MDPI

Article

A Semiparametric Tilt Optimality Model

Chathurangi H. Pathiravasan [1] and Bhaskar Bhattacharya [2,*]

1 Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA
2 School of Mathematical and Statistical Sciences, Southern Illinois University Carbondale,
   Carbondale, IL 62901, USA
* Correspondence: bhaskar@siu.edu

Abstract: Practitioners often face the situation of comparing any set of k distributions, which may follow neither normality nor equality of variances. We propose a semiparametric model to compare those distributions using an exponential tilt method. This extends the classical analysis of variance models when all distributions are unknown by relaxing its assumptions. The proposed model is optimal when one of the distributions is known. Large-sample estimates of the model parameters are derived, and the hypotheses for the equality of the distributions are tested for one-at-a-time and simultaneous comparison cases. Real data examples from NASA meteorology experiments and social credit card limits are analyzed to illustrate our approach. The proposed approach is shown to be preferable in a simulated power comparison with existing parametric and nonparametric methods.

Keywords: constraints; exponential tilt; goodness-of-fit tests; information projection; Kullback–Leibler divergence; maximum entropy

1. Introduction

Classical applied statistical techniques often depend heavily on the assumption that observations are normally distributed. The benefit of this assumption is that it helps to produce exact inferences in many popular methods, such as, t-test, F-test, chi-squared tests, analysis of variance (ANOVA) models and multivariate analysis. In reality, however, observations often show departures from normality (or near-normality). Statistical texts address this issue and discuss remedies, such as the Box-Cox transformations of data to normality. Developing techniques that use fewer assumptions (normality or others) is an important area of statistical research. When comparing different populations, this paper proposes an alternative to the one-way ANOVA by relaxing some of the assumptions.

Previous study described experimental radar reflectivity data obtained from independent radars deployed during NASA's Tropical Rainfall Measuring Mission Kwajalein Experiment in the republic of the Marshall Islands on 15 July to 12 September 1999 [1]. The data are skewed, and we investigate whether the data from two radars are from identical populations. A credit limit data set [2] with three different education levels; graduate school, university and high school was obtained from the University of California Irvine repository. The data are skewed with high variability, and we studied whether the three credit limit populations are identical. Such data have unknown statistical distributions, and known statistical procedures may fail to work properly.

If $g_i(x)$ is the probability density function (pdf) of $N(\mu_i, \sigma^2)$ for $i = 1, \ldots, k$ [3], then one can write

$$g_i(x) = g_k(x) \exp(\alpha_i + \beta_i x), \quad i = 1, \ldots, k-1, \tag{1}$$

where

$$\alpha_i = \frac{\mu_k^2 - \mu_i^2}{2\sigma^2}, \quad \beta_i = \frac{\mu_i - \mu_k}{\sigma^2}, \quad i = 1, \ldots, k-1. \tag{2}$$

From (1), denoting $g_k(x)$ as a reference distribution, one can think of $g_i(x)$ as an exponential distortion or tilt of the reference. Furthermore, using (2), the test of equality of μ_i's (as needed in the one-way ANOVA) reduces to the test of equality of the β_i's to zero, or

$$H_0 : \mu_1 = \cdots = \mu_k \iff H_0 : \beta_1 = \cdots = \beta_{k-1} = 0. \tag{3}$$

Motivated by (3), the same authors proposed a generalization of (1) by replacing $g_k(x)$ with any pdf $g(x)$, x in the exponent of (1) with any known function $h(x)$ and considered all pdfs $g_i(x)$ that can be expressed as exponential tilt pdfs of $g(x)$. In this way, (1)–(2) are updated as

$$g_i(x) = g(x) \exp(\alpha_i + \beta_i h(x)), \tag{4}$$

where

$$\alpha_i = -\ln\left(\int_{-\infty}^{\infty} e^{\beta_i h(x)} g(x)\, dx\right),\ i = 1, \ldots, k-1, \tag{5}$$

for some $\beta_i \in \mathbb{R}$. Then, $H_0 : \beta_1 = \cdots = \beta_{k-1} = 0$ can also be used to test for the equality of any k pdfs $g(x)$, $g_i(x)$, $1 \le i \le k-1$ satisfying (4), which are not necessarily normal.

In classical ANOVA, the parameters of the normal distributions are estimated using the maximum likelihood method. To estimate the β_i parameters in (4), Consider the restricted class of distributions obtained by multiplicative exponential distortions with $g_k = g$ as a reference and, based on k independent samples, used the profile maximum likelihood method to estimate $g, g_i, \forall i$ in that class [3]. This paper, instead, considers the class $\mathbb{C}$ of all distributions (which are restricted by a given mean for $h(X)$) and estimates the β_i by minimizing the Kullback–Leibler divergence between $g(x)$ and the class $\mathbb{C}$.

Often the criterion of comparison between distributions may be clear from the context of the data, which helps to formulate the constraint $\mathbb{C}$. In the radar data, we are interested to know if the mean rain rate is equal in two different radars. In the credit limit data, we are interested to know if the mean credit limit is the same in three education groups. The constraints can be multiple criteria as well. Once the constraints are fixed in $\mathbb{C}$, only those aspects are considered in the comparison between distributions.

This approach matches with the maximum entropy (ME) principle, which may be stated as follows: when selecting a model for a given situation, it is often appropriate to express the prior information in terms of the constraints. However, one must be careful so that no information other than these specified constraints are used in model selection. That is, other than the constraints that we have, the uncertainty associated with the probability distribution to be selected should be kept at its maximum [4]. In this paper, we extended the ME principle to general information projection using the Kullback–Leibler divergence.

We show in Section 2 that the solution (4) is optimum under specified constraints using h when g is known and, in this way, the proposed approach yields an optimality interpretation for the exponential tilt models. The proposed approach extends the comparisons of means in ANOVA to comparisons of means and variances for the normal and other known distributions using duality. In Section 3, we develop a semi-parametric approach when $g(x)$ is unknown and derive asymptotic test statistics for testing the equality of populations for the cases when sample sizes are equal or different.

In Section 4, we present simulation studies that evaluate the performance of the β_i-parameters with respect to the classical ANOVA methods. We also compare the test statistics developed in Section 4 with existing parametric and nonparametric procedures. Section 5 shows the details of the proposed methods for the applications with radar data and credit limit data sets. Section 6 contains discussions on the choice of the function(s) h for particular cases of reference distributions $g(x)$. The Appendix contains additional results and proofs.

2. Tilt Optimality Models

Kullback–Liebler (KL) discriminant information or divergence is a measure of 'distance' between two probability distributions. For pdf's $f(x), g(x)$, the KL discriminant

information for $f(x)$ against $g(x)$ is given by $\int_{-\infty}^{\infty} f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx$, which is always nonnegative, and $= 0$ if and only if $f \equiv g$.

Let $\mathbb{C}$ be a convex set of pdf's with $g \notin \mathbb{C}$. If $f^*(x)$ is the solution to

$$\inf_{f(x) \in \mathbb{C}} \quad \int_{-\infty}^{\infty} f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx \tag{6}$$

then $f^*(x)$ (*information projection*) is the *closest* to $g(x)$ among all pdf's in $\mathbb{C}$ with respect to the KL distance. Let $h_i(x)$ be arbitrary but known functions of x. Define $\mathbb{C}$ to be the class of all pdf's f where $E_f(h_i(X)) = 0, \forall i$—that is,

$$\mathbb{C} = \left\{ f(x) : \int_{-\infty}^{\infty} h_i(x) f(x) \, dx = 0, \forall i \right\}. \tag{7}$$

In order to solve (6), Fenchel's duality theorem can be applied as shown by [5]. The corresponding dual cone is $\mathbb{C}^* = \{ \sum_{i=1}^{k} \beta_i h_i(x) : \beta_i \in \mathbb{R} \}$.

The dual problem can be shown to be equivalent to

$$\inf_{\beta_i \in \mathbb{R}, \forall i} \quad \int_{-\infty}^{\infty} \exp\left(\sum_{i=1}^{k} \beta_i h_i(x)\right) g(x) \, dx. \tag{8}$$

As the dual problem (8) is a function of scalars only, it could be substantially easier to solve than the primal problem (6), depending on the form of $h_i(x)$. In particular, setting the derivative of the integral quantity in (8) with respect to β_i equal to zero, we obtain

$$\int_{-\infty}^{\infty} h_i(x) \exp\left(\sum_{i=1}^{k} \beta_i h_i(x)\right) g(x) \, dx = 0, 1 \le i \le k, \tag{9}$$

which can be solved easily for β_i's by the Newton–Raphson method. If the solution of (9) is $\hat{\beta}_i$, then, from [5], the solution of the primal problem (6), say $f^*(x)$, is of the form

$$f^*(x) = \frac{g(x) \exp(\sum_{i=1}^{k} \hat{\beta}_i h_i(x))}{\int_{-\infty}^{\infty} g(x) \exp(\sum_{i=1}^{k} \hat{\beta}_i h_i(x)) \, dx}. \tag{10}$$

When $k = 1$, setting $\hat{\beta} = \beta_i, f^*(x) = g_i(x)$, (10) can be simplified as (4). The above derivation explains the exponential structure of $g_i(x)$, and (9) verifies that the exponential tilt model $g_i(x)$ is in $\mathbb{C}$ (see (7)). These developments are summarized in the following theorem.

