The Parametric t-test's Latent Weakness

Daniel P. Gaile, Jeffrey C. Miecznikowski *

03/06/2015

Abstract

When a latent class structure is present, parametric t-tests conducted on the observed continuous variable data can be anti-conservative. This problem is exacerbated by: A) test multiplicity across large numbers of manifest assays, each with a plausible latent structure, and B) increased accuracy of the manifest assays to discriminate underlying latent structures. This result is relevant in many modern experimental settings where an underlying latent structure is either known to be true (e.g., methylation and array CGH) or plausible (e.g., down/up-regulated gene networks). In such settings, the discrepancy between the actual null distribution and the commonly assumed null t-distribution can lead to a breakdown in the type I error control and gross overstatements of statistical significance.

KEY WORDS: parametric t-test, Student's t-test, small sample, latent structure, mixture model, Fisher's Exact Test

*Daniel P. Gaile is Assistant Professor, Jeffrey C. Miecznikowski is Associate Professor, University at Buffalo, Department of Biostatistics, 706 Kimball Tower, 3435 Main St., Buffalo, NY 14214-3000.

1 Introduction

We consider the analysis, with a parametric two-sample t-test, of data of the type presented in Figure 1. Namely, we consider the two sample comparison of groups with small sample size (e.g. $n \approx 4$). Small samples sizes of $n = 3, 4, 5$, or 6 per group are common in pilot studies involving cell lines, mouse or rat lines, and/or high throughput biotechnologies such as microarrays and next generation sequencing platforms. We further restrict our interest to cases in which it is plausible to assume that the observed assay values are manifest observations governed by an underlying latent state.

We suppose that the observed assay values are gathered with the hopes of testing the following biological hypotheses:

$$\begin{aligned} H_0 &: \text{No between group difference in the latent biological condition} \\ H_A &: \text{A between group difference exists in the latent biological condition} \end{aligned} \tag{1.1}$$

where, for example, the latent biological condition could be methylation status, copy number state, or gene expression state (e.g., up- versus down-regulated).

In the case of a methylation study, the manifest assay values could be the β or M values from high density arrays (Bock, 2012) and the latent structure would be the actual methylation state of the targeted genomic region for each sample. Figure 1(a) depicts level 3 (i.e., β) assay values for probe cg24881834 as assayed for The Cancer Genome Atlas (i.e., http://cancer genome.nih.gov/) using 27k methylation arrays. Data for eight Gliobalstoma multiforme samples are included with four assigned to Group A (i.e., TCGA-06-2565 TCGA-06-2566, TCGA-27-1836, and TCGA-32-1982), and the balance assigned to Group B (i.e., TCGA-02-2483, TCGA-27-1835, TCGA-27-2523, and TCGA-27-2524). It is reasonable to assume that β values greater than 0.75 indicate that the assay region is methylated and β values less than 0.025 indicate that the assay region in un-methylated. The underlying latent structure of methylation data of this type is well

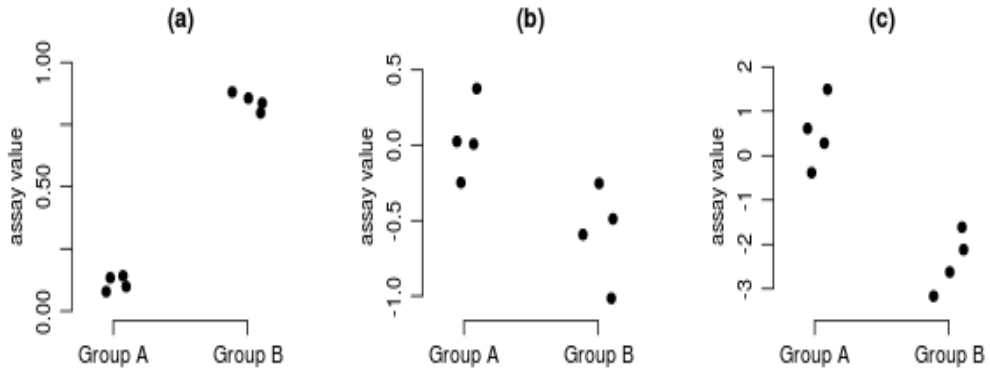

Figure 1: An illustrative example from The Cancer Genome Atlas (TCGA). Data for eight samples was segregated into two groups: Group A (i.e., TCGA-06-2565 TCGA-06-2566, TCGA-27-1836, and TCGA-32-1982), and Group B (i.e., TCGA-02-2483, TCGA-27-1835, TCGA-27-2523, and TCGA-27-2524). Plotted points correspond to assay values (y-coordinate) and are segregated according to group membership (x-coordinate). (a) Level 3 methylation values for probe cg24881834 as assayed using 27k methylation arrays. (b) Level 2 array comparative genomic hybridization values for probe A_16_P17612086 as assayed using Agilent Human Genome CGH Microarray 244A arrays. (c) Level 3 expression values for ME1 as assayed using Agilent 244K Custom Gene Expression G4502A-07 arrays.

accepted and has been modeled with a mixture of beta distributions (Teschendorff et al., 2013; Laurila et al., 2011). The assignment of the eight samples to groups was arbitrary and probe cg24881834 was selected due to the observed distribution of β values, as well as the availability of matched expression array and array CGH assay values which also suggest a latent structure.

In the case of a copy number study, the manifest assay values could be the $-\log_2$ Tumor/Control values as quantified by micro-arrays (Snijders et al., 2001) and the latent structure would be the underlying change in copy number status as the targeted genomic region for each sample. Figure 1(b) depicts level 2 array comparative genomic hybridization (aCGH) values (i.e., normalized signals for copy number alterations) for probe A_16_P17612086 as assayed using Agilent Human Genome CGH Microarray 244A arrays. The values for the subjects in Group A suggest that they all have normal copy number. The values for the subjects in Group B, save perhaps the sample with the highest value, suggest a copy number loss.

In the case of a gene expression study, the manifest variable could be the expression

values as quantified by micro-array (Schena et al., 1995) or RNA-seq technologies (Chu and Corey, 2012) and the underlying latent structure could be whether or not the expression levels of the gene of interest are down-regulated or not. Figure 1(c) depicts level 3 normalized expression values for ME1 as assayed using Agilent 244K Custom Gene Expression G4502A-07 arrays. The data suggests that ME1 is down-regulated in Group B compared to Group A. The expression values may be controlled by the combination of the underlying methylation and aCGH states for cg24881834 and A_16_P17612086, respectively, as both have genomic locations which suggest functional relationships to ME1. Namely, ME1 expression in group B could be suppressed due to either copy number loss, methylation events, or both, within the targeted region.

1.1 A Simple Two Sample Two Latent State Model

We consider the analysis of data generated under a two sample two latent state model. Specifically, we consider a set of assay values measured upon samples from two populations. We define

$$W_{ij} \sim Bernoulli(\pi_i), \quad i = 1, 2 \; ; \; j = 1, 2, 3, n \tag{1.2}$$

where W_{ij} is the latent state of the j^{th} observation from the i^{th} sample population, and π_i is the population proportion of the second latent state in the i^{th} sample population. We define the manifest variable (i.e., the observed assay values) to be conditionally distributed as:

$$X_{ij}|W_{ij} = w_{ij} \sim Normal(\tau_1(1 - w_{ij}) + \tau_2 w_{ij}, (\sigma_1^2(1 - w_{ij}) + \sigma_2^2 w_{ij})) \tag{1.3}$$

where τ_1 and τ_2 denote the populations means for the first and second latent states, respectively. Similarly, σ_1 and σ_2 denote the population standard deviations. For ease of exposition, we consider the manifest variable data to have been standardized such that $\tau_1 = 0$, $\tau_2 = \tau$, and $\sigma_1^2 = \sigma_2^2 = 1$. For the balance of this manuscript, we specify a simple

two sample two latent state model to be:

$$X_{ij}|W_{ij} = w_{ij} \sim Normal(\tau w_{ij}, 1) \;\; ; \;\; W_{ij} \sim Bernoulli(\pi_i) \tag{1.4}$$

for $i = 1, 2$ and $j = 1, 2, 3, n$. Under this model, τ represents the difference in population means between latent states, measured in unit standard deviations. The model presented in equation (1.4) constitutes a simple two sample Gaussian mixture model of a general type described in (McLachlan and Peel, 2000).

