Hypothesis testing procedures for two sample means with applications to gene expression data

Michail Tsagris¹, Abdulaziz Alenazi² and Kleio-Maria Verrou³

¹ Department of Economics, University of Crete, Greece, mtsagris@uoc.gr
² Department of Mathematics, Northern Border University, Saudi Arabia, a.alenazi@nbu.edu.sa
³ Department of Computer Science, University of Crete, Greece, kleioverrou@yahoo.com

December 2018

Abstract

In Bioinformatics, the number of available variables for a few tens of subjects, is usually in the order of tens of thousands. As an example is the case of gene expression data, where usually two groups of subjects exist, cases and controls or subjects with disease and subjects without disease. The detection of differentially expressed genes between the two groups takes place using many 2 independent samples (Welch) t-tests, one test for each variable (probeset). Motivated by this, the present research examines the empirical and exponential empirical likelihood, asymptotically, and provides some useful results revealing their relationship with the James’s and Welch t-test. By exploiting this relationship, a simple calibration based on the $t$ distribution, applicable to both techniques, is proposed. Then, this calibration is compared to the classical Welch t-test. A third, more famous, non parametric test subject to comparison is the Wilcoxon-Mann-Whitney test. As an extra step, bootstrap calibration of the aforementioned tests is performed and the exact p-value of the Wilcoxon-Mann-Whitney test is computed. The main goal is to examine the size and the power behaviour of these testing procedures, when applied on small to medium sized datasets. Based on extensive simulation studies we provide strong evidence for the Welch t-test. We show, numerically, that the Welch t-test has the same power abilities with all other testing procedures. It outperforms them though in terms of attaining the type I error. Further, it is computationally extremely efficient.

Keywords: Hypothesis testing, t-test, Welch test, empirical likelihood, exponential tilting, bootstrap

2010 MATHEMATICS SUBJECT CLASSIFICATION: 62G10, 62G09

1 Introduction

One of the most common tasks in the field of Biology is to identify properties that are link with specific phenotypes. A fundamental process that affects the biological behaviour of the
cell is the expression levels of its genes. Finding the genes that are expressed in a different abundance between two phenotypes has lead to the identifications of properties that are linked with specific conditions, like diseases. For this reason, experiments that measure the expression levels between two conditions (disease or no disease for example) are widely applied, leading to the construction of high-dimensional gene expression data, that include the information of the condition from which they derive. This experiments, mainly based on DNA microArray (Wheelan et al., 2008) or RNAseq (Denoeud et al., 2008) techniques for measuring the expression levels are typically returning datasets that consist of 55,000 variables (genes, probesets) and a few tens of observations. The identification of those variables that are important is then shouldered to the field of Bioinfomatics.

Therefore, one of the most common tasks in the field of Bioinformatics is the identification of genes that are significantly differentially expressed between phenotypes. The goal of those analysis is to identify genes that are expressed in different levels between conditions, and may provide richer information about the examined phenotype or are linked with a disease, when this is the examined phenotype. A gene is defined as differentially expressed if a statistically significant difference is observed in the expression levels between the two experimental conditions. To find differentially expressed genes between the conditions, it is important to identify a statistical distributional property of the data, like the mean value, to approximate and compare. The most frequently tests used for the comparison of these gene expression levels are the Welch t-test and the Wilcoxon-Mann-Whitney test.

Empirical likelihood (EL) is a non-parametric method of statistical inference (Owen, 1988, 1990, 2001). It allows the analysis to be carried out via likelihood methods, without the assumption of the data coming from any particular family of distributions. A substantial property of the empirical likelihood is that it is Bartlett correctable (DiCiccio et al., 1991). Exponential empirical likelihood (EEL), or non-parametric tilting or exponential tilting, was first introduced by Efron (1981) and is similar to EL. However, it is not Bartlett correctable (Jing and Wood, 1996).

Simulation studies have shown that the empirical likelihood ratio test, in the case of both a univariate and multivariate population mean, does not estimate the correct probability of type I error, when the sample size is relatively small. In fact, it tends to overestimate the probability of type I error. This is because the EL method is asymptotically size correct, which means that this method attains the probability of type I error when the sample size is large enough. Emerson (2009) produced a helpful table, where the small sample improvement of different methods in the case of a listed population mean. The improvement is in terms of the estimated probability of type I error. Evidence of the under-coverage problem when constructing confidence intervals for the mean, can be found in (Qin and Lawless, 1994). The Bartlett correction improves the coverage in general, but does not improve the performance significantly. As for the EEL, an adjustment was suggested by Zhu et al. (2008).

With regards to the power of EL, Owen (1990) showed that the distribution of the test statistic under local alternatives is asymptotically a non-central $\chi^2$ distribution. Chen (1994) stated that the power of the bootstrap and of EL have the same first order term in their power functions. Thus, he compared the higher order terms of these two ways in the univariate and bivariate cases. In the univariate case, Chen (1994) proposed a rule using either the bootstrap
or EL based upon the direction of the alternative and the sign of the skewness parameter. The rule to choose either testing procedure in the bivariate case relies upon the difference of the second order terms of these two procedures.

Hypothesis testing for two means has been the debate of many papers (Brunner and Munzel, 2000; Fagerland and Sandvik, 2009; Fagerland, 2012; Derrick et al., 2016; Chen et al., 2016; Dwivedi et al., 2017; Sudhir et al., 2018). Yet, a solid and concrete large scale Monte Carlo study covering many aspects and scenarios has not been performed.

In this work, we examine these two non parametric likelihoods as tools for hypothesis testing of two sample means. Their theoretical properties are examined and their relationship to the Welch and James test statistics are presented. The Wilcoxon-Mann-Whitney test, as another non parametric competitor to the parametric t-test is, furthermore, discussed. We implemented a large scale Monte Carlo simulation study comparing the performance of parametric and non parametric methods for the means of two samples. For this open problem, bootstrap calibration is proved useful. We included various scenarios of asymmetric or bi-modal distributions and used data from constrained spaces, e.g. percentages and circular data, as well as discrete distributions.

The choice of these distributions aims at providing answers and directions-recommendations to two issues. At first, to give an introduction to these two non-parametric likelihoods, examine some of their properties and show their relationship to the Welch t-test (Sections 3.1 and 3.2). Secondly and most important, to give recommendations, based on simulation studies, as for which test statistic or method is to be preferred when two sample means are subject to comparison (Section 5). These comparisons cover the issues of probability of type I error, power of the test. We, furthermore, pinpoint the robustness of the Welch t-test, not only to large deviations from the normal distribution, but also to the space of the data and the type of variables (continuous or discrete). To highlight the computational benefit for a bioinformatician, we compare the computational cost of each test using 4 real gene expression datasets. Finally, we summarise our conclusions and findings.

2 Classical tests for two population means

It is well known that, when the distributional assumptions hold true, the non-parametric likelihood test statistics converge to the quadratic test statistics, in the sense that the asymptotic form of the test statistic in the non-parametric likelihood procedures is the same as the form of the quadratic tests. The two quadratic tests used are the t-test and the Welch modification of it, when the two population variances cannot be assumed to be equal.

2.1 Welch t-test

Welch (1947) proposed a test for linear form of hypotheses of the population means when the variances are not known and cannot be assumed equal. The test statistic for two samples is

$$T_w = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$  (1)
Under $H_0$, $T_w \sim t_\nu$, where (Satterthwaite, 1946; Welch, 1947)

$$\nu \approx \left( \frac{s_1^2}{n_1(n_1-1)} + \frac{s_2^2}{n_2(n_2-1)} \right).$$

(2)

### 2.2 Wilcoxon-Mann-Whitney rank based test

The MannWhitney U test (Mann and Whitney, 1947), also called Wilcoxon-Mann-Whitney test or Wilcoxon rank-sum test, is a rank based test which implicitly (indirectly) assumes (or implies underneath) a normal distribution. Nonetheless, the test belongs in the non parametric family of tests and is a competitor of the t-test. The procedure is described below. Since Wilcoxon (1945) was the first to propose it, we denote it by WMW throughout the paper. The procedure Wilcoxon (1945) proposed can be described as follows.

1. Merge the observations from both samples in one vector.
2. Assign ranks to the whole vector of observations.
3. Calculate $W = \sum_{i=1}^{n_1} R_{1i}$, where $R_{1i}$ is the vector of ranks of the observations belonging to the first sample and $n_1$ is its sample size.