Theorem 1. *When $g(x)$ is known, the exponential tilt model (4) is the* optimum *model in the sense that it is the closest to $g(x)$ among all probability distributions in $\mathbb{C}$ from (7) in the KL distance.*

The model (10) will be referred to in the sequel as the *tilt optimality* (TO) model. Note that, when $\mathbb{C}$ in (7) specifies $h(x) = x - \mu_i$ and $g(x) = g_k(x) \sim N(\mu_k, \sigma^2)$, then the solution $g_i(x)$ is found to be $N(\mu_i, \sigma^2)$ as was seen in (1)–(2). When $g(x)$ in (6) is uniform (or Lebesgue measure), then the minimizing f^* is known as the *maximum entropy* model (distribution) in $\mathbb{C}$ [4].

Theorem A1 in the Appendix A shows that closed form solutions for the dual problem are obtained for the normal distributions with constraints on both the mean and variance. However, this may not always be the case. The final solution depends on the form of $g(x)$ and the restrictions in $\mathbb{C}$.

While (1) compares each $g_i(x)$ with $g_k(x)$ one-at-a-time, another approach would be to compare all k distributions simultaneously. For $x = (x_1, \ldots, x_k)$, let $X = (X_1, \ldots, X_k) \sim g(x) = N_k(\mu, \Sigma)$ with known $\mu = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\Sigma = (\sigma_{ij})$ with

$\sigma_{ii} = \sigma_i^2$ and $\sigma_{ij} = 0$, elsewhere. Considering equality of the k means (unspecified) with possibly unequal variances, we define

$$\mathbb{C} = \{f(x) : E(X_i) = E(X_{i+1}),\ 1 \le i \le k-1\}. \tag{11}$$

Following (10), the solution is given by

$$f^*(x) = \frac{g(x)\, e^{\sum_{i=1}^{k} \hat{\beta}_i (x_i - x_{i+1})}}{\int_{-\infty}^{\infty} g(x)\, e^{\sum_{i=1}^{k} \hat{\beta}_i (x_i - x_{i+1})}\, dx}. \tag{12}$$

However, a closed form expression is only available when $k = 3, 4$.

Theorem 2 (proof in Appendix A) considers equality of means as in classical ANOVA (deals with unknown means and unknown but equal variances) but with unrestricted variances.

Theorem 2. *For $k = 3$ with known reference pdf $g(x) = N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the above $\boldsymbol{\Sigma}$, we consider the minimization problem*

$$\inf_{f(x) \in \mathbb{C}} \int_x f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx \tag{13}$$

with $\mathbb{C}$ in (11). The solution to (13) is given by

$$f^*(x) = g(x)\, e^{\hat{\alpha}_1 + \hat{\beta}_1 (x_1 - x_2) + \hat{\beta}_2 (x_2 - x_3)} = N_3(\mu^* \mathbf{1}, \boldsymbol{\Sigma}) \tag{14}$$

where $\hat{\alpha}_1 = -\ln\left[\int g(x)\, e^{\hat{\beta}_1 (x_1 - x_2) + \hat{\beta}_2 (x_2 - x_3)} dx\right]$,

$$\mu^* = \frac{\mu_1 \sigma_2^2 \sigma_3^2 + \mu_2 \sigma_1^2 \sigma_3^2 + \mu_3 \sigma_1^2 \sigma_2^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2}, \quad \mathbf{1} = (1,1,1)', \tag{15}$$

$$\hat{\beta}_1 = -\left(\frac{\mu_1 \sigma_2^2 + \mu_1 \sigma_3^2 - \mu_2 \sigma_3^2 - \mu_3 \sigma_2^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2}\right), \quad \hat{\beta}_2 = -\left(\frac{\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2 - \mu_3 \sigma_1^2 - \mu_3 \sigma_2^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2}\right). \tag{16}$$

Note that the solution $f^*(x)$ has the same covariance $\boldsymbol{\Sigma}$ as the reference $g(x)$; nonetheless, $\boldsymbol{\Sigma}$ influences the mean μ^* in solution (15) as a weighted average of its elements. When $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ as in one-way ANOVA, then $\mu^* = \frac{\mu_1 + \mu_2 + \mu_3}{3}$,

$$\hat{\beta}_1 = -\left(\frac{2\mu_1 - \mu_2 - \mu_3}{3\sigma^2}\right), \quad \hat{\beta}_2 = -\left(\frac{\mu_1 + \mu_2 - 2\mu_3}{3\sigma^2}\right). \tag{17}$$

Theorem 2 can be extended to $k = 4$ with closed form solutions for $\boldsymbol{\beta}, \mu^*$, which is tedious [6]. Unique solutions exist for higher k but obtaining their closed forms seems intractable. This is also the case of extension to a general $\boldsymbol{\Sigma}$ with $\sigma_{ij} \ne 0$, $i \ne j$.

Beyond the one-way ANOVA, the above approach allows us to *simultaneously* compare both the means and variances of k independent normal distributions (Theorem A2).

3. Semiparametric Approach

For an unknown data generating process, however, the true form of $g(x)$ in (13) (with $\mathbb{C}$ in (11)) may *not* be known. Then, $g_i(x)$ in the solution (4) becomes not well-defined (replacing x with x). Note that the model (10) now becomes a 'semiparametric tilt optimality restricted model' because, along with the parametric component $\boldsymbol{\beta}$, there is also the nonparametric component $g(x)$, about which no distributional assumption is made. Using the sample, we define a discrete version of $\mathbb{C}$ expressed as moment constraints. Assuming that these sample (moment) constraints represent the corresponding population (moment) constraints efficiently and consistently, the resulting model is expected to perform well.

The dual problem corresponding to (13) is

$$\inf_{\beta_i \in \mathbb{R}, \forall i} \int_x e^{\sum_{i=1}^{k-1} \beta_i(x_i - x_{i+1})} g(x) dx. \tag{18}$$

The relevant score equations are

$$\int_x (x_i - x_{i+1}) e^{\sum_{i=1}^{k-1} \beta_i(x_i - x_{i+1})} g(x) dx = 0, \ 1 \le i \le k-1. \tag{19}$$

To study the asymptotic properties of the model (similar in spirit to [7]), we consider the cases when the sample sizes are equal and when they are different.

3.1. Equal Sample Sizes

Suppose k independent random samples $\{x_{ij} \ 1 \le j \le n\}$, $1 \le i \le k$, each of size n, are available from independent populations with unknown means μ_i and unknown variances σ_i^2, $1 \le i \le k$, respectively. If we rearrange all the nk sample values as $(x_1, \ldots, x_n)$, where $x_j = (x_{1j}, \ldots, x_{kj})'$, $1 \le j \le n$, then x_j's form a random sample from a multivariate distribution (say, pdf $g(x)$), with mean $\mu = (\mu_1, \ldots, \mu_k)'$ and covariance $\text{diag}(\sigma_1^2, \ldots, \sigma_k^2)$.

Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)'$ be the empirical distribution that has mass $\hat{p}_j = \frac{1}{n}$ at each x_j, $1 \le j \le n$. The constraint of equality of k means, $\mathbb{C}$ in (11), is discretized below as $\mathbb{K}$, appropriately, using the probability mass function (pmf) $q = (q_1, \ldots, q_n)$ and sample values x_{ij}'s as

$$\mathbb{K} = \left\{ q : \sum_{j=1}^n (x_{ij} - x_{i+1,j}) q_j = 0, 1 \le i \le k-1 \right\}.$$

Here, the primal problem (6) becomes (replacing pdfs with pmfs) $\inf_{q \in \mathbb{K}} \sum_{j=1}^n q_j \ln \left(\frac{q_j}{\hat{p}_j} \right)$, and the dual problem is $\inf_{\beta_i \in \mathbb{R}, \forall i} \sum_{j=1}^n \hat{p}_j \exp \left[\sum_{i=1}^{k-1} \beta_i(x_{ij} - x_{i+1,j}) \right]$. The score equations for the dual problem are

$$\sum_{j=1}^n \hat{p}_j (x_{ij} - x_{i+1,j}) \exp \left[\sum_{i=1}^{k-1} \beta_i(x_{ij} - x_{i+1,j}) \right] = 0, \ 1 \le i \le k-1. \tag{20}$$

Suppose $\beta_n = (\beta_{ni}, 1 \le i \le k-1)$ solve the score Equation (20). Then, the primal solution $\hat{q}_j^*$ is given by

$$\hat{q}_j^* = \frac{\hat{p}_j \exp \left[\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j}) \right]}{\sum_{j=1}^n \hat{p}_j \exp \left[\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j}) \right]}, \ j = 1, \ldots, n.$$