1.2 Surrogate Hypothesis Tests

The biological hypotheses mentioned above can be tested via the formal test of the following hypotheses:

$$\begin{aligned} H_0 &: \text{No difference in manifest assay means} \\ H_A &: \text{Difference in manifest assay means .} \end{aligned}$$

These hypotheses can be tested using an equal variance parametric t-test statistic (Gosset, 1908; Fisher, 1925):

$$T = \frac{\sqrt{n}(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_1^2 + S_2^2}} \tag{1.5}$$

where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ and $S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$. The parametric t-test of the set of assay values tests the hypotheses:

$$\begin{aligned} H_0 &: E[\bar{X}_1] = \pi_1\tau = \pi_2\tau = E[\bar{X}_2] \\ H_A &: E[\bar{X}_1] = \pi_1\tau \neq \pi_2\tau = E[\bar{X}_2] \end{aligned}$$

which is clearly equivalent to:

$$\begin{aligned} H_0 &: \pi_1 = \pi_2 \\ H_A &: \pi_1 \neq \pi_2 \ . \end{aligned} \tag{1.6}$$

Another possibility would be to test the latent state frequencies, and hence the test of hypotheses in (1.6), more directly. In such a case, the latent state counts must be estimated (e.g., with a cut-off rule), with the possibility of committing classifications errors. While there are numerous test statistics that could be utilized in this case, we consider the popular Fisher's Exact Test (Fisher, 1934). The Fisher's Exact Test is included briefly in our subsequent analyses in order to contrast the performance of a test that acknowledges the presence of an underlying structure compared to the parametric t-test, which does not. This comparison of approaches is of secondary importance to the primary focus of this manuscript, which is to prove the inadequacy of the parametric t-distribution as a null distribution under certain experimental conditions.

Both the parametric t-test and Fisher's Exact Test appear to test equivalent hypotheses, only with different data (i.e., continuous manifest variable observations and estimated categorical latent state counts) and with different null distributions (i.e., the t-distribution and the hyper-geometric distribution). An argument for the preferential use of the t-test can be made based upon the conventional wisdom that dichotomizing continuous data reduces power. Additionally, the parametric t-test, being a continuous measure, appears to have the advantage of finer resolution of evidence against then null distribution when compared to the discrete valued Fisher's Exact Test. As a practical matter, in some cases, such as when a Bonferroni corrected testing level is employed, the smallest attainable Fisher's Exact Test p-value might exceed the proposed per-comparison test level making rejection of the null hypothesis impossible. By way of contrast, the parametric t-test can deliver arbitrarily small p-values, even for small

sample sizes.

In this manuscript, we make the argument that utilizing the parametric t-test to analyze continuous manifest data governed by a latent structure can lead to an inflation of power with a loss of type I error control. We demonstrate that the parametric t-test does not follow a t-distribution under a latent state null model for which latent state frequencies are set equal. While it is true that the biological null hypothesis of "no difference in assayed biological condition" might map to the null hypothesis of "no difference in manifest assay means", the parametric t-test quantification of that evidence can be wildly inaccurate when a t-distribution is used to model the null distribution. If a latent structure is plausible, then we advocate the use of testing methods which either explicitly (e.g., Fisher's Exact Test) or implicitly (e.g., permutation/exact t-test) account for it.

In Section 2, we provide a simple motivating example data-set. We demonstrate that the p-values obtained via the parametric t-test and Fisher's Exact Test can differ to a large degree. Additionally, we provide conditions where the parametric t-test p-values are anti-conservative with a corresponding loss of type I error control. In Section 3, we derive distributional properties of the sample means and variances under a simple one and two sample two latent state models. In Section 4, we provide simulation results that demonstrate that the t-distribution can be an inadequate null distribution when a latent structure is present. We illustrate how our simulated empirical null distribution differs from a t-distribution and demonstrate how those results are a consistent with theoretical results presented in Section 3. In Section 5, we conclude the manuscript with a discussion of our findings.

2 A Motivating Example Data-set

We consider an example data-set simulated under the simple two sample two latent state model given by equation 1.4 and with $n = 4$, $\pi_1 = \pi_2 = \frac{1}{2}$, and $\tau = 5$. We label

the two sample populations "Group A" and "Group B", corresponding to $i = 1$ and $i = 2$, respectively. We consider the rare outcome: $w_{11} = w_{12} = w_{13} = w_{14} = 0$ and $w_{21} = w_{22} = w_{23} = w_{24} = 0$, which occurs with a probability $(1 - \frac{1}{2})^4(\frac{1}{2})^4 = 0.00390625$ and a corresponding odds of 255:1. Conditioned on those particular latent states, the deviations from the group means (i.e., $X_{ij} - w_{ij}\tau$), were simulated from a standard normal distribution and then rounded to two decimal places (for convenience of exposition). Group A was simulated to have deviations of $\{0.59, 0.71, -0.11, -0.45\}$ and Group B was simulated to have deviations of $\{0.61, -1.82, 0.63, -0.28\}$. The corresponding two sample data is given as: $x_{11} = 0.59, x_{12} = 0.71, x_{13} = -0.11, x_{14} = -0.45, x_{21} = 0.61, x_{22} = 3.18, x_{23} = 5.63, x_{24} = 4.72$. Plotted points in Figure 2(a) denote the observed assay values (y-coordinate) and sample group memberships (x-coordinate, labeled "Group A" and "Group B") for our motivating example data-set.

We are interested in the extent to which the observed data can be considered as evidence against the null hypothesis of equal latent population proportions across Groups A and B. To that end, an equal variance parametric two-sided t-test of the null hypothesis of equal Group A and Group B means, provides a test statistic of -7.195116 and a t_6-distribution p-value of 0.0003647. The t_6-distribution derived odds against the null hypothesis (quantified as $\frac{1-(\text{p-value})}{\text{p-value}}$) is approximately 2741:1.

Suppose that the horizontal dashed line depicted in Figure 2(b) constitutes a well accepted boundary for estimating a sample's latent state status given its observed assay value. In such a case, the observed data could be recast in the form of a two by two table: with the four observations for Group A classified as "state 1" and the four observations for Group B classified as "state 2". When applied to the 2 by 2 table data depicted in Figure 2(b), the Fisher's Exact Test provides a p-value of 0.029 and an odds against the null hypothesis of 34:1. The ratio of the estimated odds against the null hypothesis for the t-test compared to the Fisher's Exact Test is given by: $\frac{2741}{34} = 80.62$. Which is to say, the parametric t-test quantifies the odds against the null hypothesis to be 80 times greater than the Fisher's Exact Test quantification.

