Electronic Journal of Statistics
Vol. 3 (2009) 1161–1192
ISSN: 1935-7524
DOI: 10.1214/09-EJS518

# Calibration of the empirical likelihood method for a vector mean

**Sarah C. Emerson**

*Department of Biostatistics, Harvard School of Public Health*
*Boston, MA 02115*
*e-mail:* semerson@hsph.harvard.edu

**Art B. Owen**

*Department of Statistics, Stanford University*
*Stanford, CA 94305*
*e-mail:* art@stat.stanford.edu

**Abstract:** The empirical likelihood method is a versatile approach for testing hypotheses and constructing confidence regions in a non-parametric setting. For testing the value of a vector mean, the empirical likelihood method offers the benefit of making no distributional assumptions beyond some mild moment conditions. However, in small samples or high dimensions the method is very poorly calibrated, producing tests that generally have a much higher type I error than the nominal level, and it suffers from a limiting convex hull constraint. Methods to address the performance of the empirical likelihood in the vector mean setting have been proposed in a number of papers, including a contribution that suggests supplementing the observed dataset with an artificial data point. We examine the consequences of this approach and describe a limitation of their method that we have discovered in settings when the sample size is relatively small compared with the dimension. We propose a new modification to the extra data approach that involves adding two points and changing the location of the extra points. We explore the benefits that this modification offers, and show that it results in better calibration, particularly in difficult cases. This new approach also results in a small-sample connection between the modified empirical likelihood method and Hotelling's T-square test. We show that varying the location of the added data points creates a continuum of tests that range from the unmodified empirical likelihood statistic to Hotelling's T-square statistic.

**AMS 2000 subject classifications:** 62G10, 62H15.
**Keywords and phrases:** Empirical likelihood, nonparametric hypothesis testing, multivariate hypothesis testing.

Received October 2009.

## 1. Introduction

Empirical likelihood methods, introduced by Owen (1988), provide nonparametric analogs of parametric likelihood-based tests, and have been shown to perform remarkably well in a wide variety of settings. Empirical likelihood tests have been proposed for many functionals of interest, including the mean of a dis-

tribution, quantiles of a distribution, regression parameters, and linear contrasts in multisample problems.

In this paper, we focus on the use of the empirical likelihood method for inference about a vector mean, and investigate some of the small sample properties of the method. It has been widely noted (see, for example, Owen (2001), Tsao (2004a), or Chen, Variyath and Abraham (2008)) that in small samples or high dimensional problems, the asymptotic chi-square calibration of the empirical likelihood ratio statistic produces a test that generally does not achieve the nominal error rate, and can in fact be quite anti-conservative. Many authors have proposed adjustments to the empirical likelihood statistic or to the reference distribution in an attempt to remedy some of the small sample coverage errors. We briefly examine the ability of some these adjustments to correct the behavior of the empirical likelihood ratio test, and focus in particular on the method of Chen, Variyath and Abraham (2008) which involves adding an artificial data point to the observed sample. This approach offers several key benefits in both ease of computation and accuracy. We explore the consequences of the recommended placement of the extra point, and we demonstrate a limitation of the method that results in confidence regions equal to $\mathbb{R}^d$ in some settings. We propose a modification of the data augmentation that involves adding two balanced points rather than just one and changing the location of the added points. The balanced points preserve the sample mean of the augmented data set, which maintains the comparison between the sample mean and the hypothesized value. This modification addresses both the under-coverage issue of the original empirical likelihood method and the limitation of the Chen, Variyath and Abraham (2008) method. The locations of the new extra points are determined according to a parameter $s > 0$ which tunes the calibration of the resulting statistic. With an appropriate choice of $s$, these adjustments result in greatly improved calibration for small samples in high dimensional problems. Further, as $s \to \infty$, we find a small sample connection to Hotelling's T-square test. Simulation results demonstrate the effectiveness of the modified augmented empirical likelihood calibration.

We begin in Section 2 with a description of the basic setting and introduce some notation. We then outline the empirical likelihood method, and discuss the small sample issues of the method. In Section 3 we present previous proposals for calibrating the empirical likelihood method and compare the abilities of these proposals to address the various challenges for empirical likelihood in small samples. Section 4 introduces a modification of the data-augmentation strategy, and presents a result regarding the change in sample space ordering as the location of the extra points varies, connecting the empirical likelihood method to Hotelling's T-square test. We illustrate the improvement in calibration for several examples in Section 5, and conclude in Section 6 with a discussion of the results, and some ideas for future work and extensions of the methods presented here.

## 2. Background and notation

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sample of $n$ independent, identically distributed $d$-vectors, distributed according to $F_0$. We want to test a hypothesis regarding the value of $\mu_0 = \mathrm{E}_{F_0}(X_i)$, i.e., to test

$$H_0 : \mu_0 = \mu. \tag{1}$$

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ denote the sample mean, and let $\mathbf{S}$ denote the sample covariance matrix, which we assume to be full rank:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T.$$

Finally, let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be an invertible matrix satisfying $\mathbf{A}\mathbf{A}^T = \mathbf{S}$. Define the following standardized quantities: $Z_i = \mathbf{A}^{-1}\left(X_i - \bar{X}\right)$, $\bar{Z} = 0$, and $\eta = \mathbf{A}^{-1}\left(\mu - \bar{X}\right)$. We will use the standardized quantities to simplify notation in later sections.

### 2.1. Hotelling's T-square statistic

For the setting and hypothesis test described by (1), Hotelling's T-square statistic (Hotelling, 1931) is given by

$$T^2(\mu) = n\left(\bar{X} - \mu\right)^T \mathbf{S}^{-1}\left(\bar{X} - \mu\right).$$

Hotelling's T-square statistic is invariant under the group of transformations defined by $X \mapsto \tilde{X} = \mathbf{C}X$, where $\mathbf{C}$ is a full-rank matrix of dimension $d \times d$. The hypothesis being tested is then $H_0 : \mathrm{E}(\tilde{X}_i) = \tilde{\mu} = \mathbf{C}\mu$. In terms of standardized variables defined in the previous section, Hotelling's T-square statistic simplifies to

$$T^2(\mu) = n\eta^T\eta.$$

For testing the mean of a multivariate normal distribution, Hotelling's T-square test is uniformly most powerful invariant, and has been shown to be admissible against broad classes of alternatives (Stein, 1956; Kiefer and Schwartz, 1965). In the Gaussian case, the resulting statistic has a scaled $F_{d,n-d}$ distribution under the null distribution, given by:

$$\frac{n-d}{(n-1)d}T^2(\mu_0) \sim F_{d,n-d},$$

and therefore a hypothesis test of level $\alpha$ is obtained by rejecting the null hypothesis when

$$\frac{n-d}{(n-1)d}T^2(\mu) > F_{d,n-d}^{(1-\alpha)}.$$

The multivariate central limit theorem, along with Slutsky's theorem, justifies the use of this test for non-Gaussian data in large samples, and even in relatively small samples it is reasonably robust. Highly skewed distributions will of course require larger sample sizes to produce accurate inference using Hotelling's T-square test.

### 2.2. Empirical likelihood statistic

Owen (1988) and Owen (1990) proposed the ordinary empirical likelihood method, which we will denote by EL, for testing the hypothesis (1). It proceeds as follows: let

$$\mathcal{R}(\mu) = \max\left\{ \prod_{i=1}^{n} nw_i \;\middle|\; \sum_{i=1}^{n} w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}.$$

The log empirical likelihood ratio statistic is then given by

$$\mathcal{W}(\mu) = -2\log \mathcal{R}(\mu).$$

When positive weights $w_i$ satisfying the condition $\sum_{i=1}^{n} w_i X_i = \mu$ with $\sum_{i=1}^{n} w_i = 1$ do not exist, the usual convention is to set $\mathcal{R}(\mu) = -\infty$, and thus $\mathcal{W}(\mu) = \infty$. Under some mild moment assumptions, $\mathcal{W}(\mu_0) \xrightarrow{d} \chi^2_d$ as $n \to \infty$ (Owen, 1990), where $\mu_0$ is the true mean of the underlying distribution. The proof of this asymptotic behavior proceeds by showing that $\mathcal{W}(\mu_0)$ converges in probability to Hotelling's T-square statistic $T^2(\mu_0)$ as $n \to \infty$.