For arbitrary $\delta = (\delta_1, \ldots, \delta_{k-1})'$, define

$$h_j(\delta) = \sum_{i=1}^{k-1} \delta_i(x_{ij} - x_{i+1,j}), \ 1 \le j \le n,$$

$$m_n(\delta) = (m_{nr}(\delta) : 1 \le r \le k-1), \Sigma_n(\delta) = (\sigma_{nrs}(\delta); 1 \le r, s \le k-1),$$

$$m_{nr}(\delta) = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{rj} - x_{r+1,j})e^{h_j(\delta)}}{\frac{1}{n}\sum_{j=1}^{n}e^{h_j(\delta)}}, \text{ and}$$

(21)

$$\sigma_{nrs}(\delta) = \frac{\frac{1}{n}\sum_{j=1}^{n}\left[(x_{rj} - x_{r+1,j}) - m_{nr}(\delta)\right]\left[(x_{sj} - x_{s+1,j}) - m_{ns}(\delta)\right]e^{h_j(\delta)}}{\frac{1}{n}\sum_{j=1}^{n}e^{h_j(\delta)}}.$$

By (20) and (21), $m_n(\boldsymbol{\beta}_n) = \mathbf{0}$, a vector of zeroes of length $k-1$. Furthermore, by Taylor's expansion of $m_n(\boldsymbol{\beta}_n)$ around $\boldsymbol{\beta}$ ($\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k-1})$),

$$\mathbf{0} = m_n(\boldsymbol{\beta_n}) = m_n(\boldsymbol{\beta}) + \Sigma_n(\boldsymbol{w_n})(\boldsymbol{\beta_n} - \boldsymbol{\beta}),$$

(22)

where $\boldsymbol{w}_n$ satisfies $\max\{|\boldsymbol{w}_n - \boldsymbol{\beta}_n|, |\boldsymbol{w}_n - \boldsymbol{\beta}|\} \leq |\boldsymbol{\beta}_n - \boldsymbol{\beta}|$. As $n \to \infty$, by the strong law of large numbers, $\boldsymbol{\beta}_n \to \boldsymbol{\beta}, \boldsymbol{w}_n \to \boldsymbol{\beta}$, and

$$\frac{1}{n}\sum_{j=1}^{n}(x_{rj} - x_{r+1,j})^a (x_{sj} - x_{s+1,j})^b e^{h_j(\boldsymbol{w}_n)} \longrightarrow \int (x_r - x_{r+1})^a (x_{sj} - x_{s+1,j})^b e^{h(\boldsymbol{\beta})} g(x)dx,$$

(23)

$a, b = 0, 1$, $1 \leq r, s \leq k-1$, with probability 1 where $h(\boldsymbol{\beta}) = \sum_{i=1}^{k-1}\beta_i(x_i - x_{i+1})$.

By (21) and (23), it follows that, when $\mu_1 = \ldots = \mu_k$,

$$m_{nr}(\boldsymbol{w_n}) \longrightarrow \frac{\int (x_r - x_{r+1})e^{h(\boldsymbol{\beta})}g(x)dx}{\int e^{h(\boldsymbol{\beta})}g(x)dx} = 0,$$

(24)

$$\sigma_{nrs}(\boldsymbol{w_n}) \longrightarrow \sigma_{rs} = \frac{\int (x_r - x_{r+1})(x_s - x_{s+1})e^{h(\boldsymbol{\beta})}g(x)dx}{\int e^{h(\boldsymbol{\beta})}g(x)dx},$$

$1 \leq r, s \leq k-1$, with probability 1. Let $\Sigma = (\sigma_{rs})$. As $n \to \infty$, by the central limit theorem

$$\sqrt{n}(m_n(\boldsymbol{\beta}) - \mathbf{0}) \longrightarrow N(\mathbf{0}, \Sigma^*),$$

(25)

where $\Sigma^* = (\sigma_{rs}^*)$,

$$\sigma_{rs}^* = \frac{\int (x_r - x_{r+1})(x_s - x_{s+1})e^{2h(\boldsymbol{\beta})}g(x)dx}{\left(\int e^{h(\boldsymbol{\beta})}g(x)dx\right)^2}.$$

(26)

By (22)–(26),

$$\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}) = \left[\Sigma_n(\boldsymbol{w_n})\right]^{-1}\sqrt{n}(m_n(\boldsymbol{\beta}_n) - m_n(\boldsymbol{\beta})) \xrightarrow{D} N(\mathbf{0}, \Sigma^{-1}\Sigma^*\Sigma^{-1}),$$

as $n \to \infty$. The above developments are summarized in the following theorem. This theorem establishes the asymptotic normality of the parameters of the proposed model when sample sizes are equal.

Theorem 3. *For a general reference pdf g, assume that the solution of (13), $f^*(x)$, exists and $\int e^{h(\boldsymbol{\beta})}f^*(x)dx < \infty$ where $h(\boldsymbol{\beta}) = \sum_{i=1}^{k-1}\beta_i(x_i - x_{i+1})$ for $\boldsymbol{\beta}$ in an open neighborhood of $\mathbf{0}$. When $\int (x_i - x_{i+1})(x_j - x_{j+1})e^{2h(\boldsymbol{\beta})}g(x)dx < \infty, \forall i, j$, and all sample sizes are equal to n,*

$$\sqrt{n}(\boldsymbol{\beta_n} - \boldsymbol{\beta}) \xrightarrow{D} N(\mathbf{0}, \Sigma^{-1}\Sigma^*\Sigma^{-1})$$

as $n \to \infty$ where Σ, Σ^* are defined in (24) and (26), respectively.

When $\mu_1 = \ldots = \mu_k$, or equivalently, $\boldsymbol{\beta} = \mathbf{0}$, the quantity $n\boldsymbol{\beta}_n'(\Sigma^{-1}\Sigma^*\Sigma^{-1})^{-1}\boldsymbol{\beta}_n$ has an asymptotic chi-square distribution with $k-1$ degrees of freedom as $n \to \infty$.

Thus, the test statistic

$$\chi_1 = n\boldsymbol{\beta}_n' (\hat{\Sigma}^{-1}\hat{\Sigma}^*\hat{\Sigma}^{-1})^{-1}\boldsymbol{\beta}_n \tag{27}$$

can be used for testing the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ where $\sigma_{rs}, \sigma_{rs}^*$ are estimated by $\sigma_{nrs}(\boldsymbol{\beta}_n)$, $\sigma_{nrs}^*(\boldsymbol{\beta}_n)$, respectively, (since $\Sigma = \Sigma^*$, we replaced $\sigma_{nrs}(\boldsymbol{\beta}_n), \sigma_{nrs}^*(\boldsymbol{\beta}_n)$ by $(\sigma_{nrs}(\boldsymbol{\beta}_n) + \sigma_{nrs}^*(\boldsymbol{\beta}_n))/2)$.

$$
\begin{aligned}
\sigma_{nrs}(\boldsymbol{\beta}_n) &= \frac{\frac{1}{n}\sum_{j=1}^n \left[(x_{rj} - x_{r+1,j}) \right] \left[(x_{sj} - x_{s+1,j}) \right] e^{\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j})}}{\frac{1}{n}\sum_{j=1}^n e^{\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j})}}, \\
\sigma_{nrs}^*(\boldsymbol{\beta}_n) &= \frac{\frac{1}{n}\sum_{j=1}^n \left[(x_{rj} - x_{r+1,j}) \right] \left[(x_{sj} - x_{s+1,j}) \right] e^{2\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j})}}{\left(\frac{1}{n}\sum_{j=1}^n e^{\sum_{i=1}^{k-1} \beta_{ni}(x_{ij} - x_{i+1,j})} \right)^2}.
\end{aligned}
\tag{28}
$$

Clearly, the above developments can be extended for simultaneous mean and variance comparisons for k populations by modifying the $\mathbb{C}$ in (11).

3.2. Different Sample Sizes

The simultaneous approach developed above for equal sample sizes does not allow the sample sizes to be different. To that end, we consider $k - 1$ independent one-at-a-time population optimization problems by adopting the development in Section 2 reversing the roles of $g(x) = g_k(x)$ and $g_i(x)$ in (4), along with setting $h(x) = x$, and, finally, we combine the $k - 1$ results. Both procedures work for equal sample sizes.

For $1 \le i \le k - 1$, consider the ith problem as finding the pdf $f(x)$ in

$$\mathbb{C} = \{ f(x) : E[X] = \mu_k \}, \tag{29}$$

which is the closest to $g_i(x)$, assuming that μ_k is *known*. Following similar steps as in Section 2, the pdf in (29), which is the closest to $g_i(x)$, is given by

$$f_i^*(x) = g_i(x) \exp(\gamma_i + \eta_i x), \quad \gamma_i = -\ln\left(\int_{-\infty}^{\infty} e^{\eta_i x} g_i(x)\, dx \right), \tag{30}$$

$i = 1, \ldots, k - 1$, where η_i solves

$$\int_{-\infty}^{\infty} (x - \mu_k) \exp(\eta_i x) g_i(x)\, dx = 0. \tag{31}$$

If $\eta_i = 0$, then $f_i^*(x) = g_i(x), \forall x, \forall i$.