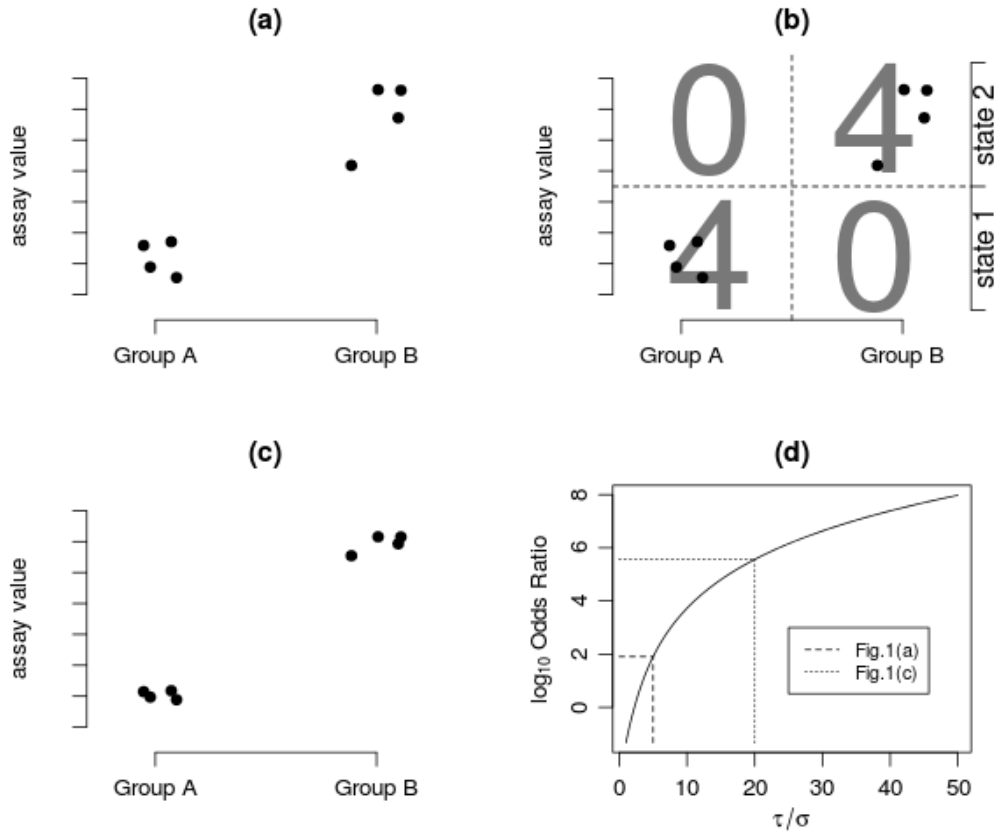

Figure 2: Motivating Example Data-set. (a) Plotted points correspond to assay values (y-coordinate) and are segregated according to group membership (x-coordinate). (b) The horizontal dashed line demarcates a well accepted boundary for the estimation of the latent state conditioned on the observed assay values. Applying the proposed cut-off rule yields an estimated count of 4 latent state 1 observations for Group A and 4 latent state 2 observations for Group B; counts which correspond to the true conditions under which the data was simulated. (c) The assay values in (a) measured with greater accuracy, i.e., deviations multiplied by a factor of 0.25. (d) The $\log_{10}$ odds ratio for the odds against the null based upon the parametric t-test to the odds based upon the Fisher's exact test of the true latent states as function of the accuracy of the assays. Simulated deviations were multiplied by σ and then added to their respective means of 0 and $\tau = 5$, for latent groups 1 and 2, respectively. The dashed line corresponds to the result in (a) where $\tau/\sigma = 5/1 = 1$ and a $\log_{10}$ odds ratio of $\log_{10}((2741:1)/(34:1)) = \log_{10} 80.59 = 1.9$ The dotted line corresponds to the result in (b) where $\tau/\sigma = 5/0.25 = 20$ and a $\log_{10}$ odds ratio of $\log_{10}\left((1.25 \times 10^7 : 1)/(34:1)\right) = \log_{10} 3.68 \times 10^5 = 5.6$.

If the experimental data in Figure 2(a) is measured with greater accuracy, the differences in the t-test and Fisher's Exact Test p-values can become more profound. For example, suppose that the simulated deviations were generated from a $N(0, 0.25^2)$ distribution rather than a standard normal. Multiplying the observed deviations by 0.25 and adding them to their respective population means yields the data depicted in Figure 2(c). The Fisher's exact test for this data provides identical results as for Figure 2(b).

However, the parametric t-test provides a p-value of 8.00×10^{-08} and an odds against the null hypothesis of 12,507,415:1. The ratio of the estimated odds against the null hypothesis for the t-test compared to the Fisher's Exact Test for the data depicted in Figure 2(c) is $\frac{12,507,415}{34} = 367865.1$. Which is to say, the parametric t-test quantifies the odds against the null hypothesis to be greater than 367 thousand times the Fisher's Exact Test quantification.

As the manifest variable becomes a more accurate discriminant of the latent states, the t-test results do not converge to the Fisher's Exact Test result. Figure 2(d) provides a plot of the $\log_{10}$ odds ratio for the odds against the null based upon the parametric t-test to the odds based upon the Fisher's exact test of the true latent states. The $\log_{10}$ odds ratio is plotted as a function of the accuracy of the assays and clearly increases with improved accuracy, becoming much larger than 0. The figure was generated as follows: simulated deviations were multiplied by σ and then added to their respective means of 0 and $\tau = 5$, for latent states 1 and 2, respectively. The dashed line in Figure 2(d) corresponds to the result in Figure 2(a) where $\tau/\sigma = 5/1 = 1$ and a $\log_{10}$ odds ratio of $\log_{10}((2741:1)/(34:1)) = \log_{10} 80.59 = 1.9$ The dotted line corresponds to the result in Figure 2(b) where $\tau/\sigma = 5/0.25 = 20$ and a $\log_{10}$ odds ratio of $\log_{10}((1.25 \times 10^7 : 1)/(34:1)) = \log_{10} 3.68 \times 10^5 = 5.6$.

If the parametric t-test is used on the pilot data in Figure 2(c) then subsequent verification/validation studies may yield grossly deflated results (as compared to the t-test quantification). For example, suppose that the simulated data in Figure 2(c) was actual experimental data and that the investigators that generated it were sufficiently encouraged by the results to perform a gold standard assay on the samples. Further suppose that the gold standard assay definitively classified (i.e., eliminating the possibility of mis-classifications errors) the observations into their true latent State categories (i.e., the gold standard assay concluded all 4 observations in Group A were in latent State 1 and all 4 observations in Group B were in latent State 2). With that new gold standard information in hand, the Fisher's Exact Test (or any other table based testing approach)

would clearly be more appropriate than the t-test applied to the original manifest variable data. After all, who would choose to test the variable manifest assay data when the definitive latent state data is in hand? Note that the gold standard results that we assume are the most favorable possible towards supporting the decision to reject the null hypothesis. Yet under that most favorable development, the investigators would see the strength of their experimental evidence diminish from 12.5 million to 1 (i.e., based upon the t-test of original data) down to 34 to 1 (i.e., based upon the Fisher's Exact Test of Gold Standard Verification of sample latent states).

It should be noted that permutation t-test p-values were in close agreement with the Fisher's Exact Test p-values, and also estimated odds against the null hypothesis as 34:1 for both example data-sets (i.e., data depicted in Figures 2(a) and (c)). The permutation of group labels with respect to the manifest variable outcomes will also permute those labels with respect to the underlying latent structure. Hence, the permutation t-test null distribution should provide an adequate approximation to the true null distribution (assuming, of course, that exchangeablity conditions are truly satisfied).

2.1 Error Control Can Be Lost

In later sections we prove that the equal variance two sample parametric t-statistic is not distributed according to a t-distribution under our assumed null model conditions (i.e., when $\pi_{h1} = \pi_{h2} = \pi$, with $\pi > 0$ and $\tau > 0$). In this section, we prove that using the t-distribution as a null distribution can lead to loss of even conservative Bonferronni family-wise-error-rate (FWER) control.

We extend our simple two sample two latent state model to model a collection of H different assays that are measured across n samples from each of two sample population groups. We model the observed manifest assay values as:

$$X_{hij}|W_{hij} = w_{hij} \sim Normal(\tau w_{hij}, 1) \;\; ; \;\; W_{hij} \sim Bernoulli(\pi) \tag{2.7}$$

for $h = 1, \ldots, H$; $i = 1, 2$; $j = 1, 2, 3, n$ and where π and τ are as defined previously. Note that this model is a true null model as π is identical for both sample groups. In real settings, the π and τ values would likely vary across the assays, and would be better parameterized as π_h and τ_h. For purposes of identifying the boundaries for which error control is lost, we simplify the model and assume that τ and π do not vary across the assays. The boundaries we provide are symmetric about $\pi = \frac{1}{2}$ and we demonstrate that error control is lost with increasing τ. Hence, the results we provide can be extended to the case where π and τ values vary across the assays. For example, if error control is lost for a set of values H^*, n^*, π^*, and τ^* then it can be said that error control would be lost for a collection of H^* tests with equal sample size n^* as long as all $\tau_h \geq \tau^*$ and all $|\pi_h - \frac{1}{2}| \leq |\pi^* - \frac{1}{2}|$.

Theorem 2.1 *Consider the simple two sample two latent state model given by equation 2.7 and the set $h = 1, \ldots, H$ of one-sided tests of hypotheses: $H_{h0} : \mu_{h1} \geq \mu_{h2}$ versus $H_{h1} : \mu_{h1} < \mu_{h2}$, where $\mu_{h1} = \pi_{h1}\tau$ and $\mu_{h2} = \pi_{h2}\tau$ with $\tau > 0$. Let the corresponding members of the set of equal variance parametric t-test statistics be given by:*

$$T_h = \frac{\sqrt{n}(\bar{X}_{h1} - \bar{X}_{h2})}{\sqrt{S_{h1}^2 + S_{h2}^2}} \tag{2.8}$$

for $h = 1, \ldots, H$, and where $\bar{X}_{hi} = \frac{1}{n}\sum_{j=1}^{n} X_{hij}$ and $S_{hi}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{hij} - \bar{X}_{hi})^2$. Assume a global null condition of $\pi_{h1} = \pi_{h2} = \pi$ for all $h = 1, \ldots, H$. For a given sample size n, latent state 2 population proportion π, and a specified family-wise error rate α, there exists combinations of H and τ such that the family-wise-error-rate (FWER) is not controlled using the standard Bonferroni Bounds applied to the collection of H one-sided equal variance parametric t-tests given in equation 2.8.