Using (2) one can calculate the exact p-value of WMW. R (R Core Team, 2018) calculates a modified version of $W$, denoted by $U = W - \frac{n_1(n_1+1)}{2}$. The standardised test statistic is given by

$$z = \frac{U - 0.5n_1n_2}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}}.$$  

(3)

If there are ties in the ranks, the denominator of (3) should be corrected and (3) be written as

$$z_c = \frac{U - 0.5n_1n_2}{\sqrt{\frac{n_1n_2}{12} \left( n_1 + n_2 + 1 - \sum_{i=1}^{k} \frac{t_i^2 - t_i}{(n_1+n+2)(n_1+n-1)} \right)}},$$

(4)

where $t_i$ is the number of observations sharing rank $i$, and $k$ is the number of (distinct) ranks. Under $H_0$, asymptotically, $z$ and $z_c$ follow the standard normal distribution. R (R Core Team, 2018) computes the exact p-value, if the samples contain less than 50 finite values and there are no ties. Otherwise, the normal approximation is used.

### 3 Non parametric likelihoods

#### 3.1 Empirical likelihood

Empirical likelihood is a non-parametric likelihood developed by Owen (1988, 1990) in order to perform non-parametric hypothesis testing. The rationale is that a positive probability $p_i$ is assigned to each observation such that $\sum_{i=1}^{n} p_i = 1$ and thus the form of the non-parametric
likelihood is
\[ L(F) = \prod_{i=1}^{n} p_i, \quad (5) \]
where \( F \) stands for the cumulative distribution function.

The task is to maximise (5) with respect to the \( p_i \)'s, subject to some constraints regarding the nature of the application. In the one sample hypothesis testing about the true value of a univariate mean (under \( H_0 \) its value is \( \mu \)), the constraint is
\[ \sum_{i=1}^{n} p_i (x_i - \mu) = 0. \quad (6) \]
Moving in the same context, Owen (2001) applied Wilks’s theorem in the empirical likelihood setting.

\[ R(F) = \frac{L(F_0)}{L(F_n)} = \prod_{i=1}^{n} \frac{p_i}{\frac{1}{n}} = \prod_{i=1}^{n} np_i, \quad (7) \]
where \( F_0 \) stands for the constrained distribution, which under \( H_0 \) assigns a probability \( p_i \) to each of the observations. \( F_n \) is the empirical distribution which under the alternative hypothesis \( H_1 \) assigns the same probability \( \frac{1}{n} \) to all observations. It is known that under \( H_0 \), \(-2 \log R(F) \sim \chi^2\) with some degrees of freedom. Therefore, the problem is transformed to the conditional maximisation of (7)
\[ \max \left\{ \prod_{i=1}^{n} np_i \mid \sum_{i=1}^{n} p_i x_i = \mu, p_i > 0, \sum_{i=1}^{n} p_i = 1 \right\}. \quad (8) \]
Introducing Lagrangian parameters and after some algebra, we can derive the following form of the \( p_i \)'s
\[ p_i = \frac{1}{n \left[ 1 + \lambda (x_i - \mu) \right]}, \]
Thus, the constraint (6) becomes
\[ \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \mu}{1 + \lambda (x_i - \mu)} = 0. \quad (9) \]
A numerical search over \( \lambda \) is necessary to obtain the \( p_i \)'s which maximise (7). This is a trivial task, since in the univariate case we can simply search over an interval close to zero. An important issue that has to be highlighted, is that (8) exists if and only if \( \mu \) lies within the convex hull of the data.

3.1.1 EL in the two sample case
Jing (1995) and Liu et al. (2008) showed how the two-sample hypothesis testing can be performed using EL. A proof of the asymptotic distribution of the test statistic, when more than two
samples in the univariate case exist, can be found in Owen (2001). In this work, a different approach to prove the asymptotic \( \chi^2 \) distribution in the (without loss of generality to more samples) two-sample case (see Appendix) was applied.

The probabilities of each of the \( j (=1, 2) \) samples have the following form

\[
p_{ji} = \frac{1}{n_j} [1 + \lambda_j (x_{ji} - \mu)]^{-1}.
\]

(10)

The log-likelihood ratio test statistic, thus, takes the following form

\[
\Lambda = \sum_{j=1}^{2} \frac{n_j (\bar{x}_j - \mu)^2}{s_j^2(\mu)} + \text{O}_p \left( n_0^{-1/2} \right).
\]

(11)

Asymptotically, the final term is the sum of 2 independent \( \chi^2 \) variables under \( H_0 \) (that each population mean is \( \mu \)). This is true, since \( s_j^2(\mu) \xrightarrow{P} \sigma^2 \), where \( \sigma^2 \) is the population variance. The degrees of freedom are equal to 1 (Owen, 2001).

### 3.1.2 Constraint of the \( \lambda_j \)'s at the optimum

Jing (1995) and Liu et al. (2008) used a sum-to-zero constraint for the \( \lambda \)'s but without reasoning or proving its correctness. This constraint should, moreover, be used when the empirical likelihood for a two-sample problem is applied. The \( \lambda \)'s are linearly dependent at the maximised log-likelihood ratio test statistic with coefficients given by the sample sizes

\[
\sum_{j=1}^{k} n_j \lambda_j = 0.
\]

It is known that, the two sample EL method can be obtained from minimising, with respect to the common mean, the sum of the two one-sample likelihood ratio test statistics from the common mean (Emerson, 2009). The same procedure was followed by Amaral and Wood (2010) and Tsagris et al. (2017) as well. This means that the goal is to minimise the sum of test statistics of whether each sample mean is equal to the common mean \( \mu \). The above constraint is derived by differentiating the (A.3) with respect to \( \mu \). This constraint holds true, asymptotically, at the maximum of the test statistic. This constraint is important and a remark stemming from it is showed below for the two samples case

\[
n_1 \lambda_1 + n_2 \lambda_2 = 0 \Rightarrow n_1 s_1^2 (\mu)^{-1} (\bar{x}_1 - \mu) = -n_2 s_2^2 (\mu)^{-1} (\bar{x}_2 - \mu) \Rightarrow \text{(by using A.2)}
\]

\[
\hat{\mu} = \left[ n_1 s_1^2 (\hat{\mu})^{-1} + n_2 s_2^2 (\hat{\mu})^{-1} \right]^{-1} \left[ n_1 s_1^2 (\hat{\mu})^{-1} \bar{x}_1 + n_2 s_2^2 (\hat{\mu})^{-1} \bar{x}_2 \right].
\]

(12)

The estimate for the common mean is the same as the Gaussian maximum likelihood estimate of the common mean of two samples without assuming equal variances. This is also the starting point for the Behrens-Fisher problem: testing equality of means with no assumptions about the variances. The formula for the common mean (12) can, of course, be generalised to more than two samples. Furthermore, it has to be noted, that the above estimate is found iteratively, since the common mean \( \hat{\mu} \) appears on both sides of (12).
3.1.3 Power of the empirical likelihood

The power of EL in the one sample was presented in Owen (1990) and has been studied by Lazar and Mykland (1998). Owen (1990) showed that the distribution of the test statistic under a particular $H_1$ (see (14) for an assumed alternative hypothesis) is asymptotically a non-central $\chi^2$ distribution. In (11) it is clear, that the test statistic asymptotically is the sum of the one sample test statistics from the common mean and has the following form

$$\Lambda = \sum_{j=1}^{2} \frac{n_j (\bar{x}_j - \mu)^2}{s_j^2(\mu)}. \quad (13)$$

Let us assume that under $H_1$ each mean $\mu_j$ deviates from the common mean $\mu$ as a function of the sample variance and the sample size of each sample plus a constant vector. Then the mean under $H_1$ can be written as a function of the variance and of the sample size

$$\mu_j = \mu + \frac{\sigma_j}{\sqrt{n_j}} \tau_j, \quad (14)$$

where $\sigma_j$ is the true variance of the $j$-th sample. By substituting (14) in the place of $\mu$ in (13), and after some tedious algebra, the asymptotic form of the test statistic under $H_1$ becomes

$$\Lambda = \sum_{j=1}^{2} \left( \frac{\sigma_j^{-1}}{\sqrt{n_j}} (\bar{x}_j - \mu) - \tau_j \right)^2 \left( 1 + \frac{T_j^2 - 1}{n_j} \right)^{-1},$$

where $T_j$ is the $j$-th one sample t-test statistic about $\mu$. Asymptotically, the scalar factor $\left( 1 + \frac{T_j^2 - 1}{n_j} \right)^{-1}$ disappears and thus, the asymptotic form of the test statistic under $H_1$ (14) becomes equal to the sum of 2 independent non-central $\chi^2$ variables, where each of them has a non-centrality parameter equal to $\tau_j^2$. Consequently, $\Lambda$ follows asymptotically a non-central $\chi^2$ with non-centrality parameter $\sum_{j=1}^{2} \tau_j^2$. Owen (1990) proved that in the two-sample case, the test statistic follows asymptotically a non central $\chi^2$ distribution. In the Appendix can be found a rather less complicated and easy to understand approach, under this sequence of alternative hypotheses.