To develop the corresponding $k - 1$ sample optimization problems, suppose independent random samples $\{x_{ij}\ 1 \le j \le n_i\}$, $1 \le i \le k$ are available from k populations with the means μ_i, $1 \le i \le k$, respectively. Although we assume that μ_k is known, in reality, it maybe unknown. Thus, we suggest choosing the kth sample as the one that is the largest in size. Let $\bar{x}_k = \frac{\sum_{j=1}^{n_k} x_{ij}}{n_k}$ be the mean of the kth sample. Then, take $\mu_k = \bar{x}_k$ (see Section 6).

For the ith $(1 \le i \le k - 1)$ sample optimization problem, let $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{i,n_i})$ be an empirical distribution that has mass $\hat{p}_{ij} = \frac{1}{n_i}$ at each x_{ij}, $1 \le j \le n_i$. Let the ith sample version of $\mathbb{C}$ in (29), say $\mathbb{K}_i$, containing $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{i,n_i})$, be defined as

$$\mathbb{K}_i = \left\{ \boldsymbol{q}_i : \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_k) q_{ij} = 0 \right\}, \quad 1 \le i \le k - 1.$$

The ith $(1 \le i \le k - 1)$ sample version of (6) and its dual problem becomes

$$\inf_{\boldsymbol{q}_i \in \mathbb{K}_i} \sum_{j=1}^{n_i} q_{ij} \ln\left(\frac{q_{ij}}{\hat{p}_{ij}} \right), \quad \text{and,} \quad \inf_{\eta_i \in \mathbb{R}} \sum_{j=1}^{n_i} \hat{p}_{ij} \exp\left[\eta_i (x_{ij} - \bar{x}_k) \right],$$

respectively. The ith score equation is

$$\sum_{j=1}^{n_i} \hat{p}_{ij}(x_{ij} - \bar{x}_k) \exp\left[\eta_i(x_{ij} - \bar{x}_k)\right] = 0. \tag{32}$$

Suppose η_{n_i} solves (32). Then, the primal solution $\hat{q}_{ij}^*$ is given by

$$\hat{q}_{ij}^* = \frac{\hat{p}_{ij} \exp\left[\eta_{n_i}(x_{ij} - \bar{x}_k)\right]}{\sum_{j=1}^{n} \hat{p}_{ij} \exp\left[\eta_{n_i}(x_{ij} - \bar{x}_k)\right]}, \quad j = 1, \ldots, n_i.$$

For arbitrary $\psi_i \in \mathbb{R}$, $1 \le i \le k - 1$, define

$$h_{ij}(\psi_i) = \psi_i(x_{ij} - \bar{x}_k), \qquad m_{n_i}(\psi_i) = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_k)e^{h_{ij}(\psi_i)}}{\frac{1}{n_i}\sum_{j=1}^{n_i}e^{h_{ij}(\psi_i)}}, \text{ and}$$

$$\sigma_{n_i}^2(\psi_i) = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left[(x_{ij} - \bar{x}_k) - m_{n_i}(\psi_i)\right]^2 e^{h_{ij}(\psi_i)}}{\frac{1}{n_i}\sum_{j=1}^{n_i}e^{h_{ij}(\psi_i)}}. \tag{33}$$

By (32), $m_{n_i}(\eta_{n_i}) = 0, \forall i$. Using Taylor's expansion of $m_{n_i}(\eta_{n_i})$ around η_i,

$$0 = m_{n_i}(\eta_{n_i}) = m_{n_i}(\eta_i) + \sigma_{n_i}^2(w_{n_i})(\eta_{n_i} - \eta_i), \tag{34}$$

where w_{n_i} is such that $\max\{|w_{n_i} - \eta_{n_i}|, |w_{n_i} - \eta_i|\} \le |\eta_{n_i} - \eta_i|$. With $h_i = \eta_i(x - \bar{x}_k)$, by the strong law of large numbers, as $n_i \to \infty$, $\eta_{n_i} \to \eta_i$, $w_{n_i} \to \eta_i$ $\forall i$ and

$$\frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_k)^a e^{h_{ij}(w_{n_i})} \longrightarrow \int(x - \bar{x}_k)^a e^{h_i}g_i(x)dx, \tag{35}$$

$a = 0, 1, 2$, $1 \le i \le k - 1$, with probability 1.

From (31), $\int(x - \bar{x}_k)e^{h_i}g_i(x)dx = 0$. When $\mu_i = \mu_k$, by (33) and (35), it follows that, as $n_i \to \infty$,

$$m_{n_i}(w_{n_i}) \longrightarrow \frac{\int(x - \bar{x}_k)e^{h_i}g_i(x)dx}{\int e^{h_i}g_i(x)dx} = 0, \quad \sigma_{n_i}^2(w_{n_i}) \longrightarrow \sigma_i^2 = \frac{\int(x - \bar{x}_k)^2 e^{h_i}g_i(x)dx}{\int e^{h_i}g_i(x)dx}, \tag{36}$$

$1 \le i \le k - 1$, with probability 1. As $n_i \to \infty$, by the central limit theorem

$$\sqrt{n_i}(m_{n_i}(\eta_i) - 0) \longrightarrow N(0, \sigma_i^{2*}), \tag{37}$$

where

$$\sigma_i^{*2} = \frac{\int(x - \bar{x}_k)^2 e^{2h_i}g_i(x)dx}{\left[\int e^{h_i}g_i(x)dx\right]^2}. \tag{38}$$

By (34), (35) and (37),

$$\sqrt{n_i}(\eta_{n_i} - \eta_i) = \frac{\sqrt{n}(m_{n_i}(\eta_{n_i}) - m_{n_i}(\eta_i))}{\sigma_{n_i}^2(w_{n_i})} \xrightarrow{D} N\left(0, \frac{\sigma_i^{*2}}{\sigma_i^4}\right), \tag{39}$$

as $n_i \to \infty$, $1 \le i \le k - 1$.

Next, we combine the results from the $k - 1$ sample-optimization problems.

Theorem 4. *For a general reference pdf g, assume that the solution of (13) subject to (29), $f_i^*(x)$, exists and $\int e^{\eta_i x}f_i^*(x)dx < \infty$ for some η_i in an open neighborhood of 0, $1 \le i \le k - 1$. Assume*

$\int x^2 e^{\eta_i x} f_i^*(x) dx < \infty, \forall i$. Since all k samples are independent, $\frac{\sqrt{n_i}(\eta_{n_i} - \eta_i)}{\sigma_i^* / \sigma_i^2} \xrightarrow{D} N(0, 1)$, as $n_i \to \infty$.

Then, under $H_0 : \eta_i = 0$, $1 \le i \le k-1$, $\sum_{i=1}^{k-1} \frac{n_i \eta_{n_i}^2}{\sigma_i^{*2} / \sigma_i^4}$ has an asymptotic chi-square distribution with $k-1$ degrees of freedom as $n_i \to \infty, \forall i$. Thus, the test statistic

$$\sum_{i=1}^{k-1} \frac{n_i \eta_{n_i}^2}{\sigma_{n_i}^{*2}(\eta_{n_i}) / \sigma_{n_i}^4(\eta_{n_i})} \tag{40}$$

can be used for testing the hypothesis $H_0 : \eta_i = 0$, $1 \le i \le k-1$, where we replaced $\sigma_i^{*2}, \sigma_i^2$ by $\sigma_{n_i}^{*2}(\eta_{n_i}), \sigma_{n_i}^2(\eta_{n_i})$, respectively, with

$$\sigma_{n_i}^{*2}(\eta_{n_i}) = \frac{\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_k)^2 e^{2\eta_{n_i}(x_{ij} - \bar{x}_k)}}{\left[\frac{1}{n_i} \sum_{j=1}^{n_i} e^{\eta_{n_i}(x_{ij} - \bar{x}_k)} \right]^2}, \quad \sigma_{n_i}^2(\eta_{n_i}) = \frac{\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_k)^2 e^{\eta_{n_i}(x_{ij} - \bar{x}_k)}}{\frac{1}{n_i} \sum_{j=1}^{n_i} e^{\eta_{n_i}(x_{ij} - \bar{x}_k)}}. \tag{41}$$

Since $\sigma_i^{*2} = \sigma_i^2$ under H_0, we replace each of $\sigma_{n_i}^{*2}(\eta_{n_i}), \sigma_{n_i}^2(\eta_{n_i})$ by a common estimate $\sigma_{n_i}^{**2}(\eta_{n_i}) = (\sigma_{n_i}^{*2}(\eta_{n_i}) + \sigma_{n_i}^2(\eta_{n_i}))/2$, and then calculate a pooled estimate of variance $\hat{\sigma}^2 = (\sum_{i=1}^{k-1} n_i \sigma_{n_i}^{**2}(\eta_{n_i}) / \sum_{i=1}^{k-1} n_i$ over $k-1$ populations. With this substitution, the test statistic from (40) is simplified as

$$\chi_2 = \hat{\sigma}^2 \sum_{i=1}^{k-1} n_i \eta_{n_i}^2. \tag{42}$$

The above developments can be extended for simultaneous mean and variance comparisons for k populations by modifying the $\mathbb{C}$ in (29).