The motivation for the empirical likelihood ratio statistic is, as the name implies, an empirical likelihood ratio. The denominator is the likelihood of the observed mean under the empirical distribution: $\prod_{i=1}^{n} \left(\frac{1}{n}\right)$. The numerator is the maximized likelihood for a distribution $F$ that is supported on the sample and satisfies $E_F[X] = \mu$. It is easy to show that the empirical likelihood ratio statistic is invariant under the same group of transformations as Hotelling's T-square test, and this is a property that we will seek to maintain as we address the calibration issues of the test.

The asymptotic result above allows us to test hypotheses regarding the mean and to construct confidence intervals using the appropriate critical values arising from the chi-square distribution. However, the small sample behavior of this statistic is somewhat problematic for several reasons. First, if $\mu$ is not inside the convex hull of the sample, the statistic is undefined, or by convention taken to be $\infty$. A paper by Wendel (1962) calculates the probability $p(d, n)$ that the mean of a $d$-dimensional distribution is not contained in the convex hull of a sample of size $n$. The result is for distributions that are symmetric under reflections through the origin, and is found to be $p(d, n) = 2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}$. That is, the probability that the convex hull of the points does not contain the mean is equal to the probability that $W \leq d-1$ for a random variable $W \sim \text{Bin}(n-1, \frac{1}{2})$. (Note: an isomorphism between the binomial coin-flipping problem and this convex hull problem has still not been identified.) In small samples this convex hull constraint can be a significant problem, and even when the sample does contain the mean, the null distribution will be distorted somewhat by the convex hull effect.

A second issue that affects the small sample calibration of the empirical likelihood statistic is the fact that the first order term of the asymptotic expansion for the statistic is clearly not chi-square for small $n$, and is in fact bounded, as

we now demonstrate. Analogous to the definition of $\mathbf{S}$, define

$$\widetilde{\mathbf{S}}(\mu) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^T.$$

In the asymptotic expansion of the statistic $\mathcal{W}(\mu_0)$, the first order term is

$$\widetilde{T}^2(\mu_0) = n(\bar{X} - \mu_0)^T \widetilde{\mathbf{S}}(\mu_0)^{-1}(\bar{X} - \mu_0),$$

which is related to Hotelling's T-square statistic by

$$\widetilde{T}^2(\mu_0) = \frac{nT^2(\mu_0)}{T^2(\mu_0) + n - 1} \leq n$$

(Owen, 2001).

It is difficult to quantify the effect of the deviation of this term from its chi-square limit because the higher order terms clearly have a non-ignorable contribution in this setting since the EL statistic is unbounded. This does, however, indicate that the asymptotic approximation may be very far from accurate for small samples.

Together, these issues result in a generally very anti-conservative test in small samples. This is illustrated in the quantile-quantile and probability-probabilty plots shown in Figure 1, which are generated by simulating 5000 datasets consisting of 10 points from the multivariate Gaussian distribution in four dimensions, and then calculating the value of the EL statistic for the true mean $\mu_0 = \vec{0}$ for

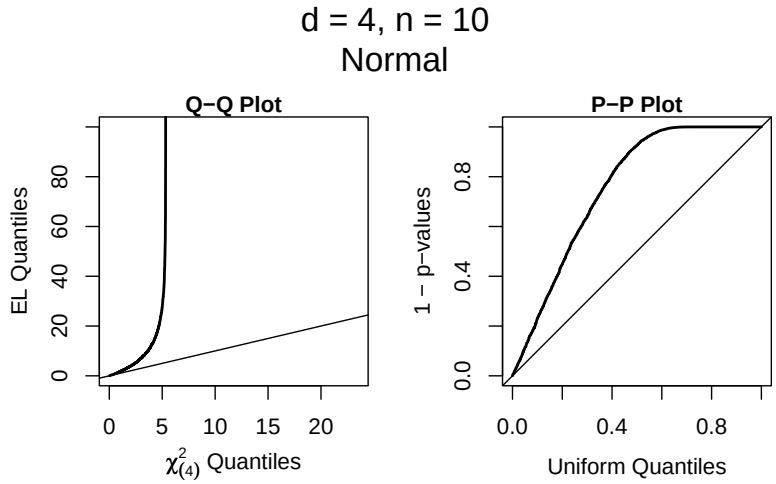


FIG 1. *Quantile-quantile and probability-probability plots for the null distribution of the empirical likelihood method (EL) statistic versus the reference $\chi^2$ distribution when the data consists of 10 points sampled from a 4 dimensional multivariate Gaussian distrubution. The x-axis corresponds to quantiles (left) or p-values (right) for the $\chi^2$ distribution and the y-axis is quantiles (left) or p-values (right) of the EL statistic.*

each dataset. We use this extreme setting of 10 points in four dimensions to make the calibration flaws readily apparent; these flaws persist, to a lesser degree, even in more reasonable settings. From these plots we can see the extremely anti-conservative behavior of this test: a test with nominal level $\alpha = 0.05$ would in fact result in a type I error rate of about 0.47. The example shown here is a difficult one, but even in more reasonable problems there can be a sizeable discrepancy between the nominal and actual type I error rates.

## 3. Calibration of empirical likelihood for a vector mean

There have been a number of suggestions for improving the behavior of the empirical likelihood ratio statistic in small samples. We give a brief description of several such calibration methods here; more in-depth discussion may be found in the references listed with each method. The simplest of these methods is to use an appropriately scaled F distribution (Owen, 2001) in place of the usual $\chi^2$ reference distribution calibration. This approach is motivated by the first order term of the empirical likelihood ratio statistic, which closely resembles Hotelling's T-square statistic. However, in many examples there is no improvement in the resulting calibration, and the convex hull issue is clearly not addressed.

Owen (1988) proposes using a bootstrap calibration, which involves resampling from the original data set to get new data sets $\{X_1^{(b)}, \ldots, X_n^{(b)}\}$ for $b = 1, \ldots, B$. Then for each bootstrap sample, the empirical likelihood ratio statistic $\mathcal{W}^{(b)}(\bar{X})$ is computed for the sample mean of the original data set using the resampled data. This resampling process is performed $B$ times, and the statistic $\mathcal{W}(\mu)$ is then compared to the distribution of values $\mathcal{W}^{(b)}(\bar{X}), b = 1, \ldots, B$ to give a bootstrap $p$-value. The bootstrap calibration does not directly address the convex hull problem, but if the empirical likelihood function is extended beyond the hull of the data in some way, the boostrap calibration can produce usable results even when $\mu$ is not in the convex hull of the data. The calibration resulting from this bootstrap process is generally reasonably good, but it is quite computationally intensive. As with most bootstrap processes, the performance is improved with a higher number of bootstrap repetitions.

DiCiccio, Hall and Romano (1991) show that the empirical likelihood method is Bartlett-correctable, and therefore the asymptotic coverage errors can be reduced from $O(n^{-1})$ to $O(n^{-2})$. They further demonstrate that even in small samples an estimated Bartlett correction offers a noticeable improvement. The Bartlett correction involves scaling the reference $\chi^2$ distribution by a factor that can be estimated from the data or computed from a parametric model, and therefore offers no escape from the convex hull. Since the Bartlett correction corresponds to shifting the slope of the reference line in the quantile-quantile plot, it is also clear that in the examples we consider here it will offer only a marginal benefit in improving calibration.