3.2 Exponential empirical likelihood

Exponential empirical likelihood or exponential tilting was first introduced by Efron (1981) as a way to perform a ”tilted” version of the bootstrap for the one sample mean hypothesis testing. Similarly to EL, the idea is to put some positive weights $p_i$, which sum to one, on the observations such that the weighted sample mean $\bar{x}$ is equal to a population mean $\mu$ under the null hypothesis. Under the alternative hypothesis the weights are equal to $\frac{1}{n}$, where $n$ is the sample size. The choice of $p_i$s minimises the Kullback-Leibler distance from $H_0$ to $H_1$ (Efron, 1981):

$$D(L_0, L_1) = \sum_{i=1}^{n} p_i \log (np_i), \quad (15)$$

7
subject to the same constraint as with EL (6). With EEL, the probabilities take the following form:

\[ p_i = \frac{e^{\lambda x_i}}{\sum_{j=1}^{n} e^{\lambda x_j}}. \]  

(16)

The constraint, then, becomes equal to

\[ \sum_{i=1}^{n} e^{\lambda x_i} (x_i - \mu) = \sum_{j=1}^{n} e^{\lambda x_j} - \mu = 0. \]  

(17)

Similarly to EL, a numerical search over \( \lambda \) is necessary to find the probabilities that minimise the Kulback-Leibler distance from \( H_0 \) to \( H_1 \) (15).

### 3.3 The two sample means case

Assuming 2 samples and \( \mu \) being the common mean, the form of the probabilities is

\[ p_{ji} = \left( \sum_{m=1}^{n} e^{\lambda_j x_{jm}} \right)^{-1} e^{\lambda_j x_{ji}} = \left[ \sum_{m=1}^{n} e^{\lambda_j x_{jm} - \mu} \right]^{-1} e^{\lambda_j (x_{ji} - \mu)} \quad (j = 1, 2). \]  

(18)

Again, it is necessary to impose a constraint onto the Lagrangian parameters \( \lambda_j \). Note that here the constraint can be expressed as \( \sum_{i=1}^{n} p_i (x_i - \mu) = 0 \), whereas in the empirical likelihood the constraint must stay \( \sum_{i=1}^{n} p_i (x_i - \mu) = 0 \). In the exponential tilting no difference is derived. However, in the empirical likelihood a difference is indicated, since the resulting probabilities in the empirical likelihood depend upon \( \mu \).

Under \( H_0 \) the log-likelihood ratio test statistic, asymptotically, becomes

\[ \Lambda = \sum_{j=1}^{2} n_j \frac{(\bar{x}_j - \mu)^2}{s_j^2(\bar{x}_j)} + O_p \left( n_0^{-1/2} \right), \]  

(19)

where the \( s_j^2(\bar{x}_j) \) denotes the sample biased variance. Under \( H_0 \) \( \Lambda \) (19) follows asymptotically a \( \chi^2 \) distribution with 1 degree of freedom. This is true by combining Slutsky’s theorem with \( s^2(\bar{x}) \sim \sigma^2 \).

### 3.3.1 Constraint of the \( \lambda_j \)'s at the optimum

Similarly to EL, the sum of a linear combination of the \( \lambda_j \) is set to zero. Asymptotically, the \( \lambda_j \) is a function of the common mean \( \mu \) and thus, by differentiating the sum (third equation of (A.4)) with respect to \( \mu \), we obtain

\[ \hat{\mu} = \left[ \sum_{j=1}^{2} \frac{n_j}{s_j^2(\bar{x}_j)} \right]^{-1} \sum_{j=1}^{2} \frac{n_j \bar{x}_j}{s_j^2(\bar{x}_j)}. \]

This estimate of the common mean is the same as in (12 for the two samples case) only that different estimators of the variances are used. However the differentiation is small, when the sample sizes are large.
3.3.2 Power of the exponential empirical likelihood

Closing the examination of EEL, the asymptotic distribution of the two-sample test statistic under an assumed form of $H_1$ is here examined. The test statistic, asymptotically, is the sum of the one sample tests from the common mean

$$
\Lambda = \sum_{j=1}^{2} \frac{n_j (\bar{x}_j - \mu)^2}{s_j^2(\bar{x}_j)}
$$

(20)

and we show that this follows the $\chi^2$ distribution with 1 degree of freedom when $H_0$ is true. In the one sample case the form of the asymptotic test statistic (20) can be written as

$$
\Lambda = n \frac{(\bar{x} - \mu)^2}{s^2(\bar{x})} = n \left( \frac{n - 1}{n} s^2 \right)^{-1} (\bar{x} - \mu)^2 = T \left( 1 - \frac{1}{n} \right)^{-1},
$$

where $T$ is the one sample t-test statistic and $s^2$ is the unbiased sample variance estimate. Therefore, it is clear that the leading terms of the asymptotic form of the two statistics (EL and EEL test statistics) are the same. However, a comparison between the two likelihoods using simulated data under different situations when two samples are used, is described in Section 5. The asymptotic test statistic (20) is written as

$$
\Lambda = \sum_{j=1}^{2} T_j \left( 1 - \frac{1}{n_j} \right)^{-1}
$$

where $T_j^2$ is the one sample t-test from the true common mean of the $j$-th sample. We assume that, under the alternative hypothesis, each mean $\mu_j$ deviates from the common mean $\mu$ by a quantity which is a function of the sample variance and the sample size of each sample plus a constant vector. Let us, moreover, assume that we can write the mean under $H_1$ as a function of the variance and of the sample size

$$
\mu_j = \mu + \frac{\sigma_j^{1/2}}{\sqrt{n_j}} \tau_j.
$$

(21)

The asymptotic form of the test statistic under $H_1$ is

$$
\Lambda = \sum_{j=1}^{2} \frac{n_j (\bar{x}_j - \mu - \frac{\sigma_j^{1/2}}{\sqrt{n_j}} \tau_j)^2}{s_j^2(\bar{x}_j)} \left( 1 - \frac{1}{n_j} \right)^{-1} = \sum_{j=1}^{2} \left[ \frac{\sqrt{n_j}(\bar{x}_j - \mu)}{\sigma_j} - \tau_j \right]^2 \left( 1 - \frac{1}{n_j} \right)^{-1},
$$

since $s^2(\mu) \xrightarrow{p} \sigma^2$. The non-centrality parameter equals $\sum_{j=1}^{2} \tau_j^2$ since it is a weighted sum of independent scaled $\chi^2$ distributions. The scale factor $\left( 1 - \frac{1}{n_j} \right)^{-1}$ shows the relationship with the James test statistic. When the sample sizes are large, these factors disappear, and thus, the asymptotic distribution is the sum of 2 independent non-central $\chi^2$ distributions, where each of them has a non central parameter equal to $\tau_j^2$. Thus, to the leading term and under the alternative hypothesis, the distribution of the EEL test statistic is identical to that of EL.
4 Alternative calibrations of the test statistics

4.1 Calibration of the EL and EEL test statistics using the $F$ distribution

The test statistic of EL (11) and of EEL (19) is the same as the test statistic suggested by James (1954). For this reason, here it is suggested that, the exponential empirical likelihood test statistic should be calibrated with the corrected $\chi^2$ distribution suggested by James (1954). However, the correction factor of the $\chi^2$ distribution has a complex formula. To keep the complexity of the analysis low, the $t$ distribution of the Welch test is used.

In the two sample case, (11) and (19) are equal to the square of the Welch test statistic (1). The square root of (11) and (19) is calibrated against the $t$ distribution with the degrees of freedom given in (2). The choice of this calibration is not new. The $F$ instead of the $\chi^2$ distribution was also suggested by Owen (2001) in the one sample twovariate mean case and the same approach for the two sample twovariate mean vectors case was used by Tsagris et al. (2017).