For one h:

$$\frac{1}{n} \sum_{j=1}^{n} e^{\beta_n h(x_j)} - \frac{1}{2n} \frac{\sum_{j=1}^{n} h^2(x_j) e^{2\beta_n h(x_j)}}{\sum_{j=1}^{n} h^2(x_j) e^{\beta_n h(x_j)}}$$

For more than one h, let $h = (h_1, \ldots, h_t)'$ represent t constraints.
(i) For equal sample sizes, recall $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{kj})'$, $1 \le j \le n$.
Recall Σ, Σ^* given as $\hat{\Sigma} = (\sigma_{nrs}(\boldsymbol{\beta}_n); 1 \le r, s \le t), \hat{\Sigma}^* = (\sigma_{nrs}^*(\boldsymbol{\beta}_n); 1 \le r, s \le t)$.

$$\sigma_{nrs}(\boldsymbol{\beta}_n) = \frac{\frac{1}{n} \sum_{j=1}^{n} h_r(\boldsymbol{x}_j) h_s(\boldsymbol{x}_j) e^{\sum_{l=1}^{t} \beta_{ni} h_i(\boldsymbol{x}_j)}}{\frac{1}{n} \sum_{j=1}^{n} e^{\sum_{l=1}^{t} \beta_{ni} h_i(\boldsymbol{x}_j)}},$$

$$\sigma_{nrs}^*(\boldsymbol{\beta}_n) = \frac{\frac{1}{n} \sum_{j=1}^{n} h_r(\boldsymbol{x}_j) h_s(\boldsymbol{x}_j) e^{2\sum_{l=1}^{t} \beta_{ni} h_i(\boldsymbol{x}_j)}}{\left(\frac{1}{n} \sum_{j=1}^{n} e^{\sum_{l=1}^{t} \beta_{ni} h_i(\boldsymbol{x}_j)} \right)^2}.$$

Then, the IC is

$$\text{IC} = \frac{1}{n} \sum_{j=1}^{n} e^{\sum_{l=1}^{t} \beta_{ni} h_i(\boldsymbol{x}_j)} - \frac{1}{2n} \sum_{l=1}^{t} \hat{\lambda}_i$$

where $\hat{\lambda}_i$ is the ith eigen value of $\hat{\Sigma}^{-\frac{1}{2}} \hat{\Sigma}^* \hat{\Sigma}^{-\frac{1}{2}}$. Find $\hat{\Sigma}^{-\frac{1}{2}}$ by Cholesky decomposition.
(ii) For different sample sizes, define Σ_i, Σ_i^* given as $\hat{\Sigma}_i = (\sigma_{n_i rs}(\boldsymbol{\eta}_{n_i}); 1 \le r, s \le t)$, $\hat{\Sigma}_i^* = (\sigma_{n_i rs}^*(\boldsymbol{\eta}_{n_i}); 1 \le r, s \le t)$.

$$\sigma_{n_i rs}(\boldsymbol{\eta}_{n_i}) = \frac{\frac{1}{n_i} \sum_{j=1}^{n_i} h_r(x_{ij}) h_s(x_{ij}) e^{\sum_{l=1}^{t} \eta_{nl} h_l(x_{ij})}}{\frac{1}{n_i} \sum_{j=1}^{n_i} e^{\sum_{l=1}^{t} \eta_{nl} h_l(x_{ij})}},$$

$$\sigma^*_{n_i rs}(\boldsymbol{\eta}_{n_i}) = \frac{\frac{1}{n_i} \sum_{j=1}^{n_i} h_r(x_{ij}) h_s(x_{ij}) e^{2\sum_{l=1}^{t} \eta_{nl} h_l(x_{ij})}}{\left(\frac{1}{n_i} \sum_{j=1}^{n_i} e^{\sum_{l=1}^{t} \eta_{nl} h_l(x_{ij})}\right)^2}.$$

Then, the IC is

$$\mathrm{IC} = \frac{1}{k-1} \sum_{i=1}^{k-1} \mathrm{IC}_i$$

$$\mathrm{IC}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} e^{\sum_{l=1}^{t} \eta_{nl} h_l(x_{ij})} - \frac{1}{2n_i} \sum_{l=1}^{t} \hat{\lambda}_{il}$$

where $\hat{\lambda}_{il}$ is the lth eigen value of $\hat{\Sigma}_{il}^{-\frac{1}{2}} \hat{\Sigma}_{il}^* \hat{\Sigma}_{il}^{-\frac{1}{2}}$. Find $\hat{\Sigma}_{il}^{-\frac{1}{2}}$ by Cholesky decomposition.

4. Simulation Studies

In this section, we study the effect of violations of the assumptions of the one-way ANOVA on the performances of the beta parameters using Mean Square Errors. We also compare powers of the proposed methods with existing parametric and nonparametric methods. We used the R program for all simulation studies. The proposed methods are applicable to any type of data including under-dispersed or over-dispersed or data containing outliers as we do not assume any distribution $g(x)$ in our analysis.

4.1. Mean Square Errors of $\boldsymbol{\beta}$

Although the one-way ANOVA does not have a β-parameter as is, (14) with $\hat{\beta}_i$ replaced by β_i may be considered as the one-way ANOVA model, which shows the β-parameters when $k = 3$. To implement the assumption of equal variance, we used $\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{3}$ as the common variance in ANOVA. The true $\beta_1 = \hat{\beta}_1, \beta_2 = \hat{\beta}_2$ are in (17). In the simulation, the true β_1, β_2 are not changed when the population distribution is any different from normal to maintain the normality assumption of the one-way ANOVA.

There are four beta parameters $(\beta_1, \ldots, \beta_4)$ under the simultaneous constraints of equality of means and equality of variances. For one-way ANOVA, we set $\beta_3 = 0$ and $\beta_4 = 0$ as there are no variance constraints. In the TO model (10), however, the true $\boldsymbol{\beta}$ changes depending on the population and constraints as shown in Section 3 for selected distributions as solutions of the corresponding score equations.

For the simulation in Table 1, we considered independent normal, uniform and beta populations and formed $g(x)$ by multiplying the pdf's of independent components as listed, which were chosen with varying means and/or variances. Various sample sizes of 50, 200 and 500 with 1000 runs for each case were considered of which only selected results are reported in Table 1 (see more details [6]). We found that the total mean square error of our method decreased for larger sample sizes.

As a consequence of the way the ANOVA model is defined in (14) using the $\boldsymbol{\beta}$, both ANOVA and TO solutions (or methods) become identical when populations are identical or when the means are the same, and thus they produce the same values for true β's, solutions and MSEs. Both methods have a relatively higher TMSE (the total of all MSEs) for beta reference distributions compared to the normal reference distributions. For the cases considered, the TO method has a smaller MSE compared with the one-way ANOVA method and more so when the mean and variance constraints are simultaneously considered.

Table 1. Mean square errors of the β parameters.

| Components of $g(x_1, x_2, x_3)$ | $E(X_1) = E(X_2) = E(X_3)$ | | | | | | | | | |
| | | | | Parameters | | MSE | | TMSE | |
	Mean	Var	Sample Sizes	TO	ANOVA	TO	ANOVA	TO	ANOVA
$X_1 \sim$ N(0,1)	0	1	200	β_1 : 0.3094	0.3030	0.0044	0.0041	0.0109	0.0125
$X_2 \sim$ N(0.2,1.1)	0.2	1.1	200	β_2 : 0.4088	0.4242	0.0065	0.0084		
$X_3 \sim$ N(0.8,1.2)	0.8	1.2	200						
$X_1 \sim$ Beta(1,2)	0.33	0.06	200	β_1 : 1.3433	1.4000	0.0726	0.0787	0.1442	0.1516
$X_2 \sim$ Beta(1.33,1.99)	0.4	0.06	200	β_2 : 1.5871	1.6000	0.0716	0.0729		
$X_3 \sim$ Beta(1.75,1.75)	0.5	0.06	200						
$X_1 \sim$ Beta(1,2)	0.33	0.06	200	β_1 : 1.2424	1.2245	0.0642	0.0639	0.1205	0.1245
$X_2 \sim$ Beta(1.08,1.62)	0.4	0.065	200	β_2 : 1.3543	1.3994	0.0562	0.0607		
$X_3 \sim$ Beta(1.29,1.29)	0.5	0.07	200						
	$E(X_1) = E(X_2) = E(X_3), V(X_1) = V(X_2) = V(X_3)$								
$X_1 \sim$ N(0,1)	0	1	200	β_1 : 0.2828	0.3030	0.0059	0.0081	0.0217	0.0408
$X_2 \sim$ N(0.2,1.1)	0.2	1.1	200	β_2 : 0.3838	0.4242	0.0094	0.0166		
$X_3 \sim$ N(0.8,1.2)	0.8	1.2	200	β_3 : 0.0429	0	0.0038	0.0091		
				β_4 : 0.0404	0	0.0026	0.0069		
$X_1 \sim$ Beta(1,2)	0.33	0.067	200	β_1 : 2.8090	1.4000	0.9763	2.1442	4.5389	9.7559
$X_2 \sim$ Beta(1.33,1.99)	0.4	0.067	200	β_2 : 2.9256	1.6000	1.1376	2.1325		
$X_3 \sim$ Beta(1.75,1.75)	.5	0.067	200	β_3 : -1.6512	0	1.2599	2.9440		
				β_4 : -1.4902	0	1.1651	2.5353		
$X_1 \sim$ Beta(1.33,1.99)	0.4	0.067	500	β_1 : -1.4640	0	0.6133	2.3176	3.9619	12.9005
$X_2 \sim$ Beta(1.2,1.8)	0.4	0.06	200	β_2 : -1.7934	0	1.1961	3.3782		
$X_3 \sim$ Beta(.97,1.46)	0.4	0.07	100	β_3 : 1.6611	0	0.7408	2.9568		
				β_4 : 2.0318	0	1.4117	4.2478		