The empirical likelihood-t method, discussed in Owen (2001) and originally proposed by K. A. Baggerly in a 1999 technical report (source unavailable) is an

attempt to address the convex hull constraint by allowing the weighted mean to differ from the hypothesized mean in a constrained manner. This method does not retain the transformation invariance of the empirical likelihood method (Owen, 2001), and requires significantly more computation time as it introduces another parameter to be profiled out in the search for optimal weights.

Tsao (2001) and Tsao (2004b) discuss a calibration for the empirical likelihood method for a vector mean that involves simulating the exact distribution of the empirical likelihood ratio statistic when the underlying distribution of the data is Gaussian, and using this simulated distribution as the reference. There is no attempt to address the convex hull issue, but the resulting coverage levels do tend to be closer to the nominal levels when the convex hull constraint allows it.

Bartolucci (2007) suggests a penalized empirical likelihood that allows hypotheses outside the convex hull of the data by penalizing the distance between the mean $\nu$ of the reweighted sample distribution and the hypothesized mean $\mu$. While this approach does escape the convex hull issue, the choice of the penalty parameter is difficult to determine, and the method is very computationally intensive as it requires an extra search to minimize the penalty and it also relies on bootstrap calibration. In fact, the author recommends double bootstrap calibration, which becomes prohibitively expensive as the dimension of the problem increases. Clearly the benefit of this approach will depend on the choice of the penalty parameter, and it is unclear how much this modification improves the calibration of the test in the best case.

Finally, Chen, Variyath and Abraham (2008) suggest a calibration, which we will refer to henceforth as the adjusted empirical likelihood method (AEL), that proceeds by adding an artificial point to the data set and then computing the empirical likelihood ratio statistic on the augmented sample. The point is added in such a way as to guarantee that the hypothesized mean will be in the convex hull of the augmented data, thereby addressing the convex hull constraint. Chen, Variyath and Abraham discuss the asymptotic behavior of this modification, showing that as long as the additional point is placed in a reasonable way, the resulting statistic has the same limiting properties as the ordinary empirical likelihood ratio statistic. This approach is attractive from a computational standpoint, and appears to have good potential to influence the appropriateness of the calibration of the empirical likelihood method.

In summary, with the exception of the last two methods, these approaches do not address the convex hull constraint, and have varying degrees of success at correcting the small sample behavior of the empirical likelihood statistic. The AEL method has most convincingly overcome the convex hull issue and has further resulted in marked improvement in the calibration of the resulting statistic, so we explore their approach in greater depth.

### *3.1. Adjusted empirical likelihood*

Chen, Variyath and Abraham (2008) propose adding an additional point to the sample and then calculating the empirical likelihood statistic based on the aug-

TABLE 1

*Comparisons of the small-sample properties of the calibration methods discussed in Section 3. The first column of comparisons indicates the abilities of the methods to address the constraint that the hypothesized mean must be contained in the convex hull of the data. The second comparison column describes the degree to which the method improves the agreement between the achieved and nominal level of a hypothesis test, when a test of that level is possible given the convex hull constraint*

| Calibration Method | Escape Convex Hull | Small-sample Improvement |
|---|---|---|
| F-calibration | No | Somewhat |
| Bootstrap calibration | No | Yes |
| Bartlett correction | No | Somewhat |
| Tsao (2001) calibration | No | Yes |
| Tsao (2004) calibration | No | Yes |
| Bartolucci (2007) calibration | Yes | Somewhat |
| Chen, et al. (2008) calibration | Yes | Yes |

mented data set. Define the following quantities:

$$v^* = \bar{X} - \mu, \qquad r^* = \|v^*\|, \qquad \text{and} \qquad u^* = \frac{v^*}{r^*},$$

so $v^*$ is the vector from the sample mean to the hypothesized mean of the underlying distribution, $r^*$ is the distance between the sample mean and the hypothesized mean, and $u^*$ is a unit vector in the direction of $v^*$. In terms of these quantities, for the setting described in Section 2, the extra point $X_{n+1}$ that Chen, Variyath and Abraham suggest is

$$X_{n+1} = \mu - a_n\left(\bar{X} - \mu\right) = \mu - a_n v^* = \mu - a_n r^* u^*, \tag{2}$$

where $a_n$ is a positive constant that may depend on the sample size $n$. Then the resulting adjusted log empirical likelihood ratio statistic is

$$\mathcal{W}^*(\mu) = -2\log\mathcal{R}^*(\mu)$$

where

$$\mathcal{R}^*(\mu) = \max\left\{\prod_{i=1}^{n+1}(n+1)w_i \;\middle|\; \sum_{i=1}^{n+1} w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^{n+1} w_i = 1\right\}.$$

They recommend the choice $a_n = \frac{1}{2}\log(n)$, but discuss other options as well and state that as long as $a_n = o_p(n^{2/3})$ the first order asymptotic properities of the original log empirical likelihood ratio statistic are preserved for this adjusted statistic. It is easy to see that this modification also preserves the invariance of the ordinary empirical likelihood method. However, in the case of small samples or high dimensions, we have discovered that the AEL adjustment has a limitation that can make the chi-square calibration very inappropriate. The following Proposition describes this phenomenon.

**Proposition 3.1.** *With an extra point placed as proposed in Chen, Variyath and Abraham (2008) at* $X_{n+1} = \mu - a_n(\bar{X} - \mu)$*, the statistic* $\mathcal{W}^*(\mu) = -2\log\mathcal{R}^*(\mu)$ *is bounded above:*

$$\mathcal{W}^*(\mu) \le B(n, a_n) \equiv -2\left[n\log\left(\frac{(n+1)a_n}{n(a_n+1)}\right) + \log\left(\frac{n+1}{a_n+1}\right)\right].$$

*Proof.* We show that weights $\widetilde{w}_i$ given by

$$\begin{aligned}\widetilde{w}_i &= \frac{a_n}{n(a_n+1)} \qquad \text{for } i = 1,\dots,n\\ \widetilde{w}_{n+1} &= \frac{1}{a_n+1}\end{aligned}$$

always satisfy $\sum_{i=1}^{n+1}\widetilde{w}_i X_i = \mu$ when $X_{n+1} = \mu - a_n(\bar{X} - \mu)$:

$$\begin{aligned}\sum_{i=1}^{n+1}\widetilde{w}_i X_i &= \sum_{i=1}^{n}\widetilde{w}_i X_i + \widetilde{w}_{n+1}X_{n+1}\\ &= \sum_{i=1}^{n}\frac{a_n}{n(a_n+1)}X_i + \frac{1}{a_n+1}\left(\mu - a_n(\bar{X}-\mu)\right)\\ &= \frac{a_n}{a_n+1}\bar{X} - \frac{a_n}{a_n+1}\bar{X} + \frac{1}{a_n+1}\mu + \frac{a_n}{a_n+1}\mu\\ &= \mu.\end{aligned}$$

Then since clearly $\sum_{i=1}^{n+1}\widetilde{w}_i = 1$, we therefore have

$$\begin{aligned}\mathcal{R}^*(\mu) &= \max\left\{\prod_{i=1}^{n+1}(n+1)w_i \;\middle|\; \sum_{i=1}^{n+1}w_i X_i = \mu, w_i \ge 0, \sum_{i=1}^{n+1}w_i = 1\right\}\\ &\ge \prod_{i=1}^{n+1}(n+1)\widetilde{w}_i.\end{aligned}$$