4.2 Bootstrap calibration

The non-parametric bootstrap procedure applicable to all testing procedures, except for the WMW testing procedure can be described as follows.

1. Define the test statistic as $M$ and define $M_{\text{obs}}$ to be $M$ calculated for the available data $(x_{11}, \ldots, x_{1n_1})$ and $(x_{21}, \ldots, x_{2n_2})$ with sample means $\bar{x}_1$ and $\bar{x}_2$ and sample variances $s_1$ and $s_2$.

2. Transform the data so that the null hypothesis is true

\[ y_{1i} = x_{1i} - \bar{x}_1 + \hat{\mu}_c \quad \text{and} \quad y_{2i} = x_{2i} - \bar{x}_2 + \hat{\mu}_c, \]

where $\hat{\mu}_c = \left( \frac{n_1}{s_1} + \frac{n_2}{s_2} \right)^{-1} \left( \frac{n_1}{s_1} \bar{x}_1 + \frac{n_2}{s_2} \bar{x}_2 \right)$ is the estimated common mean under the null hypothesis.

3. Generate two bootstrap samples $y_{1}^*$ and $y_{2}^*$ by sampling with replacement from $y_1$ and $y_2$.

4. Define $M_b$ as the test statistic calculated for the bootstrap sample in step 3.

5. Repeat steps 3 and 4 $B$ times to generate bootstrap statistics $M_{b1}, \ldots, M_{bB}$ and calculate the bootstrap p-value as

\[ p\text{-value} = \frac{\sum_{i=1}^{B} \mathbb{1}(M_{bi} > M_{\text{obs}}) + 1}{B + 1}. \]

4.3 Exact p-value of the WMW test statistic

Instead of bootstrap calibration for the WMW test statistic, its exact p-value, numerically, is computed. Mann and Whitney (1947) proposed the equivalent test statistic $M_{n_1,n_2} =$
\[
\sum_{i=1}^{n_1} \# \{ j : x_{2j} < x_{1i} \}. \text{ Under } H_0, \text{ the probability generating function of the } M_{n_1,n_2} \text{ test statistic is given by (Van de Wiel et al., 1999)}
\]

\[
\sum_{k=0}^{n_1 n_2} \Pr (M_{n_1,n_2} = k) x^k = \frac{1}{(n_1 + n_2)} \prod_{i=n_1+1}^{n_1+n_2} (1 - x^i) \prod_{i=1}^{n_2} (1 - x^i). \tag{22}
\]

5 Simulations studies for the performance of the testing procedures applied to two sample means

The Welch t-test, the Wilcoxon-Mann-Whitney and the two non parametric likelihood procedures are the subject to comparison. For all test statistics except for the WMW, bootstrap calibration, with 499 bootstrap re-samples, is applied. Moreover, bootstrap with 999 re-samples was applied. However, the outcome did not differ, corroborating the results of Tsagris et al. (2017) who used 299 re-samples, and therefore indicating, that there is no need for high numbers of bootstrap replications. In addition, the reduction in the computational burden was significant. For the WMW the exact p-value is calculated. Furthermore, for the non parametric likelihood procedures, calibration with the F distribution was examined. All tests, except for the WMW were self implemented and are available in the R package Rfast (Papadakis et al., 2018).

The estimated type I error for all cases is based on 1000 simulations. When the estimated probability of type I error falls within (0.0365, 0.0635) (95% confidence interval for the true probability of type I error based on 1000 simulations), it provides evidence that the test attains the correct probability of type I error. The estimated power of the testing procedures that were size correct is, moreover, investigated. In order for the power of a test to be meaningful and comparable, the test has to be size correct. Otherwise, comparisons between tests which do not have the same probability of type I error, are not valid.

5.1 The case scenarios

The 12 case scenarios are presented in Table 1, while figure 1 visualises the densities. The distributions used are Normal \( (N(\mu, \sigma)) \), Beta \( (Be(\alpha, \beta)) \), folded normal \( (FN(\mu, \sigma)) \), Gamma \( (Ga(\alpha, \beta)) \), Laplace \( (La(\mu, \beta)) \), von Mises \( (vM(\mu, \kappa)) \), negative binomial \( (NB(\mu, N)) \) and Poisson \( (Pois(\lambda)) \).

With the exception of the first scenario, the distributions are not normal. They are skewed and have kurtosis totally different from that of the normal distribution. In some cases they are bimodal. Secondly, distributions in constrained spaces are used, such as the beta, the gamma and the von Mises. The motivation behind these choices was to show that the t-test is robust, even with small or large deviations from normality. The t-test relies on the assumption that the sample means are normally distributed. As it is known, the central limit theorem guarantees an asymptotic normal distribution for the sample mean of data that follow any distribution. As for the constrained support, it was desirable to highlight that in the whole of \( \mathbb{R} \) or any constrained subspace of it, the t-test is still valid and robust. Hence, one needs not make hard attempts to create a test for each case, parametric or non parametric.
A third axis of comparison is the computational time required for all the aforementioned procedures. This is also important with gene expression data, where the number of tests performed is in the order of tens of thousands. This point becomes evident during the comparison of the testing procedures when applied on the real data.

| Scenario | Sample 1 | Sample 2 |
|----------|----------|----------|
| (a)      | $N(3,4)$ | $N(3,0.5)$ |
| (b)      | $Be(3,4)$ | $N(3/7,0.5)$ |
| (c)      | $0.5Be(1.5) + 0.5Be(5,2)$ | $N(37/84,0.5)$ |
| (d)      | $Be(3,4)$ | $0.5N(-11/7,0.5) + 0.5N(17/7,0.5)$ |
| (e)      | $FN(3,4)$ | $Ga(4.049335,1)$ |
| (f)      | $0.5Ga(3,3) + 0.5Ga(8,1)$ | $Ga(18,4)$ |
| (g)      | $Be(1.3,1.3)$ | $0.4Be(0.9,0.9) + 0.6Be(12,12)$ |
| (h)      | $La(0,1)$ | $0.4Be(-2,2) + 0.6N(3,1)$ |
| (i)      | $vM(\mu = 2, \kappa = 10)$ | $vM(\mu = 2, \kappa = 5)$ |
| (j)      | $Skel(\lambda_1 = 8, \lambda_2 = 8)$ | $0.5NB(\mu = -5, N = 5) + 0.5NB(\mu = 5, N = 5)$ |
| (k)      | $NB(\mu = 5, N = 12)$ | $NB(\mu = 5, N = 4)$ |
| (l)      | $NB(\mu = 5, N = 10)$ | $0.6Pois(2) + 0.4Pois(9.5)$ |

Table 1: Information about the example pairs of datasets used. The sample size and the number of components for each composition is given.

5.2 Type I error

Tables 2 and 3 contain the estimated type I errors of all testing procedures and calibrations for all case scenarios. The common observation from both Tables is that the WMW test almost always produces an inflated type I error, even when both populations are normal (Scenario (a)). This is true regardless of the asymptotic or the exact p-value being computed. The Welch test is the only test that performs very well on most scenarios and sample sizes. The EL and EEL suffer from inflated type I errors, and especially EEL.

Table 4 aggregates the estimated type I error for all scenarios. The WMW, regardless of the asymptotic or the exact p-value being calculated attained the type I error only in a few cases. EEL with the $\chi^2$ approximation was also inadequate. We remind the reader that unlike EL, EEL is not Bartlett correctable (Jing and Wood, 1996). Calibration with the $t$ distribution helped both likelihoods improve, with the EEL almost doubling the proportion of times it was size correct. EL with $t$ calibration and the Welch test had the highest proportions. Finally, bootstrap calibration helped EEL perform even better, but changes only slightly the results of EEL with $t$ calibration and the Welch test.

This means, that for examining the power of the testing procedures we will focus only on the EL, EEL and Welch tests with bootstrap calibration.

5.3 Power of the Welch t-test, EL and EEL with bootstrap calibration

Only the bootstrap calibrated procedures are considered, since the testing procedures had to be applied on size correct tests. To reduce the page limit, the cases with the largest sample sizes $(50, 100)$ is presented. By examining Figure 2, one common overall conclusion can be drawn. The Welch t-test has similar power levels to the EL and EEL competitors.
5.4 Asymptotic, bootstrap and exact p-value of the Welch test

When the sample sizes are really low, say 10, an exact Welch test might seem a better alternative to the asymptotic or the bootstrap calibrated Welch test. The term *exact* stems from computing the p-value based on permutations, where all permutations have been considered.