Here **TO** = Proposed optimality model, **ANOVA** = Analysis of variance, **MSE** = Mean square errors, **TMSE** = Total mean square errors.

4.2. Power Comparison: Equal Sample Sizes

When testing the equality of means with equal sample sizes, we conducted simulation studies to compare our proposed test statistic χ_1's (in (27)) performance with χ_2 in (42) developed for different sample sizes, (χ) [3], Hotelling's tests (T_1^2 or T_2^2 depending on known (chi-square) or unknown (*F*-distribution)) variances, classical ANOVA and the nonparametric Kruskal–Wallis (KW) test statistics.

For $k = 3$, with $\mu = (0, \mu_2, \mu_3)$, we considered either $g(x) \sim N_3(\mu, \Sigma = I_3)$, or $g(x) \sim gamma(3, 1) \times gamma(3 + \mu_2, 1) \times gamma(3 + \mu_3, 1)$. The hypothesis of equality of means becomes equivalent to testing $H_0 : \mu_2 = \mu_3 = 0$. Equal sample sizes are taken as $n = 30$ or $n = 50$. For all simulations, 1000 runs were used with a nominal level of $\alpha = 0.05$.

The power results as a function of μ_2, μ_3 are shown in Table 2. Both χ_1 and χ_2 tests have higher powers than other tests under the gamma distribution, including the nonparametric test with the χ_2 test performing slightly better than the χ_1 test. However, none of the χ, χ_1, χ_2 tests was dominated by the ANOVA test under the normal case as expected (also observed by [3] for the χ test).

4.3. Power Comparison: Different Sample Sizes

We compared the performance of the proposed semi-parametric test statistic χ_2 in (42) with the test statistic χ [3], classical ANOVA and the Kruskal–Wallis test (KW).

We used the same $g(x)$, H_0, number of runs and nominal level as in the equal sample size case. The power results as a function of μ_2, μ_3 are shown in Table 3 with different sample sizes n_1, n_2, n_3. For the calculation of χ_2, the value of the sample mean with the largest sample size is taken as fixed in all cases.

Table 2. Power comparison with equal sample sizes.

			Normal							Gamma						
n	μ_2	μ_3	χ	χ_1	χ_2	T_1^2	T_2^2	**ANOVA**	**KW**	χ	χ_1	χ_2	T_1^2	T_2^2	**ANOVA**	**KW**
30	0	0	0.054	**0.080**	**0.096**	0.043	0.058	0.040	0.042	0.069	**0.087**	**0.126**	0.045	0.068	0.041	0.038
	0.2	0.2	0.126	**0.184**	**0.368**	0.101	0.145	0.109	0.099	0.057	**0.140**	**0.104**	0.074	0.101	0.060	0.059
	0.1	0.4	0.309	**0.381**	**0.587**	0.267	0.330	0.284	0.278	0.083	**0.182**	**0.175**	0.107	0.145	0.120	0.110
	0.2	0.5	0.428	**0.495**	**0.805**	0.379	0.440	0.398	0.377	0.103	**0.239**	**0.250**	0.145	0.188	0.143	0.138
	0.5	0.5	0.527	**0.598**	**0.963**	0.474	0.546	0.499	0.480	0.115	**0.267**	**0.418**	0.161	0.218	0.170	0.183
	0.7	0.5	0.730	**0.774**	**0.992**	0.664	0.730	0.703	0.663	0.158	**0.346**	**0.573**	0.229	0.281	0.241	0.259
50	0	0	0.056	**0.084**	**0.083**	0.050	0.060	0.046	0.047	0.063	**0.091**	**0.112**	0.051	0.070	0.054	0.051
	0.2	0.2	0.161	**0.198**	**0.483**	0.139	0.164	0.150	0.153	0.058	**0.113**	**0.116**	0.075	0.095	0.078	0.080
	0.1	0.4	0.457	**0.508**	**0.769**	0.430	0.463	0.438	0.397	0.117	**0.216**	**0.254**	0.151	0.179	0.156	0.173
	0.2	0.5	0.621	**0.648**	**0.938**	0.577	0.618	0.600	0.565	0.144	**0.268**	**0.398**	0.202	0.231	0.203	0.217
	0.5	0.5	0.739	**0.768**	**0.995**	0.717	0.741	0.730	0.689	0.183	**0.345**	**0.635**	0.246	0.278	0.254	0.261
	0.7	0.5	0.897	**0.917**	**1.000**	0.882	0.903	0.896	0.887	0.306	**0.482**	**0.831**	0.383	0.424	0.386	0.422

Here χ = Test statistic [3], χ_1 = Proposed test statistic (27) for equal sample sizes, χ_2 = Proposed test statistic (42) for different sample sizes, T_1^2 or T_2^2 = Hotelling's tests depending on known (chi-square) or unknown (*F*-distribution) variances, **ANOVA** = *F* test statistic of analysis of variance, **KW** = Kruskal-Wallis test statistic.

Table 3. Power comparison with different sample sizes.

			Normal				Gamma			
Sample Sizes	μ_2	μ_3	χ_2	χ	**ANOVA**	**KW**	χ_2	χ	**ANOVA**	**KW**
$n_1 = 200$	0.0	0.0	**0.068**	0.050	0.048	0.048	**0.085**	0.055	0.040	0.040
$n_2 = 100$	0.2	0.2	**0.594**	0.333	0.330	0.306	**0.182**	0.113	0.161	0.159
$n_3 = 40$	0.1	0.4	**0.719**	0.531	0.520	0.498	**0.244**	0.120	0.202	0.208
	0.2	0.5	**0.916**	0.785	0.772	0.755	**0.382**	0.222	0.337	0.327
	0.5	0.5	**1.000**	0.987	0.988	0.981	**0.807**	0.514	0.602	0.629
	0.7	0.5	**1.000**	1.000	1.000	1.000	**0.968**	0.766	0.839	0.854
$n_1 = 200$	0.0	0.0	**0.071**	0.055	0.054	0.055	**0.078**	0.053	0.046	0.043
$n_2 = 100$	0.2	0.2	**0.714**	0.370	0.363	0.349	**0.219**	0.105	0.141	0.153
$n_3 = 100$	0.1	0.4	**0.960**	0.835	0.828	0.813	**0.475**	0.269	0.346	0.347
	0.2	0.5	**0.999**	0.960	0.959	0.962	**0.735**	0.443	0.542	0.569
	0.5	0.5	**1.000**	0.998	0.998	0.997	**0.933**	0.606	0.673	0.703
	0.7	0.5	**1.000**	1.000	1.000	1.000	**0.990**	0.819	0.867	0.893

Here χ_2 = Proposed test statistic (42) for different sample sizes, χ = Test Statistic [3], **ANOVA** = *F* test statistic of analysis of variance, **KW** = Kruskal-Wallis test statistic.

It can be seen that the χ_2 test performed better than the other tests. However, it is not dominated by the ANOVA test as expected under the normal distribution.

5. Applications

We considered two applications. Equal sample sizes were used in the radar meteorology example, and the sample sizes were different in the credit limit example.

5.1. Radar Meteorology

We considered radar reflectivity data [1] from two independent radars deployed during NASA's Tropical Rainfall Measuring Mission Kwajalein Experiment. An S-band radar was located on Kwajalein Island at the southern end of the Kwajalein Atoll, and a C-band radar was on board the NOAA ship Ronald H. Brown. Different calibrations were applied to two radars, and their spherical data were recorded to the same 1 km cube Cartesian grid.