So taking logarithms and multiplying by $-2$ we find that:

$$\begin{aligned}\mathcal{W}^*(\mu) &\le -2\sum_{i=1}^{n+1}\log\left[(n+1)\widetilde{w}_i\right]\\ &= -2n\log\left(\frac{(n+1)a_n}{(a_n+1)n}\right) - 2\log\left(\frac{n+1}{a_n+1}\right).\end{aligned}$$

□

This result clearly indicates the poor performance of the chi-square calibration for this statistic with small $n$ or large $d$, as this bound will in some cases be well below the $1-\alpha$ critical value of the $\chi^2_{(d)}$ reference distribution, which will make the chi-square calibrated $1-\alpha$ confidence intervals equal $\mathbb{R}^d$. Table 2

TABLE 2

*Maximum possible confidence level for a non-trivial chi-square calibrated confidence interval using the AEL method of Chen, Variyath and Abraham (2008). Confidence intervals with nominal level greater than the given values will include the entire parameter space. These numbers are for the case when $n = 10$ and $a_n = \frac{\log(n)}{2}$, for dimension ranging from 1 to 9. The upper bound for the adjusted log empirical likelihood ratio statistic for this $n$ and $a_n$ is $B(n, a_n) = 7.334$*

| Dimension $d$ | $P\left(\chi^2_{(d)} \leq B(n, a_n)\right)$ |
|---|---|
| 1 | 0.993 |
| 2 | 0.974 |
| 3 | 0.938 |
| 4 | 0.881 |
| 5 | 0.803 |
| 6 | 0.709 |
| 7 | 0.605 |
| 8 | 0.499 |
| 9 | 0.398 |

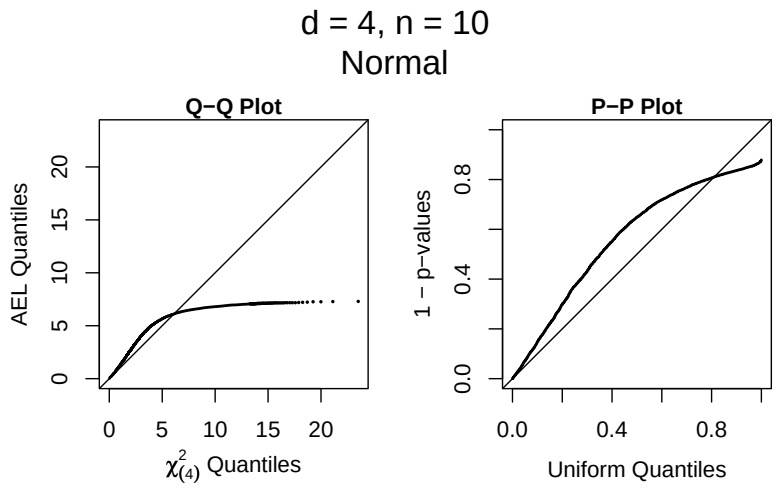


FIG 2. *Quantile-quantile and probability-probability plots for the null distribution of the adjusted empirical likelihood (AEL) statistic versus the reference $\chi^2$ distribution when the data consists of 10 points sampled from a 4 dimensional multivariate Gaussian distrubution. The x-axis corresponds to quantiles (left) or p-values (right) for the $\chi^2$ distribution and the y-axis is quantiles (left) or p-values (right) of the AEL statistic.*

displays the largest possible coverage level that does not result in the trivial parameter space confidence region using the AEL method, for the situation where 10 observations in $d$ dimensions. For small values of $d$ or large values of $n$, the bound will not cause much of a problem. For larger values of $d$ relative to $n$, the bound can be rather restrictive: from Table 2, we see that for $d \geq 3$, a 95% confidence region based on the $\chi^2_{(3)}$ reference distribution will include the entire space. Predictably, as $d$ increases for a fixed $n$, this issue becomes more pronounced. Figure 2 illustrates the bound phenomenon for 10 points in 4 di-

mensions, and also demonstrates suboptimal calibration even for values of $\alpha$ for which the boundedness of the statistic is not an issue.

## 4. Modified sample augmentation

Inspired by the approach of the AEL method, we propose augmenting the sample with artificial data to address the challenges mentioned above. However there are several key differences between their approach and ours. In contrast to the one point, placed at $X_{n+1} = \mu - \frac{1}{2}\log(n)(\bar{X} - \mu)$ as suggested by Chen, Variyath and Abraham, we propose adding two points to preserve the mean of the augmented data at $\bar{X}$. We also modify the placement of the points to

$$X_{n+1} = \mu - sc_{u^*}u^* \tag{3}$$
$$X_{n+2} = 2\bar{X} - \mu + sc_{u^*}u^* \tag{4}$$

where $c_{u^*} = \left(u^{*T}\mathbf{S}^{-1}u^*\right)^{-1/2}$. This choice of $c_{u^*}$ may be recognized as the inverse Mahalanobis distance of a unit vector from $\bar{X}$ in the direction of $u^*$, and will result in the points being placed closer to $\mu$ when the covariance in the direction of $\bar{X} - \mu$ is smaller, and farther when the covariance in that direction is larger. We will assume that $P(\bar{X} = \mu) = 0$ and therefore we do not have to worry about the case when $u^*$ is undefined because $v^*$ is zero.

With the points placed as described, the sample mean of the augmented dataset is maintained at $\bar{X}$. The scale factor $s$ can be chosen based on considerations that will be investigated in the next section. Having determined the placement of the extra points, we then proceed as if our additional points $X_{n+1}$ and $X_{n+2}$ were part of the original dataset, and compute $\widetilde{\mathcal{W}}(\mu) = -2\log(\widetilde{\mathcal{R}}(\mu))$ where

$$\widetilde{\mathcal{R}}(\mu) = \max\left\{\prod_{i=1}^{n+2}(n+2)w_i \;\middle|\; \sum_{i=1}^{n+2} w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^{n+2} w_i = 1\right\}.$$

We will refer to this statistic and method as the balanced augmented empirical likelihood method (BAEL) throughout the paper, to distinguish it from the unadjusted empirical likelihood statistic (EL) and the adjusted empirical likelihood statistic (AEL) of Chen, Variyath and Abraham. By the arguments of Chen, Variyath and Abraham (2008), it is easy to show that with a fixed value of $s$ this approach to augmenting the dataset has the same asymptotic properties as the ordinary empirical likelihood statistic. Other desirable properties of the EL statistic are retained as well, as addressed in the following Proposition.

**Proposition 4.1.** *Placing the points according to (4) preserves the invariance property of the empirical likelihood method under transformations of the form* $X \mapsto \tilde{X} = \mathbf{C}X$, *where* $\mathbf{C}$ *is an arbitrary full-rank matrix of dimension* $d \times d$.