Detecting significant differences, with so small sample sizes, is really difficult. Hence, we will concentrate our attention to the type I error. Also, we will ease the difficulty of the scenario to equal sample sizes of 10. The 12 previous scenarios will be used again and the Welch test will be subject to comparison. The asymptotic p-value, the bootstrap based p-value and the exact p-value are reported in Table 5.

To our surprise, the estimated type I error based on the asymptotic p-value was outside the permissible limits only once. The estimated type I error of the bootstrap based p-value was outside the permissible limits twice. Finally, the estimated type I error based on the exact p-value fell outside the acceptable limits half of the times.

6 Real gene expression datasets

Time is important for researchers and practitioners. Why wait hours instead of minutes or even seconds? In order to highlight this, 4 gene expression datasets (GSE annotated) were downloaded from BioDataome (Lakiotaki et al., 2018). From a biological point of interest, the data have been uniformly preprocessed, curated and automatically annotated. We use them to evaluate the computational cost of all testing procedures except for the WMW test which did not perform equally well. The time (in seconds) required by each testing procedure to perform tens of thousands of hypothesis tests is computed. Bootstrap calibration, with only 499 resamples, is only applicable to the Welch t-test due to the excessively high computational cost required by the EEL and EL. The Welch t-tests requires microseconds for a single test and hundreds of seconds for 54,675. It becomes obvious that EEL and EL with bootstrap calibration would require a few thousands of seconds and a few hours respectively. We also compute the computational cost of the (new) efficient bootstrap calibration of the Welch t-test (Chatzipantsiou et al., 2019). The execution times appear on Table 6. All functions are self written with no C++ implementation of any test and no use of parallel in order to be as fair as possible.
Figure 1: The densities for each case scenario.
Table 2: Estimated probability of Type I error using different tests and a variety of calibrations. The nominal level of the Type I error was equal to 0.05. The numbers in bold indicate that the estimated probability was within the acceptable limits (0.0365, 0.0635).

| Sample sizes | EL(χ²) | EEL(χ²) | EL(t) | EEL(t) | Welch | WMW | EL(boot) | EEL(boot) | Welch(boot) | WMW(exact) |
|--------------|--------|---------|-------|--------|-------|-----|----------|-----------|-------------|------------|
| (20, 30)     | 0.065  | 0.076  | 0.050 | 0.061  | 0.048 | 0.120| 0.046    | 0.046     | 0.047       | 0.120      |
| (20, 40)     | 0.081  | 0.085  | 0.065 | 0.073  | 0.065 | 0.138| 0.064    | 0.062     | 0.056       | 0.138      |
| (20, 30)     | 0.060  | 0.069  | 0.049 | 0.057  | 0.052 | 0.121| 0.048    | 0.048     | 0.049       | 0.121      |
| (20, 40)     | 0.048  | 0.052  | 0.039 | 0.043  | 0.039 | 0.122| 0.042    | 0.044     | 0.038       | 0.122      |
| (20, 50)     | 0.057  | 0.062  | 0.055 | 0.056  | 0.052 | 0.149| 0.052    | 0.052     | 0.051       | 0.150      |
| (20, 30)     | 0.058  | 0.066  | 0.053 | 0.057  | 0.051 | 0.033| 0.052    | 0.052     | 0.048       | 0.033      |
| (20, 40)     | 0.043  | 0.049  | 0.039 | 0.047  | 0.038 | 0.024| 0.038    | 0.034     | 0.038       | 0.024      |
| (20, 50)     | 0.055  | 0.062  | 0.051 | 0.056  | 0.052 | 0.038| 0.047    | 0.051     | 0.038       | 0.038      |
| (20, 60)     | 0.060  | 0.069  | 0.060 | 0.067  | 0.058 | 0.029| 0.060    | 0.060     | 0.057       | 0.029      |
| (20, 30)     | 0.076  | 0.083  | 0.066 | 0.074  | 0.062 | 0.071| 0.065    | 0.055     | 0.071       |           |
| (20, 40)     | 0.062  | 0.083  | 0.053 | 0.073  | 0.048 | 0.068| 0.052    | 0.051     | 0.068       |           |
| (20, 50)     | 0.053  | 0.058  | 0.047 | 0.054  | 0.047 | 0.052| 0.043    | 0.044     | 0.052       |           |
| (20, 60)     | 0.065  | 0.073  | 0.060 | 0.067  | 0.058 | 0.072| 0.060    | 0.059     | 0.072       |           |
| (20, 70)     | 0.043  | 0.066  | 0.040 | 0.063  | 0.041 | 0.063| 0.038    | 0.043     | 0.063       |           |
| (20, 30)     | 0.050  | 0.052  | 0.037 | 0.037  | 0.049 | 0.102| 0.041    | 0.033     | 0.102       |           |
| (20, 40)     | 0.055  | 0.057  | 0.039 | 0.044  | 0.052 | 0.113| 0.048    | 0.039     | 0.113       |           |
| (20, 50)     | 0.054  | 0.054  | 0.042 | 0.042  | 0.054 | 0.108| 0.050    | 0.045     | 0.107       |           |
| (20, 60)     | 0.045  | 0.047  | 0.031 | 0.034  | 0.042 | 0.195| 0.044    | 0.036     | 0.195       |           |
| (20, 70)     | 0.050  | 0.050  | 0.042 | 0.045  | 0.048 | 0.195| 0.042    | 0.037     | 0.195       |           |
| (20, 30)     | 0.058  | 0.074  | 0.048 | 0.062  | 0.043 | 0.079| 0.043    | 0.039     | 0.079       |           |
| (20, 40)     | 0.055  | 0.076  | 0.045 | 0.060  | 0.047 | 0.098| 0.043    | 0.042     | 0.098       |           |
| (20, 50)     | 0.051  | 0.057  | 0.047 | 0.048  | 0.046 | 0.086| 0.047    | 0.045     | 0.086       |           |
| (20, 60)     | 0.039  | 0.045  | 0.034 | 0.041  | 0.032 | 0.084| 0.030    | 0.028     | 0.084       |           |
| (20, 70)     | 0.041  | 0.057  | 0.038 | 0.055  | 0.041 | 0.124| 0.038    | 0.039     | 0.125       |           |
| (20, 30)     | 0.070  | 0.074  | 0.047 | 0.054  | 0.058 | 0.161| 0.049    | 0.049     | 0.161       |           |
| (20, 40)     | 0.071  | 0.081  | 0.055 | 0.064  | 0.066 | 0.195| 0.053    | 0.051     | 0.195       |           |
| (20, 50)     | 0.055  | 0.059  | 0.050 | 0.051  | 0.055 | 0.159| 0.051    | 0.050     | 0.159       |           |
| (20, 60)     | 0.053  | 0.059  | 0.044 | 0.049  | 0.048 | 0.197| 0.043    | 0.044     | 0.197       |           |
| (20, 70)     | 0.067  | 0.076  | 0.063 | 0.068  | 0.072 | 0.275| 0.067    | 0.069     | 0.276       |           |
Table 3: Estimated probability of Type I error using different tests and a variety of calibrations. The nominal level of the Type I error was equal to 0.05. The numbers in bold indicate that the estimated probability was within the acceptable limits (0.0365, 0.0635). The “-” indicates that the exact p-value for the WMW test statistic was not computed because of ties in the data.