Proof Let $\mathbf{W_h}$ denote the vector of latent states, $[W_{h11}, W_{h12}, \ldots, W_{h1n}, W_{h21}, W_{h22}, \ldots, W_{h2n}]^T$, for the h^{th} set of assays. Let $\mathbf{W_h} = w_{0/n}$ denote shorthand notation for the outcome that all of the observations for sample group 1 had underlying latent state 1 and all of the observations for sample group 2 had underlying latent state 2, i.e., $\mathbf{W_h} = [W_{h11} =$

$0, W_{h12} = 0, \ldots, W_{h1n} = 0, W_{h21} = 1, W_{h22} = 1, \ldots, W_{h2n} = 1]^T$. Under the specified model and a global null condition of $\pi_{h1} = \pi_{h2} = \pi$ for all $h = 1, \ldots, H$,

$$P(\mathbf{W_h} = w_{0/n}) = (1-\pi)^n \pi^n \ .$$

The probability that at least one $\mathbf{W_h} = w_{0/n}$ for $h = 1, \ldots, H$, is equal to:

$$P(\text{at least one } \mathbf{W_h} = w_{0/n}) = 1 - (1 - (1-\pi)^n \pi^n)^H \ .$$

Conditioned on the latent state $\mathbf{W}_h = w_{0/n}$, T_h in (2.8) is distributed as a non-central t-distribution with $2n-2$ degrees of freedom and non-centrality parameter equal to $-\tau\sqrt{2}$. The probability of observing a T_h statistic less than a specified cut-off, t_{cut}, is

$$P(T_h \leq t_{cut}) = F_{T_{\nu=2n-2, ncp=-\tau\sqrt{2}}}(t_{cut})$$

where $F_{T_{\nu=2n-2, ncp=-\tau\sqrt{2}}}$ is the cumulative distribution function of a non-central t-distribution with $2n-2$ degrees of freedom and non-centrality parameter $ncp = -\tau\sqrt{2}$. We note that

$$\begin{aligned} P(\text{at least} \quad & \text{one } T_h < t_{cut} | \pi, \tau) \\ & \geq \quad P(T_h < t_{cut} | \pi, \tau, \mathbf{W}_h = w_{0/n}) P(\text{at least one } \mathbf{W}_h = w_{0/n}) \\ & = \quad F_{T_{2n-2, -\tau\sqrt{2}}}(t_{cut}) \left(1 - (1 - (1-\pi)^n \pi^n)^H\right) \end{aligned}$$

Suppose that t_{cut} is selected according to a Bonferroni bound for FWER control:

$$F_{T_{2n-2}}(t_{cut}) \leq \frac{\alpha}{H}$$

and, hence,

$$t_{cut} \leq Q_{T_{\nu=2n-2}}\left(\frac{\alpha}{H}\right) \ .$$

where $Q_{T_{\nu=2n-2}}$ is the quantile function of a t-distribution with $2n-2$ degrees of freedom.

FWER control is lost when

$$\begin{aligned} P(\text{at least one } T_h < t_{cut}|\pi,\tau) &> \alpha \\ F_{T_{2n-2,-\tau\sqrt{2}}}(t_{cut})\left(1-(1-(1-\pi)^n\pi^n)^H\right) &> \alpha \\ F_{T_{2n-2,-\tau\sqrt{2}}}\left(Q_{T_{2n-2}}\left(\frac{\alpha}{H}\right)\right)\left(1-(1-(1-\pi)^n\pi^n)^H\right) &> \alpha \end{aligned}$$

which leads to

$$F_{T_{2n-2,-\tau\sqrt{2}}}\left(Q_{T_{2n-2}}\left(\frac{\alpha}{H}\right)\right) > \frac{\alpha}{(1-(1-(1-\pi)^n\pi^n)^H)} \ . \tag{2.9}$$

If we select

$$H > \frac{\log(1-\alpha)}{\log(1-(\pi(1-\pi))^n)} \tag{2.10}$$

then

$$1 > \frac{\alpha}{(1-(1-(1-\pi)^n\pi^n)^H)} \ .$$

We note that, conditioned on n, α, and H, the left side of the inequality in equation 2.9 is a strictly increasing function in τ. Hence, for any H satisfying equation 2.10, a τ can be selected such that

$$1 > F_{T_{2n-2,-\tau\sqrt{2}}}\left(Q_{T_{2n-2}}\left(\frac{\alpha}{H}\right)\right) > \frac{\alpha}{(1-(1-(1-\pi)^n\pi^n)^H)} \ .$$

Corollary 2.2 *Consider the simple two sample two latent state model given by equation 2.7 and the set, $h = 1, \ldots, H$ of two-sided test of hypotheses: $H_{h0} : \mu_{h1} = \mu_{h2}$ versus $H_{h1} : \mu_{h1} \neq \mu_{h2}$, where $\mu_{h1} = \pi_{h1}\tau$ and $\mu_{h2} = \pi_{h2}\tau$ with $\tau > 0$. Under the conditions specified in Theorem 2.1, for a given sample size n, latent state 2 population proportion π, and a specified family-wise error rate α, there exists combinations of H and τ such that the family-wise-error-rate (FWER) is not controlled using the standard Bonferroni Bounds applied to the collection of H equal variance parametric t-tests given*

in equation 2.8

Proof Similar to the proof of Theorem 2.1 except that FWER control is lost when

$$F_{T_{2n-2,-\tau\sqrt{2}}}\left(Q_{T_{2n-2}}\left(\frac{\alpha}{2H}\right)\right) > \frac{\alpha/2}{(1-(1-(1-\pi)^n\pi^n)^H)} \tag{2.11}$$

2.1.1 Example Conditions

The results of Theorem 2.1 can be applied to the conditions relevant to the motivating example data-set. Specifically, consider data generated under the simple two sample two latent state null model with $n = 4$, $H = 1,000$, $\pi_1 = \pi_2 = \frac{1}{2}$ and $\tau = 5$. Under such a model, the probability that at least one of the H assay sets will have the latent state $\mathbf{W}_h = w_{0/4}$ is $1 - (1 - (\frac{1}{2})^8)^{1000} = 0.98$. If the tests are conducted with a Bonferroni FWER of 0.05, the probability that an assay with latent state $\mathbf{W}_h = w_{0/4}$ will yield an observed test statistic less than $Q_{t_6}\left(\frac{0.05}{1000}\right) = -9.0823$, is $F_{t_{6,-5\sqrt{2}}}(-9.0823) = 0.2866$. Hence, the probability of committing a type I error is at least:

$$\begin{aligned} P(\text{at least one } t_h < t_{cut}) &= F_{t_{6,-5\sqrt{2}}}\left(Q_{t_6}\left(\frac{0.05}{1000}\right)\right)\left(1-(1-(1-\frac{1}{2})^4\frac{1}{2}^4)^{1000}\right) \\ &= 0.2866 \times 0.9800 \\ &= 0.2809 \end{aligned}$$

Which is to say, that for a set of $H = 1000$ tests generated under model conditions $\tau = 5$ and $\pi_1 = \pi_2 = \frac{1}{2}$, the probability of at least one t-test statistic falling within the .05 family-wise level Bonferroni rejection region is greater than or equal to 0.2809; a clear loss of FWER control.