*Proof.* The transformed $\tilde{u}$ is given by

$$\tilde{u} = \frac{\bar{\tilde{X}} - \tilde{\mu}}{\|\bar{\tilde{X}} - \tilde{\mu}\|} = \frac{\mathbf{C}\left(\bar{X} - \mu\right)}{\left\|\mathbf{C}\bar{X} - \mathbf{C}\mu\right\|},$$

and the transformed $\tilde{c}_{\tilde{u}}$ is given by

$$\tilde{c}_{\tilde{u}} = \left(\tilde{u}^T \left(\mathbf{CSC^T}\right)^{-1} \tilde{u}\right)^{-1/2} = \left\|\mathbf{C}\bar{X} - \mathbf{C}\mu\right\| \left[\left(\bar{X} - \mu\right)^T \mathbf{S}^{-1} \left(\bar{X} - \mu\right)\right]^{-1/2}.$$

Thus we have

$$\begin{aligned}
\tilde{c}_{\tilde{u}}\tilde{u} &= \left\|\mathbf{C}\bar{X} - \mathbf{C}\mu\right\| \left[\left(\bar{X} - \mu\right)^T \mathbf{S}^{-1} \left(\bar{X} - \mu\right)\right]^{-1/2} \frac{\mathbf{C}\left(\bar{X} - \mu\right)}{\left\|\mathbf{C}\bar{X} - \mathbf{C}\mu\right\|} \\
&= \mathbf{C}\left\|\bar{X} - \mu\right\| \left[\left(\bar{X} - \mu\right)^T \mathbf{S}^{-1} \left(\bar{X} - \mu\right)\right]^{-1/2} \frac{\bar{X} - \mu}{\left\|\bar{X} - \mu\right\|} \\
&= \mathbf{C}\left[u^{*T}\mathbf{S}^{-1}u^*\right]^{-1/2} u^*.
\end{aligned}$$

Finally, when we place $\tilde{X}_{n+1}$ based on the transformed data, we get

$$\tilde{X}_{n+1} = \tilde{\mu} - s\tilde{c}_{\tilde{u}}\tilde{u} = \mathbf{C}\mu - s\mathbf{C}\left[u^{*T}\mathbf{S}^{-1}u^*\right]^{-1/2} u^* = \mathbf{C}X_{n+1},$$

and similarly $\tilde{X}_{n+2} = \mathbf{C}X_{n+2}$. Using the fact that the original empirical likelihood method is invariant, we may conclude that this augmentation leaves the statistic invariant under the same group of transformations. □

One of the key differences between this approach and that of the AEL method is that as $\|\bar{X} - \mu\|$ increases the distance $\|\mu - X_{n+1}\|$ remains constant in our approach. This avoids the upper bound on $\mathcal{W}^*(\mu)$ that occurs using the AEL method. The other key idea in this placement of the extra points is to utilize distributional information estimated from the sample in the placement of the extra points.

The use of two points rather than just one is motivated by the original context of the empirical likelihood ratio statistic as a ratio of two maximized likelihoods: the numerator is the maximized empirical likelihood with the constraint that the weighted mean be $\mu$, and the denominator is the unconstrained maximized empirical likelihood which occurs at the sample mean $\bar{X}$. Adding just one point would necessarily change the sample mean, and therefore as different values of $\mu$ are tested, the resulting likelihood ratios are comparing the constrained maximum likelihoods to different sample means. Though the resulting weights in the denominator are the same no matter the value of the sample mean, the addition of two balanced points retains the spirit of the method and results in an interesting connection between the empirical likelihood ratio statistic and Hotelling's T-square statistic, as discussed further in Section 4.1.

In the next section we will address the choice of the scale factor $s$ on the resulting statistic, and in particular we will describe and prove a result connecting the empirical likelihood method and Hotelling's T-square test in small samples.

### 4.1. Limiting behavior of $\widetilde{W}(\mu)$ as $s \to \infty$

To reduce notation, we will work with the standardized versions of the data and the hypothesized mean as described in Section 2, so

$$\widetilde{R}(\mu) = \widetilde{R}(\mu; X_1, \ldots, X_{n+2}) = \widetilde{R}(\eta; Z_1, \ldots, Z_{n+2}) = \widetilde{R}(\eta)$$
$$\widetilde{W}(\mu) = \widetilde{W}(\mu; X_1, \ldots, X_{n+2}) = \widetilde{W}(\eta; Z_1, \ldots, Z_{n+2}) = \widetilde{W}(\eta)$$

where $Z_{n+1}$ and $Z_{n+2}$ are defined as follows. Using the transformed variables, we let

$$v = \bar{Z} - \eta = -\eta, \qquad r = \|v\| = \|\eta\|, \qquad \text{and} \qquad u = \frac{v}{r} = \frac{-\eta}{\|\eta\|}.$$

As these standardized observations have sample mean equal to zero and sample covariance matrix equal to $\mathbf{I}_d$, the extra points $Z_{n+1}$ and $Z_{n+2}$ are then given by

$$Z_{n+1} = \eta - su \qquad \text{and} \qquad Z_{n+2} = -\eta + su. \tag{5}$$

Then as the distance of these extra points from $\bar{Z} = 0$ increases, we are interested in the limiting behavior of the resulting adjusted empirical likelihood statistic, which is given by the following theorem:

**Theorem 4.2.** *For a fixed sample of size $n$*

$$\frac{2ns^2}{(n+2)^2}\widetilde{W}(\mu) \to T^2(\mu)$$

*as $s \to \infty$, where $T^2(\mu)$ is Hotelling's $T^2$ statistic.*

Here we present a brief outline of the proof; a complete and detailed proof is given in the Appendix. We will use the following notation throughout the proof of the theorem. As in Owen (2001), let $\lambda$ be the Lagrange multiplier satisfying

$$\sum_{i=1}^{n+2} \frac{1}{(n+2)} \frac{Z_i - \eta}{1 + \lambda^T (Z_i - \eta)} = 0 \tag{6}$$

so then the weights that maximize $\widetilde{R}(\eta)$ are given by

$$w_i = \frac{1}{(n+2)} \frac{1}{1 + \lambda^T (Z_i - \eta)}.$$

The proof of the theorem proceeds in the following steps:

1. First we establish that $\lambda^T u = o(s^{-1})$ using a simple argument based on the boundedness of the weights $w_i$.
2. We bound the norm of $\lambda$ by $\|\lambda\| = o(s^{-1/2})$ using the result from step 1 together with the fact that $\lambda^T(Z_i - \eta) > -1$ for all $i$, and the identity

$$\sum_{i=1}^{n+2} \lambda^T (Z_i - \eta) = \lambda^T (n+2)(-\eta).$$

3. Using the result from step 2, the unit vector in the direction of $\lambda$, given by $\theta$, is shown to satisfy $\theta^T u \to 1$. Then since from step 1 we have $\lambda^T u = o(s^{-1})$, we get $\|\lambda\| = o(s^{-1})$.
4. The limiting behavior of $\lambda$ is found to be $s^2 \lambda^T u \to \frac{(n+2)r}{2}$, using the bound from step 3 together with the constraint given by equation (6), and the identity
$$\frac{1}{1+x} = 1 - x + \frac{x^2}{1+x}.$$
This gives $\|\lambda\| = O(s^{-2})$.
5. Finally we use the limiting behavior of $\lambda$ from step 4 to get $\frac{2ns^2}{(n+2)^2}\widetilde{W}(\mu) \to T^2$. This is done by substituting the expression for $\lambda$ from step 4 into the expression for $\widetilde{W}(\eta)$:
$$\widetilde{W}(\eta) = -2\sum_{i=1}^{n+2} \log\left[(n+2)w_i\right]$$
and using the Taylor series expansion for $\log(x)$ as $x \to 1$.

This proof differs in several key ways from the usual empirical likelihood proofs, and these five steps are presented in full detail in Sections A.1–A.5 of the appendix.

We mentioned in Section 2.2 that asymptotically the empirical likelihood test becomes equivalent to Hotelling's T-square test under the null hypothesis as $n \to \infty$, but this theorem extends that relationship. This result provides a continuum of tests ranging from the ordinary empirical likelihood method to Hotelling's T-square test for any sample size. The magnitude of $s$ that is required to achieve reasonable convergence to Hotelling's test depends on the dimension and sample size.

## 5. Results

First we present the results of simulations to compare the accuracy of the chi-square calibration for the original empirical likelihood method (EL), the Chen, Variyath and Abraham (2008) adjusted empirical likelihood method (AEL), and our balanced augmented empirical likelihood method (BAEL) in Section 5.1. Then we illustrate the effect of the $s$ parameter on the relationship of the BAEL method to the original empirical likelihood method and to Hotelling's T-square test in Section 5.2.