| Scenario | Sample sizes | EL(\(\chi^2\)) | EEL(\(\chi^2\)) | EL(t) | EEL(t) | Welch | WMW | EL(boot) | EEL(boot) | Welch(boot) | WMW(exact) |
|----------|--------------|-----------------|-----------------|-------|--------|-------|-----|----------|-----------|-------------|-------------|
| (g)      | (20, 30)     | 0.055           | 0.064           | 0.046 | 0.054  | 0.046 | 0.062| 0.045    | 0.046     | 0.044       | 0.062       |
|          | (20, 40)     | 0.069           | 0.088           | 0.055 | 0.075  | 0.061 | 0.085| 0.054    | 0.057     | 0.055       | 0.085       |
|          | (30, 40)     | 0.047           | 0.054           | 0.044 | 0.047  | 0.045 | 0.059| 0.046    | 0.047     | 0.047       | 0.059       |
|          | (30, 50)     | 0.059           | 0.069           | 0.052 | 0.062  | 0.057 | 0.074| 0.051    | 0.051     | 0.052       | 0.074       |
|          | (50, 100)    | 0.069           | 0.085           | 0.064 | 0.082  | 0.067 | 0.085| 0.059    | 0.060     | 0.058       | 0.086       |
| (h)      | (20, 30)     | 0.056           | 0.071           | 0.054 | 0.064  | 0.047 | 0.119| 0.047    | 0.048     | 0.045       | 0.119       |
|          | (20, 40)     | 0.060           | 0.084           | 0.048 | 0.069  | 0.052 | 0.153| 0.048    | 0.046     | 0.045       | 0.153       |
|          | (30, 40)     | 0.058           | 0.066           | 0.051 | 0.058  | 0.056 | 0.174| 0.049    | 0.052     | 0.054       | 0.174       |
|          | (30, 50)     | 0.051           | 0.062           | 0.038 | 0.054  | 0.045 | 0.185| 0.041    | 0.041     | 0.040       | 0.185       |
|          | (50, 100)    | 0.056           | 0.077           | 0.052 | 0.070  | 0.055 | 0.287| 0.051    | 0.050     | 0.048       | 0.289       |
| (i)      | (20, 30)     | 0.051           | 0.062           | 0.049 | 0.058  | 0.051 | 0.059| 0.048    | 0.048     | 0.045       | 0.059       |
|          | (20, 40)     | 0.052           | 0.071           | 0.043 | 0.056  | 0.037 | 0.055| 0.039    | 0.040     | 0.034       | 0.055       |
|          | (30, 40)     | 0.062           | 0.067           | 0.059 | 0.062  | 0.057 | 0.056| 0.065    | 0.055     | 0.052       | 0.056       |
|          | (30, 50)     | 0.058           | 0.067           | 0.055 | 0.059  | 0.052 | 0.067| 0.042    | 0.044     | 0.046       | 0.067       |
|          | (50, 100)    | 0.058           | 0.079           | 0.050 | 0.073  | 0.052 | 0.069| 0.050    | 0.048     | 0.047       | 0.069       |
| (j)      | (20, 30)     | 0.057           | 0.062           | 0.049 | 0.058  | 0.051 | 0.059| 0.043    | 0.042     | 0.044       | 0.054       |
|          | (20, 40)     | 0.059           | 0.080           | 0.049 | 0.069  | 0.053 | 0.071| 0.049    | 0.046     | 0.045       | 0.054       |
|          | (30, 40)     | 0.059           | 0.064           | 0.054 | 0.058  | 0.053 | 0.066| 0.050    | 0.050     | 0.051       | 0.051       |
|          | (30, 50)     | 0.050           | 0.062           | 0.045 | 0.053  | 0.045 | 0.063| 0.043    | 0.042     | 0.041       | 0.041       |
|          | (50, 100)    | 0.061           | 0.073           | 0.055 | 0.070  | 0.058 | 0.082| 0.053    | 0.053     | 0.054       | 0.054       |
| (k)      | (20, 30)     | 0.055           | 0.062           | 0.049 | 0.058  | 0.051 | 0.059| 0.043    | 0.042     | 0.044       | 0.066       |
|          | (20, 40)     | 0.059           | 0.080           | 0.049 | 0.069  | 0.053 | 0.071| 0.049    | 0.046     | 0.045       | 0.066       |
|          | (30, 40)     | 0.059           | 0.064           | 0.054 | 0.058  | 0.053 | 0.066| 0.050    | 0.050     | 0.051       | 0.050       |
|          | (30, 50)     | 0.050           | 0.062           | 0.045 | 0.053  | 0.045 | 0.063| 0.043    | 0.042     | 0.041       | 0.054       |
|          | (50, 100)    | 0.061           | 0.073           | 0.055 | 0.070  | 0.058 | 0.082| 0.053    | 0.053     | 0.054       | 0.054       |
| (l)      | (20, 30)     | 0.055           | 0.062           | 0.049 | 0.058  | 0.051 | 0.059| 0.043    | 0.042     | 0.044       | 0.073       |
|          | (20, 40)     | 0.055           | 0.075           | 0.038 | 0.062  | 0.056 | 0.137| 0.039    | 0.037     | 0.042       | 0.073       |
|          | (30, 40)     | 0.056           | 0.061           | 0.048 | 0.053  | 0.049 | 0.134| 0.044    | 0.046     | 0.046       | 0.073       |
|          | (30, 50)     | 0.057           | 0.069           | 0.046 | 0.058  | 0.049 | 0.160| 0.049    | 0.044     | 0.045       | 0.073       |
|          | (50, 100)    | 0.057           | 0.073           | 0.052 | 0.067  | 0.056 | 0.263| 0.058    | 0.059     | 0.057       | 0.073       |
Table 4: Number of times each testing procedure attained the correct size.

| Testing procedure | EL(\chi^2) | EEL(\chi^2) | EL(t) | EEL(t) | Welch |
|-------------------|------------|-------------|-------|--------|-------|
|                   | 49/60      | 23/60       | 55/60 | 39/60  | 54/60 |
| Testing procedure | WMW EL(boot) | EEL(boot) | Welch(boot) | WMW(exact) |
|                   | 13/60      | 56/60       | 56/60 | 57/60  | 8/45  |

Table 5: Estimated type I error of the Welch test using the asymptotic p-value, the bootstrap p-value and the exact p-value calculated from all possible permutations. The numbers in bold indicate that the estimated probability was within the acceptable limits (0.0365, 0.0635).

| Scenario | Asymptotic | Bootstrap | Exact |
|----------|------------|-----------|-------|
| Scenario (a) | 0.056 | 0.068 | 0.077 |
| Scenario (c) | 0.053 | 0.053 | 0.057 |
| Scenario (e) | 0.052 | 0.053 | 0.054 |
| Scenario (g) | 0.054 | 0.058 | 0.064 |
| Scenario (i) | 0.055 | 0.063 | 0.060 |
| Scenario (k) | 0.053 | 0.055 | 0.073 |

Table 6: Computational time in seconds required by each method. The Welch(boot*) stands for the efficient bootstrap methodology for the Welch t-test (Chatzipantsiou et al., 2019).

| Datasets | Sample sizes | # Probesets | EL(\chi^2) | EEL(\chi^2) | Testing procedures |
|----------|--------------|-------------|------------|-------------|--------------------|
| GSE12276 | (102, 102)   | 54,675      | 126.60     | 6.41        | 0.12               |
| GSE50081 | (91, 90)     | 54,675      | 128.75     | 5.50        | 0.12               |
| GSE62452 | (65, 65)     | 33,297      | 55.87      | 2.84        | 0.06               |
| GSE28735 | (45, 45)     | 33,297      | 54.44      | 2.41        | 0.07               |
| Totals   |              |             | 365.66     | 17.16       | 0.37               |

17
Figure 2: The estimated power of EL, EEL and Welch, with bootstrap calibration, for each case scenario.
7 Conclusions

Two non-parametric likelihood methods, were examined both from a theoretical and computational point of view. A third non parametric method tested was the Wilcoxon-Mann-Whitney test and the only parametric test used was the Welch t-test. We show that, asymptotically, the non-parametric likelihood test statistic converge to the Welch t-test. Thus, when the sample size(s) is (are) large, the non parametric likelihood testing procedures are equivalent to the Welch t-test. Given their relationship, we calibrate the non parametric likelihoods using the \( t \) distribution, instead of the usual \( \chi^2 \) distribution. As a second alternative calibration, bootstrap is applied.

However, the greatest concerned felt with the finite sample, and especially the small size case. Moreover, we were concerned with the robustness of the Welch test against large deviations from the normal, or even symmetric, distributions. Further, we examined distributions of limited support and distributions of discrete data. In order to examine the type I error of all testing procedures, we conduct simulation studies covering various scenarios of distributions.