2.1.2 Estimated Bounds for the Loss of Error Control

If it is possible to lose error control when the parametric t-distribution is used as a null in the presence of a latent structure, then under what conditions might such a problem

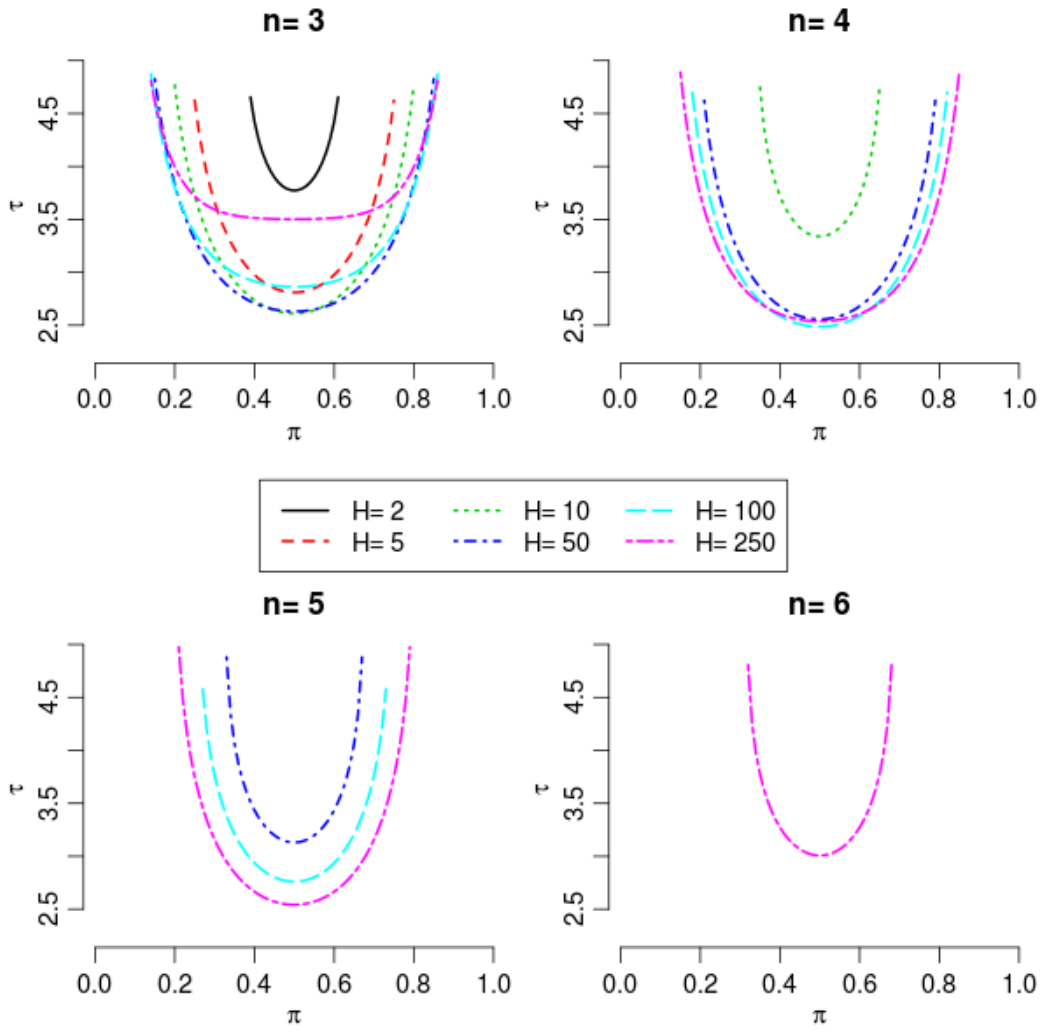

Figure 3: Estimated bounds for which error control is lost under the two-sided testing conditions of Corollary 2.2. For a given combination of n, H, and π, the curves provide an upper-bound for the lowest value of τ for which a breakdown of Bonferroni FWER control will occur. Although the bounds are conservative, they demonstrate that the loss of error control can occur under reasonable conditions with respect to sample size, effect size, and underlying latent structure proportions.

occur? Figure 3 provides estimates of the bounds for which error control might be lost under the two-sided testing conditions of Corollary 2.2. For $\alpha = 0.05$ and combinations of $n = 3, 4, 5, 6$, $H = 2, 5, 10, 50, 100, 250$, curves are provided for the condition:

$$F_{T_{2n-2,-\tau\sqrt{2}}}\left(Q_{T_{2n-2}}\left(\frac{\alpha}{2H}\right)\right) = \frac{\alpha/2}{(1-(1-(1-\pi)^n\pi^n)^H)}.$$

Hence, for a given combination of n, H, and π, the curves provide an upper-bound for the lowest value of τ for which a breakdown of Bonferroni FWER control will occur. The estimates of the upper bounds are increasingly conservative as H increases. Hence, inversions of their expected orderings can occur, as in Figure 3(a). Although the bounds are conservative, they demonstrate that the loss of error control can occur under reasonable conditions with respect to sample size, effect size, and underlying latent structure proportions.

3 Distributional Properties of $\bar{X}$ and S^2

3.1 Under A Simple One Sample Two Latent State Model

Before considering the two sample t-test, we first derive the distributional properties of the sample mean, $\bar{X} = \sum_{j=1}^{n} X_j$, and sample variance, $S_X^2 = \sum_{j=1}^{n}(X_j - \bar{X})^2$, in a single sample setting. Consider n observations, $X_1, \ldots, X_n$ a made on the manifest variables governed by a simple one sample, two latent state model:

$$X_j | W_j = w_j \sim Normal(\tau w_j, 1) \;\; ; \;\; W_j \sim Bernoulli(\pi) \tag{3.12}$$

for $j = 1, 2, 3, n$.

Theorem 3.1 (One Sample Case) *Let* $W_\cdot = \sum_{j=1}^{n} W_j$. *Under the conditions of equation 3.12 and conditioned on* $W_\cdot = w_\cdot$*, the distribution of the sample mean is:*

$$\bar{X} | W_\cdot = w_\cdot \sim N\left(\frac{w_\cdot \tau}{n}, \frac{1}{n}\right) \tag{3.13}$$

Corollary 3.2 (Two Sample Case) *Consider the simple two sample two latent state model given by equation 1.4 for* $H = 1$ *(i.e., suspend the h subscript. Let* $\mathbf{W}$ *denote the vector of latent states,* $[W_{11}, W_{12}, \ldots, W_{1n}, W_{21}, W_{22}, \ldots, W_{2n}]^T$*. Let* $\mathbf{W} = w_{p/q}$ *denote shorthand notation for the outcome that p of the n observations for sample group 1 have*

underlying latent state 1 and q of the n observations for sample group 2 had underlying latent state 2. Let $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, *for* $i = 1, 2$. *Conditioned on* $\mathbf{W} = w_{p/q}$, *the distribution of the difference in sample means is:*

$$\bar{X}_1 - \bar{X}_2 | \mathbf{W}_\cdot = w_{p/q} \sim N(\frac{(p-q)\tau}{n}, \frac{1}{n}) \tag{3.14}$$

Theorem 3.3 (One Sample Case) *Under the conditions of equation 3.12, the distribution of the sample mean* $\bar{X}$ *has expectation and variance:*

$$\begin{aligned} E[\bar{X}] &= \pi\tau \\ Var[\bar{X}] &= \frac{1}{n}(1 + \tau^2\pi(1-\pi)) \end{aligned}$$

Proof

$$\begin{aligned} E[\bar{X}] &= E_{\mathbf{W}}[E[\bar{X}|\mathbf{W}]] = \frac{1}{n}E_{\mathbf{W}}[\sum_j W_j\tau] \\ &= \frac{1}{n}n\pi\tau = \pi\tau \end{aligned}$$

$$\begin{aligned} Var[\bar{X}] &= E_{\mathbf{W}}[Var[\bar{X}|\mathbf{W}]] + Var_{\mathbf{W}}[E[\bar{X}|\mathbf{W}]] \\ &= E_{\mathbf{W}}[\frac{1}{n}] + Var_{\mathbf{W}}[\frac{1}{n}\tau\sum_j W_j] \\ &= \frac{1}{n} + \frac{\tau^2}{n}\pi(1-\pi) \end{aligned}$$

Corollary 3.4 (Two Sample Case) *Under the two sample conditions of corollary 3.2,*

the distribution of the difference in sample means has expectation and variance:

$$\begin{aligned} E[\bar{X}_1 - \bar{X}_2] &= 0 \\ Var[\bar{X}_1 - \bar{X}_2] &= \frac{2}{n}(1 + \tau^2\pi(1-\pi)) \end{aligned}$$

Theorem 3.5 (One Sample Case) *Under the conditions of equation 3.12, and For $n = 4$, the distribution of S_X^2 conditional on $W_\cdot = w_\cdot$ is as follows:*

$$\begin{aligned} (n-1)S_X^2|W_\cdot = 0 &\sim \chi^2_{\nu=3} \\ (n-1)S_X^2|W_\cdot = 1 &\sim \chi^2_{\nu=3,ncp=\frac{3}{4}\tau^2} \\ (n-1)S_X^2|W_\cdot = 2 &\sim \chi^2_{\nu=3,ncp=\tau^2} \\ (n-1)S_X^2|W_\cdot = 3 &\sim \chi^2_{\nu=3,ncp=\frac{3}{4}\tau^2} \\ (n-1)S_X^2|W_\cdot = 4 &\sim \chi^2_{\nu=3} \end{aligned}$$

Proof The proof for $W_\cdot = 0$ or $W_\cdot = 4$ is just the classic non-mixture result.