Histograms of the samples were identical if sample 1 and sample 2 both are taken from the C-band population (Figure 1). Histograms of sample 1 and 3 illustrate that the

shapes of the distributions are different for sample 1 taken from the C-band population and sample 2 taken from the S-band population (Figure 1).
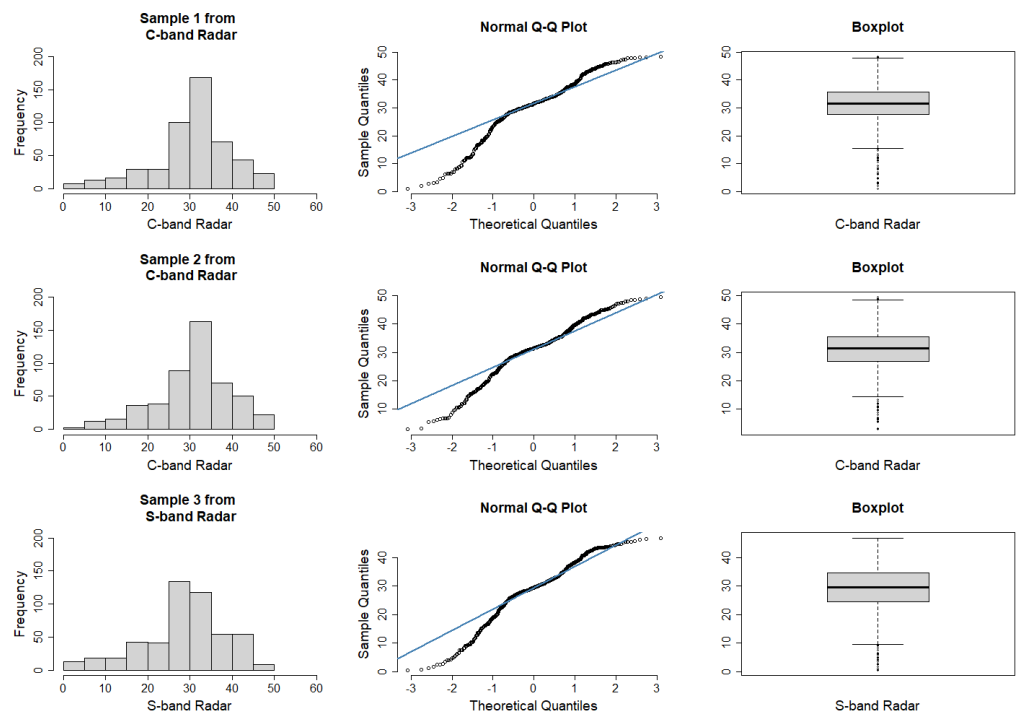

Figure 1. Radar reflectivity data: histograms, normal QQ-plots and boxplots. Sample 1 and sample 2 were taken from the C-band radar population with each sample size $n = 500$. Sample 3 was taken from the S-band Radar population with the same sample size $n = 500$.

First, two independent samples were taken from the C-band radar population to test whether they were from the same population. The population size was around 3987, and the sample sizes were 500 each. To test that the populations were equal, the method of Section 4 (in particular, the test statistic χ_1 in (27) was used with $k = 2$), and the results are given in Table 4. Using only equality of means constraints, the data fail to reject $H_0 : \beta = 0$ with the p-value = 0.71. Using simultaneous equality of means and equality of variance constraints, the data again fail to reject $H_0 : \beta_1 = 0, \beta_2 = 0$ with p-value = 0.18, and thus we concluded that both samples are from the same population (with respect to the means and variances).

Table 4. Hypothesis testing when samples are coming from the same or different populations.

Populations	H_0	Transformation	Dual Solutions	χ_1	*p*-Value
C-band, C-band	$H_0 : \beta = 0$	$h(\boldsymbol{x}) = x_1 - x_2$	$\hat{\beta} = -0.00086$	0.1382	0.71
	$H_0 : \beta_1 = 0,$	$h_1(\boldsymbol{x}) = x_1 - x_2,$	$\hat{\beta}_1 = -0.0175,$	3.3577	0.18
	$\beta_2 = 0$	$h_2(\boldsymbol{x}) = x_1^2 - x_2^2$	$\hat{\beta}_2 = 0.0003$		
	$H_0 : \beta = 0$	$h(\boldsymbol{x}) = \ln x_1 - \ln x_2$	$\hat{\beta} = -0.08541$	-0.74671	0.4552
C-band, S-band	$H_0 : \beta = 0$	$h(\boldsymbol{x}) = x_1 - x_2$	$\hat{\beta} = -0.0087$	13.24	0.0003
	$H_0 : \beta_1 = 0,$	$h_1(\boldsymbol{x}) = x_1 - x_2,$	$\hat{\beta}_1 = 0.0048,$	23.4	0.0
	$\beta_2 = 0$	$h_2(\boldsymbol{x}) = x_1^2 - x_2^2$	$\hat{\beta}_2 = -0.0002$		
	$H_0 : \beta = 0$	$h(\boldsymbol{x}) = \ln x_1 - \ln x_2$	$\hat{\beta} = 0.4414$	4.8307	0.0

Next, we considered two independent samples of size 500 each from C- and S-band populations to test if those were from the same population. With C-band radar data as the reference and when using only equality of means constraints, the data rejected $H_0 : \beta = 0$ with p-value =0.0003. Using equality of means and equality of variance

constraints simultaneously, the data again rejected $H_0 : \beta_1 = 0, \beta_2 = 0$ with p-value ≈ 0; therefore, we concluded that the samples are from different populations (with respect to the means and variances).

From the histograms, the data appear to be skewed to the left (as was noted by [3]). Thus, we considered the log transformation of the data and repeated the analysis for equality of means as reported in Table 4; however, our conclusions remain unchanged.

5.2. Credit Limit Data

Previous study considered credit limit information from three different education levels, namely, graduate school, university and high school [2]. We want to investigate whether the populations are identical using the constraints that the means of the credit limit between three different education levels are equal. The sample sizes from the three populations are 10,585, 14,030 and 4917, respectively.

The credit limit data of each groups follows positively skewed distributions (Figure 2).
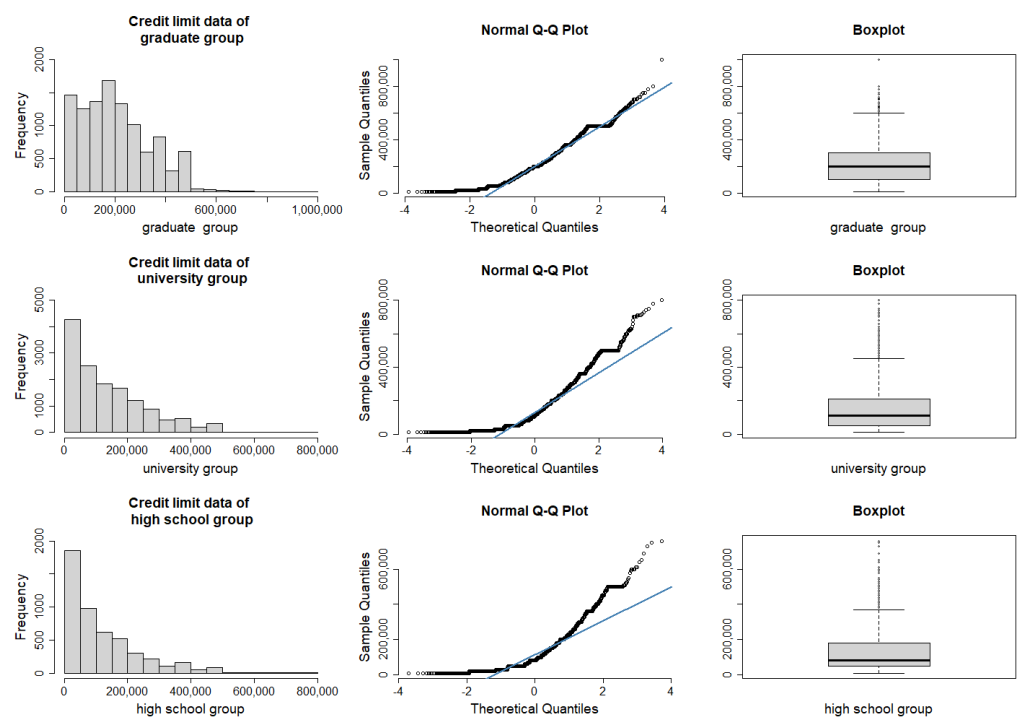
Figure 2. Credit limit data: histograms, normal QQ-plots and boxplots for the graduate ($n = 10,585$), university ($n = 14,030$) and high school ($n = 4917$) student groups.

