)$. Thus, $g(\lambda_1, \lambda_2, \rho) = P_\rho(Z_1 > t - \lambda_1, Z_2 > t + a) + P_\rho(Z_1 < t - \lambda_1, Z_2 < -t + a) \leq P(Z_1 > t - \lambda_1) P(Z_2 > t + a) + P(Z_1 > t + \lambda_1) P(Z_2 > t - a) \leq g(\lambda_1, \lambda_2, 0)$.

*Case 2 Assume* $0 \leq a \leq t < \lambda_1$.

Since $P(Z_1 > t - \lambda_1, Z_2 > t + a) = P(Z_2 > t + a) - P(Z_2 > t + a, Z_1 < t - \lambda_1)$. By Claim 2, $P(Z_1 > t - \lambda_1, Z_2 > t + a) \leq P(Z_2 > t + a) - P(Z_1 > t + a) P(Z_1 < t - \lambda_1) = P(Z_2 > t + a) P(Z_1 > t - \lambda_1)$. Thus, $P(Z_1 > t - \lambda_1, Z_2 > t + a) \leq P(Z_2 > t + a) - P(Z_2 > t + a) P(Z_1 > t - \lambda_1) P(Z_1 < t - \lambda_1) = P(Z_2 > t + a)$. By Claim 1, $P(Z_1 > t + \lambda_1, Z_2 > t - a) \leq P(Z_1 > t + \lambda_1) P(Z_2 > t - a)$. By an argument similar to Case 1, $g(\lambda_1, \lambda_2, \rho) \leq g(\lambda_1, \lambda_2, 0)$.

*Case 3 Assume* $0 \leq t < a \leq \lambda_1$.

Similar to Case 2, it can be shown that $P(Z_1 > t - \lambda_1, Z_2 > t + a) \leq P(Z_2 > t + a)P(Z_1 > t - \lambda_1)$. Using Claim 2, $P(Z_1 > t + \lambda_1, Z_2 > t - a) = P(Z_1 < -t - \lambda_1) - P(Z_1 < -t - \lambda_1, Z_2 > -t + a) \leq P(Z_1 < -t - \lambda_1) - P(Z_2 > -t + a)P(Z_1 < -t - \lambda_1) = P(Z_1 < -t - \lambda_1)P(Z_1 > t - a)$.

Overall, for any $t > 0$, we have shown that $g(\lambda_1, \lambda_2, \rho) \leq g(\lambda_1, \lambda_2, 0)$ holds for all $t > 0$. This completes the proof of Lemma 8. By an argument similar to Case 1, $g(\lambda_1, \lambda_2, \rho) \leq g(\lambda_1, \lambda_2, 0)$.

### B3. proof of Lemma 10

$$\frac{f(x+a)}{f(x)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x+a)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} = e^{-\frac{(a^2+2ax)}{2}} \geq e^{-\frac{(a^2+2ax)}{2}}\Big|_{x=t} = e^{-\frac{(a^2+2at)}{2}} \geq e^{-\frac{(a^2+2a^2)}{2}} = e^{-\frac{3a^2}{2}} \text{ for } t-a < x < t < a.$$

## Appendix C

In this appendix, we describe two naive procedures in addition to the naive procedure based on 2-sided TSTTs described in Section 2.

### C1. Naive procedure based on 1-sided TSTTs

To test for a strictly monotonic trend against the broad null, the two possible directions of the trend can be tested separately: the strictly increasing trend and the strictly decreasing trend. Each direction can be tested via a 1-sided TSTT between groups 1 and 2, and another such test between groups 2 and 3. For the increasing trend for example, we first test whether $\mu_1$ is significantly smaller than $\mu_2$, and likewise for $\mu_2$ and $\mu_3$. For the decreasing trend, a similar process can be followed. After four 1-sided TSTTs are conducted, we reject the null hypothesis in Equation (1) if for either direction both 1-sided TSTTs result in rejections of their corresponding nulls. Because we have multiple tests, the significance level for each test needs to be adjusted. Using Bonferroni correction, it can be shown that if we conduct each 1-sided TSTT at level $\frac{\alpha}{2}$, then the overall type 1 error rate is controlled at $\alpha$. Similar with 2-sided TSTTs procedure, the p-value is a linear transform from the greedy test. The p-value for 1-sided TSTTs procedure is $P_{1-sidedTSTT} = \max(p_1, p_2)$. This procedure is also laid out in Table 2.

### C2. Naive procedure based on confidence intervals

The test can be conducted using simultaneous 2-sided confidence intervals for the group means. The confidence interval for each of the sample means can be obtained based on the t-distribution with degrees of freedom equal to the associated sample size minus 1. The broad null hypothesis is then rejected if and only

| | |
|---|---|
| Step 1. Compare $\mu_1$ and $\mu_2$ | Test 1 (level $\alpha/2$):<br>$H_{01} : \mu_1 \geq \mu_2$<br>$H_{a1} : \mu_1 < \mu_2$ |
| Step 2. Compare $\mu_2$ and $\mu_3$ | Test 2 (level $\alpha/2$):<br>$H_{02} : \mu_2 \geq \mu_3$<br>$H_{a2} : \mu_2 < \mu_3$ |
| Step 3. Compare $\mu_1$ and $\mu_2$ | Test 3 (level $\alpha/2$):<br>$H_{03} : \mu_1 \leq \mu_2$<br>$H_{a3} : \mu_1 > \mu_2$ |
| Step 4. Compare $\mu_1$ and $\mu_2$ | Test 4 (level $\alpha/2$):<br>$H_{04} : \mu_2 \leq \mu_3$<br>$H_{a4} : \mu_2 > \mu_3$ |
| Reject the overall $H_0$ if one of the following happens:<br>(i) both tests 1 and 2 yield a rejection, or<br>(ii) both tests 3 and 4 yield a rejection | |

Table 3: Naive procedure based on 1-sided TSTTs

if none of the confidence intervals overlap and the sample means have a monotonic trend. To yield an overall confidence level of $1 - \alpha$, the Bonferroni method can be used to form the confidence interval for each mean at level $1 - \frac{\alpha}{3}$. It can be shown that this testing procedure guarantees an overall type 1 error control at $\alpha$.