For $W_\cdot \in \{1, 2, 3\}$, we assume, w.l.o.g., that the X_j are ordered such that the first $n - w_\cdot$ correspond to observations made with latent state 1 and the remaining X_j are made with respect to latent state 2. Let $\bar{X}_k$ and $S^2_{X,k}$ denote the sample mean and variance of the first k ordered observations. We make use of this property (Casella and Berger, 2001):

$$(k-1)S^2_{X,k} = (k-2)S^2_{X,k-1} + \left(\frac{k-1}{k}\right)(X_k - \bar{X}_{k-1})^2$$

where

$$S^2_{X,2} = \frac{1}{2}(X_2 - X_1)^2 \quad .$$

For $W_{\cdot} = 1$:

$$(4-1)S^2_{X,4} = (4-2)S^2_{X,4-1} + \left(\frac{4-1}{4}\right)(X_4 - \bar{X}_3)^2$$

where $S^2_{X,3} \sim \chi^2_2$, $\bar{X}_3 \sim N(0, \frac{1}{3})$, and $X_4 \sim (N(\tau, 1))$. Which implies that:

$$\left(\frac{3}{4}\right)^{1/2}(X_4 - \bar{X}_3) \sim N\left(\left(\frac{3}{4}\right)^{1/2}\tau, \left(\frac{3}{4}\right)\left(\frac{1}{3}+1\right)\right)$$

Hence,

$$(4-1)S^2_{X,4} \sim \chi^2_{\nu=2} + \chi^2_{\nu=1,ncp=\frac{3}{4}\tau^2} \sim \chi^2_{\nu=3,ncp=\frac{3}{4}\tau^2}$$

By a symmetry argument the above result also holds for $W_{\cdot} = 3$

For $W_{\cdot} = 2$:

$$(3-1)S^2_{X,3} = (3-2)S^2_{X,2} + \left(\frac{3-1}{3}\right)(X_3 - \bar{X}_2)^2$$

where $S^2_{X,2} \sim \chi^2_1$, $\bar{X}_2 \sim N(0, \frac{1}{2})$, and $X_3 \sim (N(\tau, 1))$. Which implies that:

$$\left(\frac{2}{3}\right)^{1/2}(X_3 - \bar{X}_2) \sim N\left(\left(\frac{2}{3}\right)^{1/2}\tau, \left(\frac{2}{3}\right)\left(\frac{1}{2}+1\right)\right)$$

Hence,

$$(3-1)S^2_{X,3} \sim \chi^2_{\nu=1} + \chi^2_{\nu=1,ncp=\frac{2}{3}\tau^2} \sim \chi^2_{\nu=2,ncp=\frac{2}{3}\tau^2}$$

and

$$(4-1)S^2_{X,4} = (4-2)S^2_{X,4-1} + \left(\frac{4-1}{4}\right)(X_4 - \bar{X}_3)^2$$

where $S^2_{X,3} \sim \chi^2_{\nu=2,ncp=\frac{2}{3}\tau^2}$, $\bar{X}_3 \sim N(\frac{1}{3}\tau, \frac{1}{3})$, and $X_4 \sim (N(\tau, 1))$. Hence,

$$\left(\frac{3}{4}\right)^{1/2}(X_4 - \bar{X}_3) \sim N\left(\left(\frac{3}{4}\right)^{1/2}\frac{2}{3}\tau, \left(\frac{3}{4}\right)\left(\frac{1}{3}+1\right)\right)$$

and

$$(4-1)S^2_{X,4} \sim \chi^2_{\nu=2,ncp=\frac{2}{3}\tau^2} + \chi^2_{\nu=1,ncp=\frac{1}{3}\tau^2} \sim \chi^2_{\nu=3,ncp=\tau^2}$$

Theorem 3.6 (One Sample Case) *Assume the conditions of equation 3.12 and* $n = 4$. *The distribution of* $(4-1)S^2_X$ *is a mixture of* χ^2 *distributions, some central and some non-central.*

$$\begin{aligned}(4-1)S^2_4 \quad \sim \quad & \left(B(0,4,\pi) + B(4,4,\pi)\right)\chi^2_{\nu=3} + \left(B(1,4,\pi) + B(3,4,\pi)\right)\chi^2_{\nu=3,ncp=\frac{3}{4}\tau^2} \\ & + B(2,4,\pi)\chi^2_{\nu=3,ncp=\tau^2}\end{aligned}$$

where $B(x,n,\pi)$ *is the pmf of the Binomial distribution with* x *successes in* n *trials with a success probability of* π.

Proof application of previous theorem results.

Theorem 3.7 (One Sample Case) *Assume the conditions of equation 3.12 with* $\pi \in (0,1)$ *and* $\tau > 0$. *Under said conditions, the distributions of* $\bar{X} - \pi\tau$ *and* S^2_X *are dependent provided* $\pi \neq 1/2$. *When* $\pi = 1/2$, $Cov(\bar{X}, S^2) = 0$.

Proof Let $Y_j = X_j - \pi\tau$ and note that $\mu_Y = E(\bar{Y}) = 0$, $S^2_Y = S^2_X$ and $Cov(\bar{Y}, S^2_Y) = Cov(\bar{X}, S^2_X)$. Using the result (Mukhopadhyaya and Son, 2011) that

$$Cov(\bar{Y}, S^2_Y) = n^{-1}\mu_3 \tag{3.15}$$

where μ_3 is the third central moment, we find:

$$\begin{aligned} E[(Y-\mu_Y)^3] = E[Y^3] &= E_W\left[E\left[Y^3|W\right]\right] \\ &= E_W\left[\mu^3_{Y|W} + 3\mu_{Y|W}\right] \qquad \text{since } Y|W \sim N((W-\pi)\tau, 1) \\ &= E_W\left[(W-\pi)^3\tau^3 + 3(W-\pi)\tau\right] \\ &= (\pi - 3\pi^2 + 2\pi^3)\tau^3 \qquad \text{since } W \sim Binomial(1,\pi) \\ &= \pi(1-3\pi+2\pi^2)\tau^3 \quad . \end{aligned}$$

If $\pi = 1/2$, then $E[Y^3] = 0$ and $Cov(\bar{X}, S^2) = 0$. If $\pi \neq 1/2$, then $Cov(\bar{X}, S^2) \neq 0$, and $\bar{X}$ and S^2 are linearly dependent.

Comment: Although $Cov(\bar{X}, S^2) = 0$ when $\pi = 1/2$, a non-linear dependency exists between $\bar{X}$ and S^2. In the two sample case, $Cov(\bar{X}_1 - \bar{X}_2, S_1^2 + S_2^2) = Cov(\bar{X}_1, S_1^2) - Cov(\bar{X}_2, S_2^2) = 0$. Simulation results in Section 4 will show that $\bar{X}_1 - \bar{X}_2$ and $S_1^2 + S_2^2$ are dependent in a non-linear fashion. One sample simulation results (not shown) display a similar relationship.

4 Simulation Results

The results of a simulation study are presented in Figure 4. For each value of $\tau \in \{0, 1, 2, 5, 10\}$, 1,000,000 replicate data-sets were simulated under model (1.4) with $n = 4$ in and $\pi_1 = \pi_2 = \frac{1}{2}$). For each of the 5,000,000 replicate data sets, the equal variance t-test statistic was calculated. Figure 4(a) displays the quantile plots of the observed equal variance t-test statistics versus the quantiles of a t_6. Note the excellent agreement between -4 and 4 and the clear anti-conservative breakdown for more extreme quantiles.