To test that the populations were identical with respect to their means, the method of Section 4 for different sample sizes, in particular, the test statistic χ_2 in (42) was used with $k = 3$. As the 'university' group had the largest sample size, the sample mean from that group was taken as fixed to be used in (29). While comparing the other groups with the university group, we found $\hat{\eta}_1 = -0.4435 \times 10^{-5}$ and $\hat{\eta}_2 = 0.1396 \times 10^{-5}$. The test statistic value was $\chi_2 = 1272.92$ with p-value ≈ 0. Thus, we reject $H_0 : \eta_1 = \eta_2 = 0$, and conclude that the populations are different. As the data were skewed, a log transformation was, again, considered. Then, the dual problem solutions were $\hat{\eta}_1 = -0.563, \hat{\eta}_2 = 0.1988$, and the test statistic was $\chi_2 = 928.75$ (p-value ≈ 0) rejecting H_0.

6. Discussions

Often prior information is known about the population. However, the sample collected may not reveal this information due to the sampling variability. Hence, it is worthwhile to build a model that satisfies the prior information and is the closest to the observed data.

In Theorem 1, $g(x)$ serves as the observed data, $\mathbb{C}$ serves as the prior information, and the distance between $g(x)$ and $\mathbb{C}$ is measured using the KL distance.

This paper proposed a method to compare between different populations based on a set of restrictions specified by the investigator. The restrictions were set in the form of moment constraints through one or more functions h. Setting different types of h compares different aspects of the distributions under consideration, e.g., $h(x) = x - c_1$ in (4) compares $g(x)$ and $g_i(x)$ regarding their means. When $\beta_i = 0$ in (4), then $g_i(x) = g(x), \forall x$. However, when $\beta_i \neq 0$, then $g_i(x)$ and $g(x)$ might differ in aspects other than only their values of $E(h(X))$.

For real data, one can obtain basic information from the data, including the shape. If any of the distributions under consideration are known to be approximately symmetric, using $h_1(x) = x - c_1$ and/or $h_2(x) = x^2 - c_2$ may be the first steps to determine if the distributions are different regarding their means and/or variances. However, if the distributions under consideration are known to be approximately skewed, then using $h(x) = \ln x$ would be more appropriate. In general, the reference distribution in (4) may be any of the k distributions, leaving the exponential distortion intact but with shifted parameters. When using the one-at-a-time method for different sample sizes, we chose the sample with the largest size as the reference, considering it to be the most trusted, and used its mean as μ_k.

Appendix A

Theorem A1. *Suppose the reference pdf is $g_2(x) = g(x) \sim N(\mu_2, \sigma_2^2)$, where μ_2, σ_2^2 are known. Then, the solution $g_1(x) = f^*(x)$ to the minimization problem (6) with $\mathbb{C} = \{f(x) : E_f(X) = \mu_1, Var_f(X) = \sigma_1^2\}$ is*

$$f^*(x) = g_1(x) = g(x)\, e^{\hat{\alpha}_1 + \hat{\beta}_1 x + \hat{\beta}_2 x^2} = N(\mu_1, \sigma_1^2), \tag{A1}$$

where $\hat{\alpha}_1 = -\ln\left[\int g(x)\, e^{\hat{\beta}_1 x + \hat{\beta}_2 x^2} dx\right], \hat{\beta}_1 = \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2},\ \hat{\beta}_2 = \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}.$

Proof. See Pathiravasan (2019). $\square$

When $\sigma_1^2 = \sigma_2^2$, then the $\mathbb{C}$ above reduces to $\mathbb{C} = \{f(x) : E_f(X) = \mu_1\}$, and then, it can be shown that the solution of the dual problem is $\hat{\beta} = \frac{\mu_1 - \mu_2}{\sigma_2^2}$. Recall, the same $\beta = \hat{\beta}$ was also observed as β_i in model (1) for $k = 2$, which justifies the optimality of the exponential tilt form of $N(\mu_1, \sigma_1^2)$ with $g(x) = N(\mu_2, \sigma_2^2)$ as the reference distribution for the one-way ANOVA model in Section 2.

Proof. [Proof of Theorem 2.] The cone dual to $\mathbb{C}$ in (11) is $\mathbb{C}^* = \{\beta_1(x_1 - x_2) + \beta_2(x_2 - x_3); \ \beta_1, \beta_2 \in \mathbb{R}\}$. Therefore, the corresponding dual problem is

$$\inf_{\beta_1, \beta_2 \in \mathbb{R}} \int_x e^{\beta_1(x_1 - x_2) + \beta_2(x_2 - x_3)} g(x)dx. \tag{A2}$$

In order to find the solution to (A2), we differentiate it with respect to β_1 and β_2 separately. The relevant score equations are

$$\int_x (x_i - x_{i+1}) e^{\beta_1(x_1 - x_2) + \beta_2(x_2 - x_3)} g(x)dx = 0, \ i = 1, 2. \tag{A3}$$

Using the expression for $g(x)$ and combining similar terms in the exponent, (A3) can be simplified as $E(X_i - X_{i+1}) = 0$, where $(X_1, X_2, X_3)'$ has a tri-variate normal distribution with a mean vector $(\mu_1 + \beta_1\sigma_1^2, \mu_2 + (\beta_2 - \beta_1)\sigma_2^2, \mu_3 - \beta_2\sigma_3^2)'$ and covariance matrix Σ. Setting the equations $E(X_i - X_{i+1}) = 0, i = 1, 2$ as

$$\mu_1 + \beta_1\sigma_1^2 = \mu_2 + (\beta_2 - \beta_1)\sigma_2^2, \ \ \mu_2 + (\beta_2 - \beta_1)\sigma_2^2 = \mu_3 - \beta_2\sigma_3^2,$$

we find the solution of the corresponding dual problem as given in (16). Using these solutions, one obtains the solution $f^*(x)$ in (14) with the expression of μ^* as stated in (15). $\square$

For $k > 3$, closed form solutions for $\hat{\beta}_i, \mu^*$ exist but are tedious or intractable (Pathiravasan, 2019). For $g(x)$ nonnormal cases, unique solutions for β_i's exist but may not be in closed forms and/or the final solution $f^*(x)$ may not be a known distribution.

Theorem A2. *For $k = 3$ with the same pdf $g(x) = N_3(\mu, \Sigma)$ as in Theorem 3, consider the minimization problem (13) with $\mathbb{C} = \{f(x); E_f(X_i) = E_f(X_{i+1}), Var_f(X_i) = Var_f(X_{i+1}), i = 1, 2\}$. The solution is given by*

$$f^*(x) = g(x) e^{\hat{\alpha}_1 + \hat{\beta}_1(x_1 - x_2) + \hat{\beta}_2(x_2 - x_3) + \hat{\beta}_3(x_1^2 - x_2^2) + \hat{\beta}_4(x_2^2 - x_3^2)} = N_3(\mu^*\mathbf{1}, \sigma^{2*}I_3),$$

where $\hat{\alpha}_1$ is the normalizing factor, μ^*, as in (15), and $\sigma^{2*} = \frac{3\sigma_1^2\sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2}$,
$\hat{\beta}_1 = \frac{-1}{3}\left(\frac{2\mu_1\sigma_2^2\sigma_3^2 - \mu_2\sigma_1^2\sigma_3^2 - \mu_3\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2\sigma_3^2}\right), \hat{\beta}_2 = \frac{-1}{3}\left(\frac{\mu_1\sigma_2^2\sigma_3^2 + \mu_2\sigma_1^2\sigma_3^2 - 2\mu_3\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2\sigma_3^2}\right)$,
$\hat{\beta}_3 = \frac{-1}{6}\left(\frac{\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 - 2\sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2\sigma_3^2}\right), \hat{\beta}_4 = \frac{-1}{6}\left(\frac{2\sigma_1^2\sigma_2^2 - \sigma_1^2\sigma_3^2 - \sigma_2^2\sigma_3^2}{\sigma_1^2\sigma_2^2\sigma_3^2}\right)$, and I_3 is the (3×3) identity matrix.

Proof. See Pathiravasan (2019). $\square$

References

1. Kedem, B.; Wolff, D.B.; Fokianos, K. Statistical comparison of algorithms, *IEEE Trans. Instrum. Meas.* **2004**, *53*, 770–776. [CrossRef]
2. Yeh, I.-C.; Lien, C-H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.* **2009**, *36*, 2473–2480. [CrossRef]
3. Fokianos, K.; Kedem, B.; Qin, J.; Short, D.A. A semiparametric approach to the one-way layout. *Technometrics* **2001**, *43*, 56–65. [CrossRef]
4. Jaynes, E.T. Information theory and statistical mechanics, *Phys. Rev.* **1957**, *106*, 620–130. [CrossRef]
5. Bhattacharya, B.; Dykstra, R. A general duality approach to I-projections, *J. Stat. Plan. Inference* **1995**, *47*, 203–216. [CrossRef]
6. Pathiravasan, C.H. Generalized Semiparametric Approach to the Analysis of Variance. Ph.D. Dissertation, Department of Mathematics, Southern Illinois University Carbondale, Carbondale, IL, USA, 2019.
7. Haberman, S.J. Adjustment by minimum discriminant information, *Ann. Stat.* **1984**, *12*, 971–988. [CrossRef]