### *5.1. Calibration results*

To compare the calibration of EL, AEL, and BAEL, we performed numerical comparisons based on simulated datasets for a variety of settings. We considered four combinations of sample size and dimension: $(d, n) = (4, 10), (4, 20), (8, 20),$

TABLE 3
*Skewness and kurtosis of example distributions*

| Marginal Distribution | Skewness | Kurtosis |
|---|---|---|
| Normal(0, 1) | 0 | 0 |
| t(3) | 0 | $+\infty$ |
| Double Exponential(1) | 0 | 3 |
| Uniform | 0 | -1.2 |
| Beta(0.1, 0.1) | 0 | -1.875 |
| Exponential(3) | 2 | 6 |
| F(4, 10) | 4 | 54 |
| Chi-square(1) | $2\sqrt{2}$ | 12 |
| Gamma(0.25, 0.1) | 4 | 24 |

and $(8, 40)$. For each combination, we simulated datasets from nine different distributions with independent margins. The distributions were chosen to represent a range of skewness and kurtosis so that we could evaluate the effects of higher moments on the calibration of the method. The skewness and kurtosis of the chosen distributions are listed in Table 3. We compared the chi-square calibrations of EL, AEL, and BAEL by creating quantile-quantile plots of the log empirical likelihood ratio statistics versus the appropriate chi-square distribution. Figures 3–10 show the resulting improvement in chi-square calibration using our BAEL method. We also plotted the $p$-values resulting from the chi-square calibration versus uniform quantiles in the corresponding probability-probability plots, to give a better indication of the coverage errors of the different methods. In each figure, the black lines or points represent the ordinary EL method; the red lines or points represent the AEL method of Chen, Variyath and Abraham; and the green lines or points are the results of our BAEL statistic. In the probability-probability plots, we have also included a blue line for the $p$-values resulting from Hotelling's T-square test. All of these figures were produced using $s = 1.9$; more discussion of the choice of $s$ will be given in Section 6.

These plots demonstrate the marked improvement in calibration achieved by our method: for symmetric distributions, the actual type I error is almost exactly the nominal level, particularly in the upper right regions of the plots where most hypothesis testing is focused. For the skewed distributions, the accuracy of the calibration depends on the degree of skewness and also on the kurtosis of the distributions. We find that it is harder to correct the behavior of empirical likelihood in skewed and highly kurtotic distributions, but even in the case of the Gamma(1/4, 1/10) distribution we have acheived distinct improvement over the other two versions of empirical likelihood. We have also essentially matched the calibration performance of Hotelling's T-square test even though the value of the scale factor $s$ is not large enough to have forced convergence to Hotelling's test, as will be addressed in Section 5.2. Thus we are still in the empirical likelihood setting, but with significantly improved accuracy for our test.

Note also that though the behavior in skewed distributions is not completely corrected by our calibration, it appears from the quantile-quantile plots that a

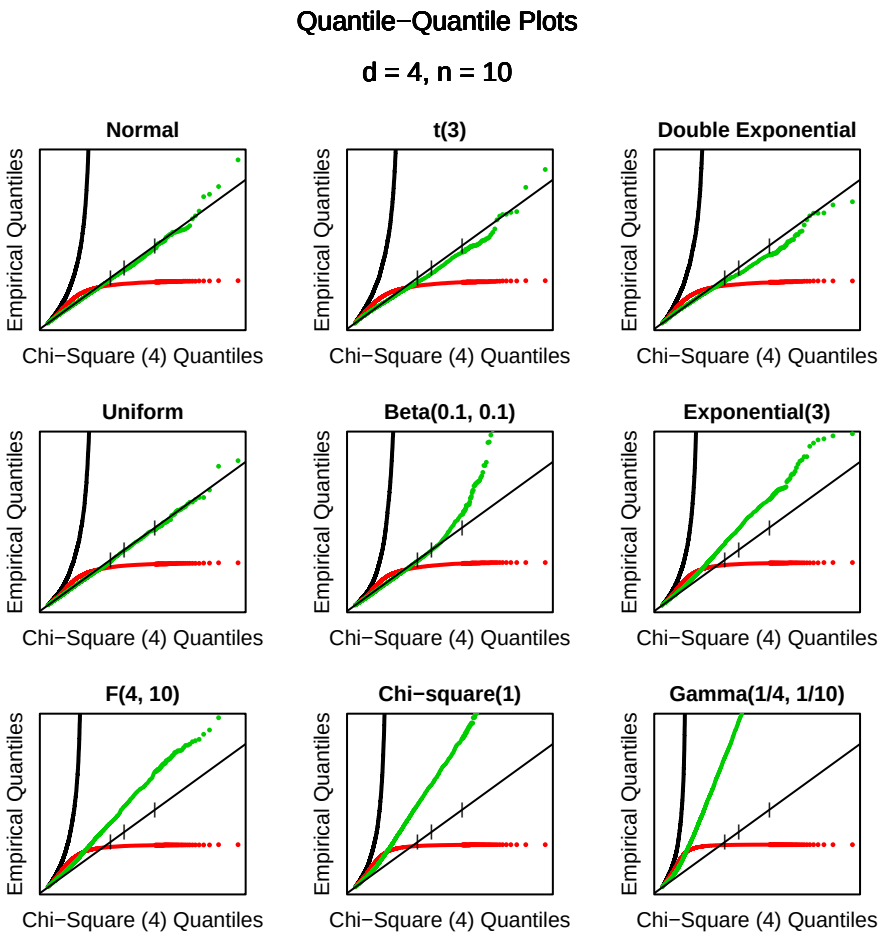


FIG 3. *Quantile-quantile plots for $d = 4$, $n = 10$. The $x$-axis has quantiles of the $\chi^2_{(4)}$ distribution, and the $y$-axis is quantiles of the ordinary EL statistic (black), the AEL statistic (red), and our BAEL statistic (green). Reading across the rows, the distributions are arranged in order of increasing skewness and then increasing kurtosis. The first five distributions are symmetric. Black tick marks on the $y = x$ line indicate the 90%, 95%, and 99% quantiles of the reference distribution.*

Bartlett correction might result in a marked improvement by shifting the slope of the reference distribution line. A Bartlett correction is clearly not as likely to result in improvement for the EL and AEL statistics, as the quantile-quantile plots for those methods versus the reference chi-square distribution are quite non-linear.