In most of the cases, the nominal level of the type I error was not attained by the procedures, unless the sample sizes were large. The use of the \( t \) distribution for calibration of the EL and EEL test statistics was tested. The results show an improvement, mainly for the EEL, compared to the use of the \( \chi^2 \). However, these alternative calibrations do not improve the size of the test substantially. Bootstrap calibrated methods, on the other hand, perform well in almost all cases. In fact, for the very small sample size cases, bootstrap calibration plays an important role in correcting the test size. The computational cost of this re-sampling procedure is evident only for the EL and EEL. The Welch test requires no numerical optimisation, only vector calculations and, by using the efficient method of (Chatzipantsiou et al., 2019), the computational cost is further reduced. With relatively larger sample sizes though, Welch test with no bootstrap calibration is size correct and computationally extremely cheap. In the era of big data and vast availability of data, computational time has become an important asset, and therefore, the above could be an evidence against the use of these non-parametric likelihoods.

A conclusion that we reach is that the characteristics of the distribution (continuous, discrete, constrained or unconstrained support) do not differentiate the results. Even when one expects that for cases like percentages the Welch t-test not to perform well, this is not happening. Let us remind ourselves though, that the means of the samples are assumed to be normally distributed, and not the distributions of the data themselves, due to the central limit theorem. An advantage of this theorem is that its validity starts from small sample sizes and this is the reason why the above conclusion stands.

Another disadvantage of EL concerns with the convex hull. It is not necessary for the candidate common mean to lie within the convex hull of each sample. If, for example, the two samples are separated, then clearly there is no intersection between them. This, of course, is an issue that commonly rises in higher dimensions. A similar situation occurs with EEL.

Questioning on whether there will always be a solution in the two sample problem, we conclude to the fact that in same cases, when EL did not find the common mean, EEL was implemented successfully. However, this was not universally observed.

The WMW test was shown to be highly non accurate in terms of type I error, even if the
exact p-value was calculated. Furthermore, the exact p-value could not be computed when discrete data were examined. Both arguments show that WMW should not be considered as a competing non parametric alternative to the Welch t-test.

The ultimate conclusion to which this research reaches is that the Welch test statistic with or without bootstrap calibration, is to be preferred. It is computationally cheap, simpler to program, understand and, most importantly, it is free of the convex hull limitations. Even if one programs EL in C++, with or without bootstrap calibration, there is little to gain. Our simulations showed that the Welch t-test is highly robust to various situations and is not affected by the sample space of the data nor the shape of the distributions. For this reason, we are proposing the use of the Welch t-test for comparing two sample means.

Appendix

Proof of the EL test statistic

We have 2 samples of sizes $n_1$ and $n_2$. The 2 constraints are

$$f_j(\lambda_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \left\{ [1 + \lambda_j (x_{ji} - \mu)]^{-1} (x_{ij} - \mu) \right\} = 0 \quad (j = 1, 2). \quad (A.1)$$

The first derivative of $f_j(\lambda_j)$ with respect to $\lambda_j$ is

$$\frac{\partial f_j(\lambda_j)}{\partial \lambda_j} = -\frac{1}{n_j} \sum_{i=1}^{n_j} \left\{ [1 + \lambda_j (x_{ji} - \mu)]^{-2} (x_{ji} - \mu) (x_{ji} - \mu) \right\}.$$

For each of the 2 constraints, the Maclaurin expansion for the vector $\lambda_j$ in (A.1) has to be applied, since when the null hypothesis is true, $\lambda_j$ is the zero vector. Keeping the leading term only, $f_j(\lambda_j)$ can be written as

$$\lambda_j = s_j^2 (\mu)^{-1} (\bar{x}_j - \mu) + O_p\left( n_0^{-1/2} \right), \quad (A.2)$$

where $n_0 = \min_{1 \leq i \leq 2} n_i$. The probabilities of each of the $j$ samples have the following form

$$p_{ji} = \frac{1}{n_j} [1 + \lambda_j (x_{ji} - \mu)]^{-1}.$$

The log-likelihood ratio test statistic has the following form

$$\Lambda = -2 \sum_{j=1}^{2} \sum_{i=1}^{n_j} \log n_j p_{ji} = 2 \sum_{j=1}^{2} \sum_{i=1}^{n_j} \log [1 + \lambda_j (x_{ji} - \mu)] = 2 \sum_{j=1}^{2} g_j(\lambda_j). \quad (A.3)$$
We will again use Maclaurin series for each of the λ_j in the g_j(λ_j) for j = 1, 2. The first and second derivatives with respect to λ_j are

$$\frac{\partial g_j(\lambda_j)}{\partial \lambda_j} = 2 \sum_{i=1}^{n_j} \left\{ [1 + \lambda_j (x_{ji} - \mu)]^{-1} (x_{ji} - \mu) \right\}$$
$$\frac{\partial^2 g_j(\lambda_j)}{\partial \lambda_j^2} = 2 \sum_{i=1}^{n_j} \left\{ [1 + \lambda_j (x_{ji} - \mu)]^{-2} (x_{ji} - \mu) (x_{ji} - \mu) \right\}.$$

Finally, keeping the first two leading terms and using (A.2) we have

$$\Lambda = 0 + s_j^2 (\mu)^{-1} (\bar{x}_j - \mu) 2n_j (\bar{x}_j - \mu) - \frac{\left[ s_j^2 (\mu)^{-1} (\bar{x}_j - \mu)^2 \right]^2}{2n_js_j^2 (\mu)} + O_p \left( n_0^{-1/2} \right).$$

Asymptotically, the final term is the sum of 2 independent χ^2 variables under H_0 (that each population mean is μ). This is true, since s^2(μ) → σ^2, where σ^2 is the population variance. The degrees of freedom are equal to 1 (Owen, 2001).

**Proof of the EEL test statistic**

Here, the two sample means case but in a simpler form will be presented, using a similar approach to the one Jing and Robinson (1997) followed. The three constraints are

$$\left( \sum_{j=1}^{n_1} e^{\lambda_1 x_{1j}} \right)^{-1} \left( \sum_{i=1}^{n_1} x_{i1} e^{\lambda_1 x_{1i}} \right) - \mu = 0$$
$$\left( \sum_{j=1}^{n_2} e^{\lambda_2 x_{2j}} \right)^{-1} \left( \sum_{i=1}^{n_2} x_{i2} e^{\lambda_2 x_{2i}} \right) - \mu = 0$$
$$n_1 \lambda_1 + n_2 \lambda_2 = 0.$$  (A.4)

The proof of the third constraint comes from minimisation of (19) with respect to μ. We can equate the first two constraints of (A.4):

$$\left( \sum_{j=1}^{n_1} e^{\lambda_1 x_{1j}} \right)^{-1} \left( \sum_{i=1}^{n_1} x_{i1} e^{\lambda_1 x_{1i}} \right) = \left( \sum_{j=1}^{n_2} e^{\lambda_2 x_{2j}} \right)^{-1} \left( \sum_{i=1}^{n_2} x_{i2} e^{\lambda_2 x_{2i}} \right).$$  (A.5)

Also, we can write the third constraint of (A.4) as λ_2 = -n_1/n_2 λ_1 and thus we can rewrite (A.5) as

$$\left( \sum_{j=1}^{n_1} e^{\lambda x_{1j}} \right)^{-1} \left( \sum_{i=1}^{n_1} x_{i1} e^{\lambda x_{1i}} \right) = \left( \sum_{j=1}^{n_2} e^{-\frac{n_1}{n_2} \lambda x_{2j}} \right)^{-1} \left( \sum_{i=1}^{n_2} x_{i2} e^{-\frac{n_1}{n_2} \lambda x_{2i}} \right).$$

This trick allows us to avoid the estimation of the common mean. However, this is not applicable in the empirical likelihood method. Instead of minimisation of the sum of the one-sample test statistics from the common mean, the probabilities by searching for the λ can be
defined, leading the last equation to hold true. The third constraint of (A.4) is a convenient constraint whose validity is proved in the Appendix. Jing and Robinson (1997) discussed this issue and mention that even if it is a simple constraint, it does not lead to second-order accurate confidence intervals, unless the two sample sizes are equal.