By Theorem 3.4, $E[\bar{X}_1 - \bar{X}_2] = 0$ and $Var[\bar{X}_1 - \bar{X}_2] = \frac{1}{2}(1 + \frac{1}{4}\tau^2)$. Figure 4(b) contains the quantile plot for $\bar{X}_1 - \bar{X}_2$ versus the quantiles of a $N(0, \frac{1}{2}(1 + \frac{1}{4}\tau^2))$. The observed agreement in the quantile plot suggests that normality is not an unreasonable assumption for the distribution of $\bar{X}_1 - \bar{X}_2$. Therefore, the discrepancy between the observed and

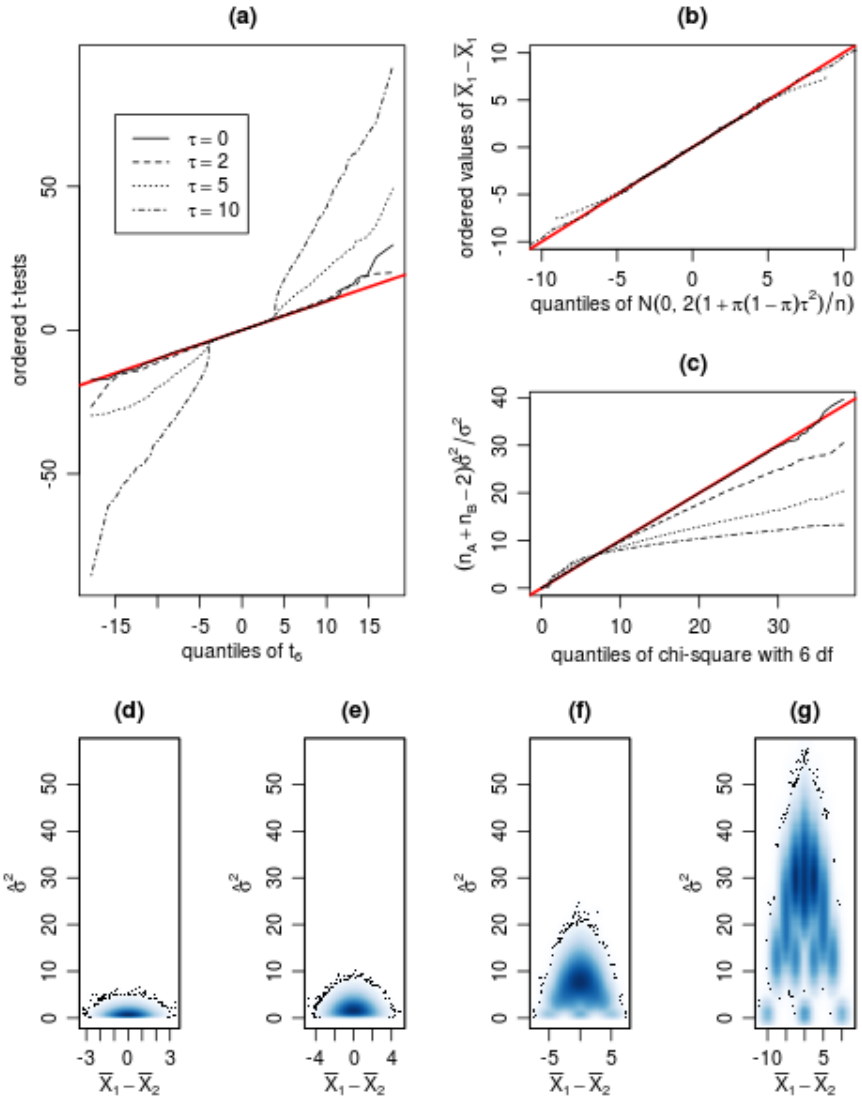

Figure 4: Simulation Results. For each value of $\tau \in \{0, 1, 2, 5, 10\}$, 1,000,000 replicate data-sets were simulated under the simple two sample two latent state model with $\pi = 2$ and $n = 4$. The the equal variance t-test statistic, and its components were calculated for each data set. (a) QQ-plot of the observed t-tests versus a t_6 distribution. Note the anti-conservative breakdown for extreme quantiles. (b) QQ-plot demonstrates reasonable agreement between the observed distributions of $\bar{X}_1 - \bar{X}_2$ and $N(0, \frac{1}{2}(1 + \frac{1}{4}\tau^2))$ distributions. (c) QQ-plot of $(2n-2)\hat{\sigma}^2/\sigma^2$ versus a χ^2_6, where $\hat{\sigma}^2$ denotes the pooled sample variance estimator and $\sigma^2 = \frac{1}{2}(1+\frac{1}{4}\tau^2)$. As τ increases, the observed quantiles fall markedly below their theoretical counterparts. (d)-(g) kernel smoothed estimated density plots of the observed $\bar{X}_1 - \bar{X}_2$ versus $\hat{\sigma}^2$ for $\tau \in \{0, 1, 2, 5, 10\}$. As τ increases, the (non-linear) dependence of $\bar{X}_1 - \bar{X}_2$ and $\hat{\sigma}^2$ becomes increasingly obvious.

theoretical quantiles for the t-test statistics does not appear to be caused by a violation of that assumption.

Figure 4(c) contains the quantile plot for $(2n-2)\hat{\sigma}^2/\sigma^2$ versus the quantiles of a chi-

square with 6 df, where $\hat{\sigma}^2$ denotes the pooled sample variance estimator. As τ increases, the observed quantiles fall markedly below their theoretical counterparts. For this plot, $\sigma^2 = \frac{1}{2}(1 + \frac{1}{4}\tau^2)$, the true theoretical variance. The lack of agreement between observed and theoretical quantiles is an expected consequence of the extension of Theorem 3.5 results for the two sample case. In the presence of the latent structure, the scaled variance estimator is not simply distributed as a chi-square random variable. Rather, it is a mixture of non-central and central chi-square distributions.

Figure 4(d)-(g) provide kernel smoothed density plots of the observed $\bar{X}_1 - \bar{X}_2$ versus $\hat{\sigma}^2$ for $\tau \in \{0, 1, 2, 5, 10\}$. The derivation of the t-distribution relies upon an assumption of independence of the difference in sample mean and the pooled variance estimator. As τ increases, the dependence of $\bar{X}_1 - \bar{X}_2$ and $\hat{\sigma}^2$ becomes increasingly obvious. This makes intuitive sense in the latent model setting. Knowledge of the mean provides knowledge of the latent states, especially for larger τ. With respect to the sample variance (i.e., the pooled estimator), latent state combinations with group observations in both states will be estimated to be more variable than those with observations in only one of the two possible states. Since the data was simulated with $\pi = \frac{1}{2}$, the distributions are symmetric about $\bar{X}_1 - \bar{X}_2 = 0$ and, as predicted by Theorem 3.7, the (linear) covariance is zero although a clear non-linear dependency exists.

Figure 5 illustrates the results of the simulation study conditioned on the underlying latent state. The underlying states are nomenclated as pq where $p \in \{0, 1, 2, 3, 4\}$ and denotes the number of observations in the first group with underlying latent state 2, and $q \in \{0, 1, 2, 3, 4\}$ and denotes the number of observations in the second group with underlying latent state 2. For example, the label 04 denotes the condition of our motivating example, provided in Figure 2.

Figure 5(a) contains the the kernel density estimated densities of the observed t-statistics for $\tau = 10$ and conditioned on latent state. Conditioned on the extreme latent

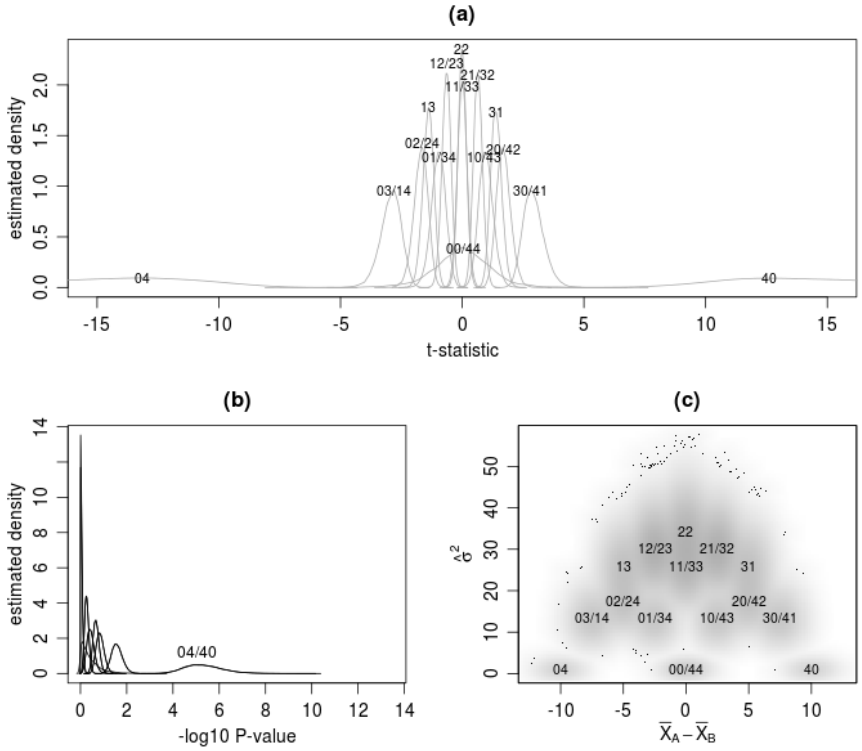

Figure 5: Latent State Annotated Simulation Results. The simulation results corresponding to $\tau = 10$ from Figure 4 are annotated with values pq where p and q denote the number of latent state 2 observations in the first and second sample groups, respectively. (a) Kernel density estimates for the observed t-statistics. Latent states 04 and 40 contribute enough mass to the tail regions of the true t-statistic null distribution to render the t_6 distribution an unsuitable approximation. (b) Kernel density estimates for the $-\log_{10}$ p-values of the two-sided equal variance t-statistics conditioned on latent state. (c) Kernel smoothed density plots of the observed $\bar{X}_A - \bar{X}_B$ versus $\hat{\sigma}^2$ for $\tau = 10$. The latent state labels are centered at the coordinates corresponding to the expected mean differences and variance estimates as predicted by the theorems provided in Section 3.