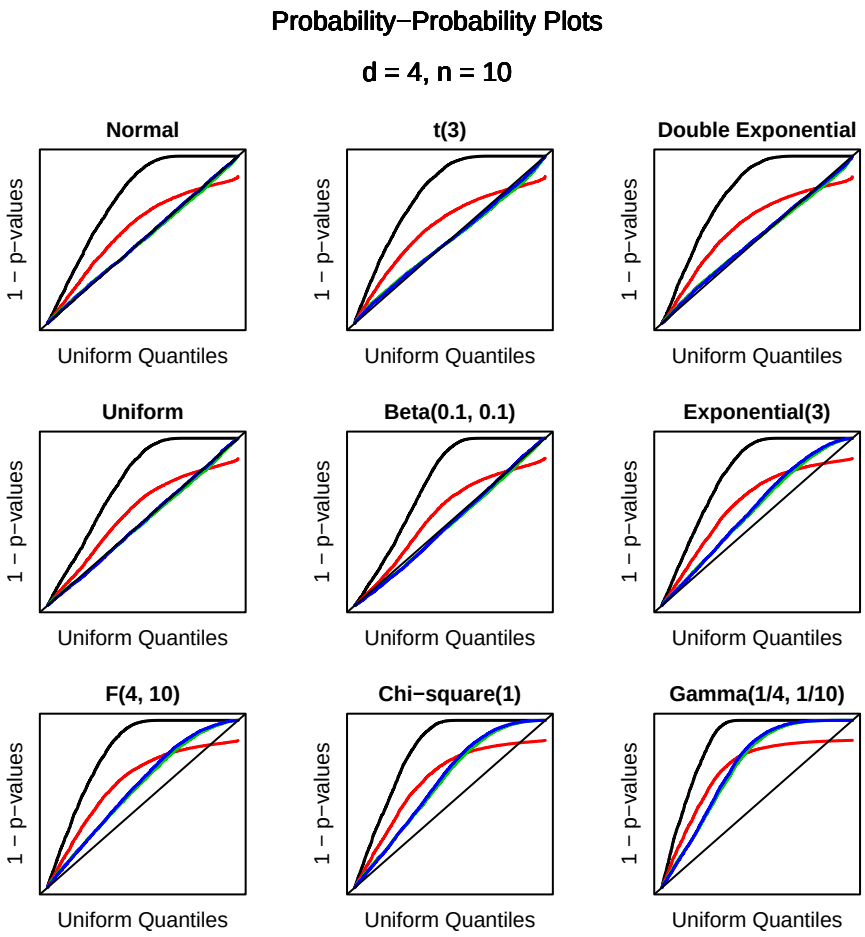


FIG 4. *Probability-Probability plots for $d = 4, n = 10$, for the same scenarios as illustrated in Figure 3. The x-axis is uniform quantiles, and the y-axis is $1 - p$-values computed from the $\chi^2(4)$ reference distribution for the ordinary EL statistic (black), the AEL statistic (red), and the BAEL statistic (green). Hotelling's T-square $1 - p$-values are also included on this plot (blue).*

### *5.2. Sample space ordering results*

Next we explored the degree to which our new calibration deviates from the ordinary empirical likelihood method to agree with Hotelling's, as a function of the scale factor $s$. Two tests are functionally equivalent if they order the possible samples in the same way, and therefore will always come to the same conclusion. Otherwise, if the tests produce different orderings of possible samples, they may make different decisions on the same dataset. For instance, the two-tailed $t$-test

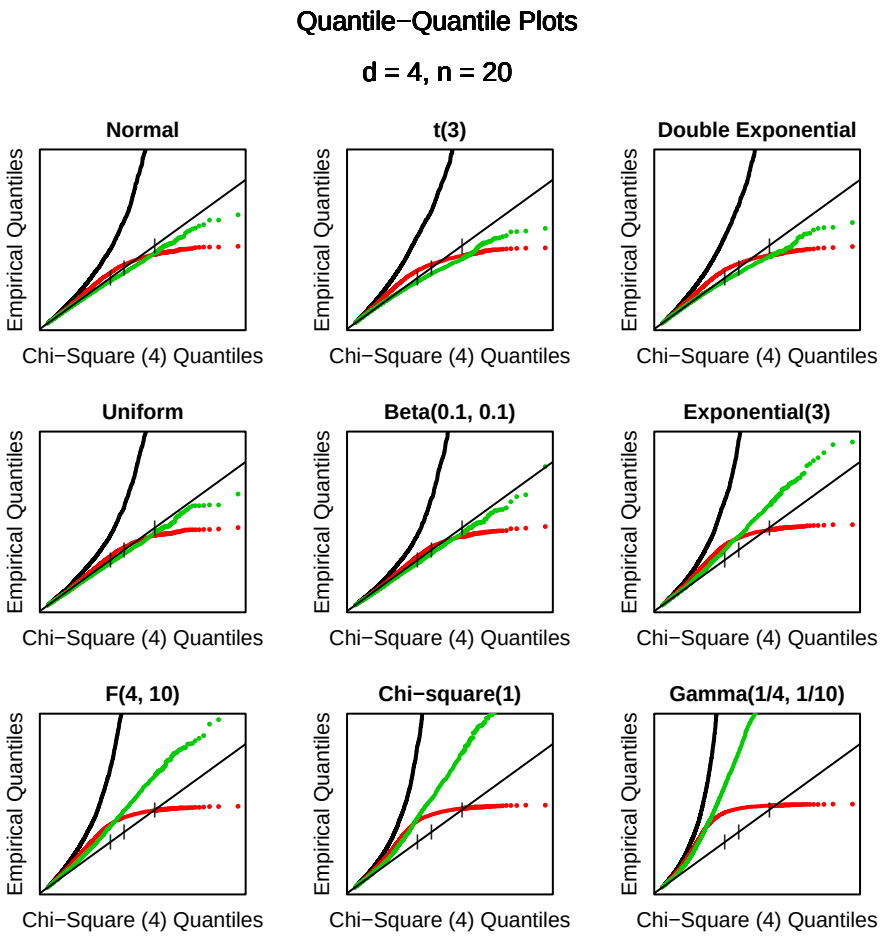


FIG 5. *Quantile-quantile plots for* $d = 4, n = 20$.

for a univariate mean is equivalent to the $F$-test that results from squaring the $t$ statistic: though these two tests have different reference distributions, they will always make the same decision for any given sample. In contrast, Pearson's chi-square test for independence in $2 \times 2$ tables orders the sample space differently than Fisher's exact test does, and thus these two tests may come to different conclusions. The important idea here is the ordering that different tests impose on the sample space determines the properties of the tests, such as their power against various alternatives.

We have shown that as $s$ increases, our BAEL statistic will become equivalent to Hotelling's T-square statistic, but we would like to explore the extent to which

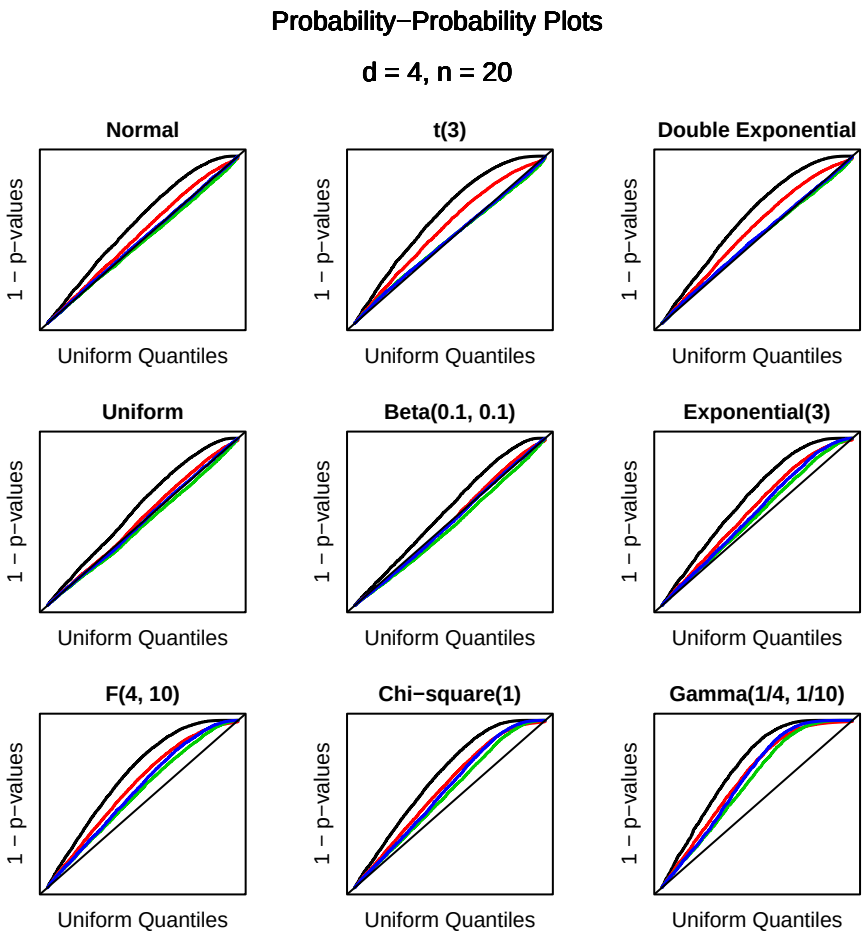


FIG 6. *Probability-Probability plots for* $d = 4, n = 20$.