The asymptotic distribution of the EEL test statistic in the two-sample case will be presented, following the same path as for EL. The special case of two samples is also explained due to a simpler version and the one-sample case is not difficult to derive. Jing and Robinson (1997) showed the two-sample case in the univariate case using a different direction. Assuming $k$ samples and $\mu$ being the common mean as before, the form of the probabilities is

$$p_{ji} = \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} e^{\lambda_j x_{ji}} = \left[ \sum_{m=1}^{n_j} e^{\lambda_j (x_{jm} - \mu)} \right]^{-1} e^{\lambda_j (x_{ji} - \mu)} \quad (j = 1, 2).$$

Again, it is necessary to impose a constraint onto the Lagrangian parameters $\lambda_s$. The constraints for the exponential tilting are

$$f_j(\lambda_j) = \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} \left( \sum_{i=1}^{n_j} x_{ji} e^{\lambda_j x_{ji}} \right) - \mu = 0 \quad (j = 1, 2).$$

(A.6)

The first derivative of $f_j(\lambda_j)$ with respect to $\lambda_j$ is

$$\frac{\partial f_j(\lambda_j)}{\partial \lambda_j} = - \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-2} \left( \sum_{i=1}^{n_j} x_{ji} e^{\lambda_j x_{ji}} \right)^2 + \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} \sum_{i=1}^{n_j} x_{ji}^2 e^{\lambda_j x_{ji}}$$

Note that here the constraint can be written as $\sum_{i=1}^{n} p_i x_i - \mu = 0$, whereas in the empirical likelihood the constraint must stay $\sum_{i=1}^{n} p_i (x_i - \mu) = 0$. In the exponential tilting no difference is derived. However, in the empirical likelihood a difference is indicated, since the resulting probabilities in the empirical likelihood depend upon $\mu$.

Under $H_0$ the value of each of the $\lambda_j$ is equal to zero. By substituting $\lambda$ with its approximation using the Maclaurin series we get

$$\lambda_j = - s_j^2 (\bar{x}_j)^{-1} (\bar{x}_j - \mu) + o_p \left( n_j^{-1/2} \right),$$

(A.7)

where the $s_j^2 (\bar{x}_j)$ denotes the sample biased variance. Finally the $j$-th constraint, keeping the leading term, becomes The log-likelihood ratio test statistic is

$$\Lambda = -2 \sum_{j=1}^{2} \sum_{i=1}^{n_j} \log \left[ n_j \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} e^{\lambda_j x_{ji}} \right] = 2 \sum_{j=1}^{2} g_j(\lambda_j),$$

(A.8)

where

$$g_j(\lambda_j) = n_j \log \sum_{i=1}^{n_j} e^{\lambda_j x_{ji}} - \sum_{i=1}^{n_j} \lambda_j x_{ji} - \sum_{i=1}^{n_j} \log n_j$$

(A.9)

We can now apply the Maclaurin series to (A.9). The first and second derivative of $g_j(\lambda_j)$ with
respect to $\lambda_j$ are

$$\frac{\partial g_j(\lambda_j)}{\partial \lambda_j} = n_j \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} \left( \sum_{i=1}^{n_j} x_{ji} e^{\lambda_j x_{ji}} \right) - \sum_{i=1}^{n_j} x_{ji}$$

$$\frac{\partial^2 g_j(\lambda_j)}{\partial \lambda_j^2} = -n_j \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-2} \left( \sum_{i=1}^{n_j} x_{ji} e^{\lambda_j x_{ji}} \right)^2 + n_j \left( \sum_{m=1}^{n_j} e^{\lambda_j x_{jm}} \right)^{-1} \sum_{i=1}^{n_j} x_{ji}^2 e^{\lambda_j x_{ji}}$$

Then, by keeping the two leading terms in the Maclaurin series, (A.9) can be written as

$$g_j(\lambda_j) \simeq \frac{n_j}{2} \frac{(\bar{x}_j - \mu)^2}{s^2_j (\bar{x}_j)}.$$ 

Finally (A.8) can be written as

$$\Lambda = 2 \sum_{j=1}^{2} g_j(\lambda_j) = \sum_{j=1}^{2} n_j \frac{(\bar{x}_j - \mu)^2}{s^2_j (\bar{x}_j)} + O_p \left( n^{-1/2}_0 \right),$$

which is asymptotically a $\chi^2$ distribution with 1 degree of freedom. This is true by combining Slutsky’s theorem with $s^2(\bar{x}) \xrightarrow{p} \sigma^2$, the true variance.

Similarly to EL the sum of a linear combination of the $\lambda$s is set to zero as presented in the previous Section. Asymptotically, the $\lambda$s is a function of the common mean $\mu$ and thus by differentiating the sum (third equation of (A.4)) with respect to $\mu$ we obtain

$$\hat{\mu} = \left[ \sum_{j=1}^{2} \frac{n_j}{s^2_j (\bar{x}_j)} \right]^{-1} \sum_{j=1}^{2} \frac{n_j \bar{x}_j}{s^2_j (\bar{x}_j)}.$$ 

This estimate of the common mean is the same as in (12) for the two samples case, with the only difference that different estimators of the variances are used, but when the sample sizes are large it makes little difference.

References

Amaral, G. J. and Wood, A. T. (2010). Empirical likelihood methods for two-dimensional shape analysis. *Biometrika*, 97(3):757–764.

Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(1):17–25.

Chatzipantsiou, C., Dimitriadis, M., Papadakis, M., and Tsagris, M. (2019). Extremely efficient permutation and bootstrap hypothesis tests using R. *Journal of Modern Applied Statistical Methods (To apper)*. arXiv preprint arXiv:1806.10947.

Chen, M., Wu, C., and Thompson, M. E. (2016). Mann–Whitney test with empirical likelihood methods for pretest–posttest studies. *Journal of Nonparametric Statistics*, 28(2):360–374.
Chen, S. (1994). Comparing empirical likelihood and bootstrap hypothesis tests. *Journal of Multivariate Analysis, 51*(2):277–293.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., et al. (2008). Annotating genomes with massive-scale rna sequencing. *Genome biology, 9*(12):R175.

Derrick, B., Toher, D., and White, P. (2016). Why Welch’s test is Type I error robust. *The Quantitative Methods in Psychology, 12*(1):30–38.

DiCiccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *The Annals of Statistics, 19*(2):1053–1061.

Dwivedi, A. K., Mallawaarachchi, I., and Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in medicine, 36*(14):2187–2205.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics, 9*(2):139–158.

Emerson, S. (2009). *Small sample performance and calibration of the Empirical Likelihood method*. PhD thesis, Stanford university.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies a paradox of statistical practice? *BMC Medical Research Methodology, 12*(1):78.

Fagerland, M. W. and Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary clinical trials, 30*(5):490–496.

James, G. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika, 41*(1/2):19–43.

Jing, B. (1995). Two-sample empirical likelihood method. *Statistics & probability letters, 24*(4):315–319.

Jing, B. and Robinson, J. (1997). Two-sample nonparametric tilting method. *Australian Journal of Statistics, 39*(1):25–34.

Jing, B. and Wood, A. (1996). Exponential empirical likelihood is not Bartlett correctable. *The Annals of Statistics, pages* 365–369.

Lakiotaki, K., Vorniotakis, N., Tsagris, M., Georgakopoulos, G., and Tsamardinos, I. (2018). Biodataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. *Database, 2018.*

Lazar, N. and Mykland, P. (1998). An evaluation of the power and conditionality properties of empirical likelihood. *Biometrika, 85*(3):523.

Liu, Y., Zou, C., and Zhang, R. (2008). Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters, 78*(5):548–556.
Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Owen, A. (2001). *Empirical likelihood*. Boca Raton: Chapman & Hall/CRC.

Papadakis, M., Tsagris, M., Dimitriadis, M., Papadakis, G., Burkardt, J., Zou, C., and Lakiotaki, K. (2018). *Rfast: A Collection of Efficient and Extremely Fast R Functions*. R package version 1.9.2.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.

Sudhir, P., Wang, Y.-G., and Ullah, I. (2018). A review of the Behrens-Fisher problem and some of its analogs: Does the same size fit all? *Revstat Statistical Journal*, To appear.

Tsagris, M., Preston, S., and Wood, A. T. A. (2017). Nonparametric hypothesis testing for equality of means on the simplex. *Journal of Statistical Computation and Simulation*, 87(2):406–422.

Van de Wiel, M., Di Bucchianico, A., and Van der Laan, P. (1999). Symbolic computation and exact distributions of nonparametric test statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(4):507–516.

Welch, B. L. (1947). The generalization of student’s problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Wheelan, S. J., Murillo, F. M., and Boeke, J. D. (2008). The incredible shrinking world of dna microarrays. *Molecular BioSystems*, 4(7):726–732.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Zhu, H., Zhou, H., Chen, J., Li, Y., Lieberman, J., and Styner, M. (2008). Adjusted exponentially tilted likelihood with applications to brain morphology. *Biometrics*, 65(3):919–927.