states, $\mathbf{W_h} = w_{0/4}$, and $\mathbf{W_h} = w_{4/0}$, we have the conditional T-distributions:

$$\begin{aligned} T|\mathbf{W_h} = w_{0/4} &\sim t - dist(\text{df} = 6, \text{ncp} = -10\sqrt{(2)}) \\ T|\mathbf{W_h} = w_{4/0} &\sim t - dist(\text{df} = 6, \text{ncp} = 10\sqrt{(2)}) \end{aligned}$$

As expected, the densities in Figure 5(a) corresponding to latent states 04 and 40 contain notable mass located away from the origin. As per the logic of Theorem 2.1 and Corollary 2.2, latent states 04 and 40 will, under certain combinations of π, τ, and n, con-

tribute enough mass to the tail regions of the true t-statistic null distribution to render the t-distribution an unsuitable approximation.

Figure 5(b) contains the kernel density estimated densities of the $-\log_{10}$ p-values of the two-sided equal variance t-statistics conditioned on latent state. As expected, the $-\log_{10}$ p-values corresponding to latent states 04 and 40 contain notable mass for large values, a result that is consistent with the motivating example presented in Section 2.

Figure 5(c) contains the kernel smoothed estimated density plots of the observed $\bar{X}_A - \bar{X}_B$ versus $\hat{\sigma}^2$ for $\tau = 10$. The latent state labels are centered at the coordinates corresponding to the expected mean differences and variance estimates as predicted by the theorems provided in Section 3. This figure explains the "paw print" nature of Figure 4(d)-(g), where different "paw pads" correspond to the collection of bivariate distributions of $(\bar{X}_A - \bar{X}_B, \hat{\sigma}^2)$ which differ according to latent state.

5 Discussion

The parametric t-test can be grossly anti-conservative when it is applied to a manifest variable that is governed by an underlying latent state. Unbalanced latent state assignments (e.g., $\mathbf{W_h} = w_{0/4}$) contribute mass to the tails of the true t-test null distribution and can render the the t-distribution an inadequate approximation. In terms of our simple two sample two latent state models, the likelihood of unbalanced states to occur is governed by π and the mass that they contribute to the tail of the distribution is governed by τ. Values of π close to $\frac{1}{2}$ and larger values of τ constitute the least favorable conditions for the t-distribution approximation to the actual null distribution.

By our standardized parameterization, the increased precision of the manifest assays to discriminate underlying latent structures corresponds to larger τ values. Hence, as technologies improve and assays are measured with greater accuracy, the problems discussed in this manuscript will become increasingly relevant. In the motivating example data-set, an effective four fold increase in τ (equivalent to a reduction in σ in equation 1.3

from 1 to $\frac{1}{4}$) changes the odds against the null hypothesis from 2741:1 to 12.5×10^6:1. Those estimated odds are in stark disagreement with the 34:1 odds derived from the Fisher's Exact Test of the true latent state counts. Figure 2(d) demonstrates how increased accuracy can dramatically inflate the odds ratio of the t-distribution based odds versus the Fisher's Exact Test odds, where the Fisher's Exact Test odds are computed assuming perfect knowledge of the latent states.

Test multiplicity across a large number of manifest assays, each with plausible latent structure, can also compound the problem. Theorem 2.1 and Corollary 2.2 demonstrate that combinations of n, H, τ, and π exist such that even conservative Bonferroni family-wise error control is lost. Figure 3 provides estimated bounds for the parameter combinations for which error control is lost. For smaller n loss of error control can occur for plausible parameter combinations, e.g., τ is not required to be implausibly large. If Bonferroni control can be lost, it is reasonable to expect that other methods of error control will also fail. For example, the p-values from the t-distribution null will be non-uniform and anti-conservative and this will adversely effect FDR control methods such as that of Benjamini and Hochberg (Benjamini and Hochberg, 1995). Mixture model based methods for the empirical control of the false discovery rate Efron (Efron, 2010) and of the generalized family wise eeror rate (Miecznikowski and Gaile, 2014), will also be adversely effected as the tail heaviness (compared to the t-distribution) of the true null distribution is confounded with true departures from the null distribution of test statistics.

An underlying latent structure can have a profound effect on the distribution of the variance estimator. Theorems 3.5 and 3.6 state that S^2 is distributed according to mixture of central and non-central chi-square distributions. Theorem 3.7 and Figures 4(d)-(g) and 5(c) demonstrate that the sample variance estimators (in both the one and two sample cases) are not independent of the sample means. For $\pi \neq \frac{1}{2}$, a linear component to the dependence exists, and for $\pi = \frac{1}{2}$ it does not (although a non-linear dependence does exist).

Although not included here, similar results were also observed for the unequal variance t-test. The equal variance t-test was used for ease of exposition.

In some, if not most, cases, the plausibility of the existence of a latent structure can only be evaluated by developing a proper understanding of the underlying biological processes at play. Inspection of the data itself may not be sufficiently informative. For example, the assay data observed for both sample groups in Figure 2(a) are normally distributed and generated under equal variances. There is nothing in the observed data to suggest that the normal distribution of the assay values are actually conditioned on unobserved realizations of Bernoulli distributed latent variables. By all appearances, the data could very well have been generated under a classic mean shift alternative model.

If a data analyst can not dismiss the plausibility of a latent structure existing in the context of a small two sample comparison of continuous assay values, then an alternative test to the parametric t-test should be considered. For example, cut-off rules could be employed and Fisher's Exact Tests could be utilized. The tests could even be conducted across a range of plausible cut-off values and the p-values could be adjusted using a minP approach (Westfall and Young, 1993). A permutation t-test provides another reasonable alternative. As we mentioned above, the permutation of group labels with respect to the manifest variable outcomes will also permute those labels with respect to the underlying latent structure. Hence, the permutation null distribution should provide an adequate approximation to the true null distribution. Although these alternative approaches may not provide the incredibly small p-values that many researchers desire, they protect against the very real possibility of reporting p-values that are misleading.

As our motivating example has demonstrated, parametric t-test p-values can provide stunning evidence against the null hypothesis and yet completely fail in subsequent validation studies. We speculate that this phenomenon may have contributed to unexpectedly high validation failure rates for putative biomarkers identified in early microarray experiments (where sample sizes were small and latent structures plausible). It could also contribute to higher than expected validation failure rates in the context of pilot

animal or cell line studies.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, 13(10):705–719.

Casella, G. and Berger, R. L. (2001). Statistical Inference, Second edition. pages 220–221.

Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*, 22(4):271–274.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Fisher, R. A. (1925). Applications of student's distribution. *Metron*, 5:90–104.

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Gosset, W. (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Laurila, K., Oster, B., Andersen, C. L., Lamy, P., Orntoft, T., Yli-Harja, O., and Wiuf, C. (2011). A beta-mixture model for dimensionality reduction, sample classification and analysis. *BMC Bioinformatics*, 12:215.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics, New York.

Miecznikowski, J. C. and Gaile, D. P. (2014). A novel characterization of the generalized family wise error rate using empirical null distributions. *Stat Appl Genet Mol Biol*, 13(3):299–322.

Mukhopadhyaya, N. and Son, M. S. (2011). On the covariance between the sample mean and variance. *Communications in Statistics - Theory and Methods*, 40(7):1142–1148.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29(3):263–264.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196.

Westfall, P. H. and Young, S. S. (1993). ReSampling Based Multiple Testing.