this is true for small values of $s$. To do this, we generated 100 datasets, each consisting of 40 observations from a standard multivariate Gaussian distribution in 8 dimensions. For each dataset, we computed Hotelling's T-square statistic $T^2(\mu_0)$, the EL statistic $\mathcal{W}(\mu_0)$, and the BAEL statistic $\widetilde{\mathcal{W}}(\mu_0)$. We considered how the three statistics ordered different samples when testing the true null hypothesis by ranking the datasets according to each of the statistics. Figure 11 plots the ranking of the samples according to the BAEL statistic on the $y$-axis versus the ranking according to Hotelling's T-square statistic on the $x$-axis. The value of $s$ increases as powers of 2 from the top left plot to the bottom right. These same samples and choices of $s$ are shown again in Figure 12, except now the $x$-axis is the rank according to the EL statistic.

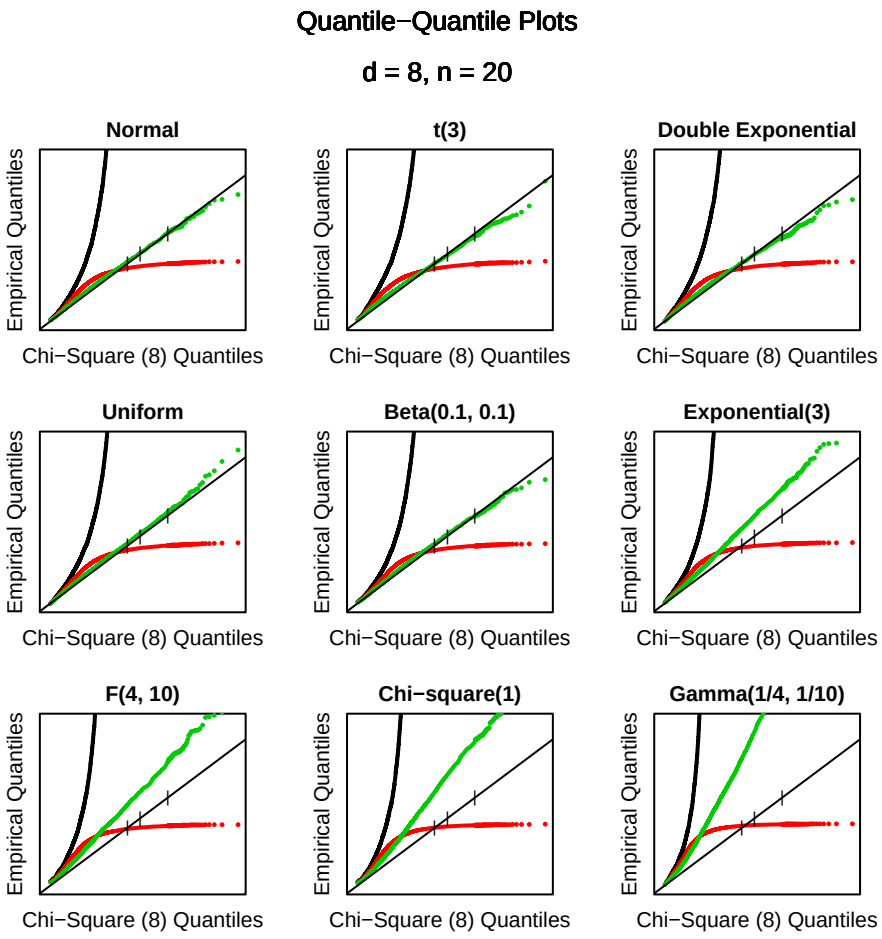


FIG 7. *Quantile-quantile plots for* $d = 8, n = 20$.

These figures demonstrate the convergence of the sample space ordering to that of Hotelling's T-square statistic as $s$ increases. From these figures we can see, for example, that for the value $s = 1.9$ used in the calibration simulations the ordering imposed by the BAEL statistic has not yet converged to the ordering produced by Hotelling's T-square statistic. It is important to note that though the sample space ordering of the new augmented empirical likelihood statistic looks to be identical to that of Hotelling's statistic when $s = 16$, this does not mean that the relationship is linear yet. We also note that for different combinations of the underlying distribution, sample size, and dimension, the same value of $s$ will produce different ordering discrepancies between the aug-

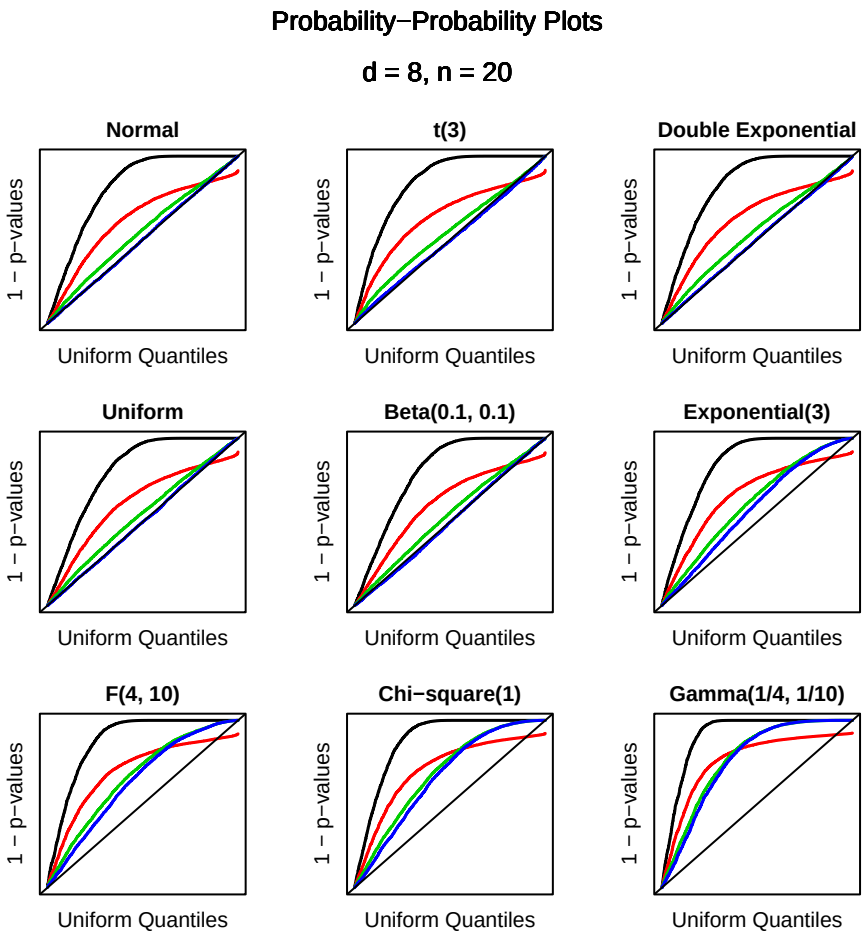


FIG 8. *Probability-Probability plots for* $d = 8, n = 20$.

mented empirical likelihood method and Hotelling's T-square statistic, but the qualitative behavior as $s$ increases will be preserved.

## 6. Discussion

We have introduced and explored many of the properties of a new augmented data empirical likelihood calibration. It has performed remarkably well in difficult problems with quite small sample sizes, and produces a versatile family of tests that allow an investigator to take advantage of both the data-driven confidence regions of the empirical likelihood method and the accurate calibration of Hotelling's T-square test.

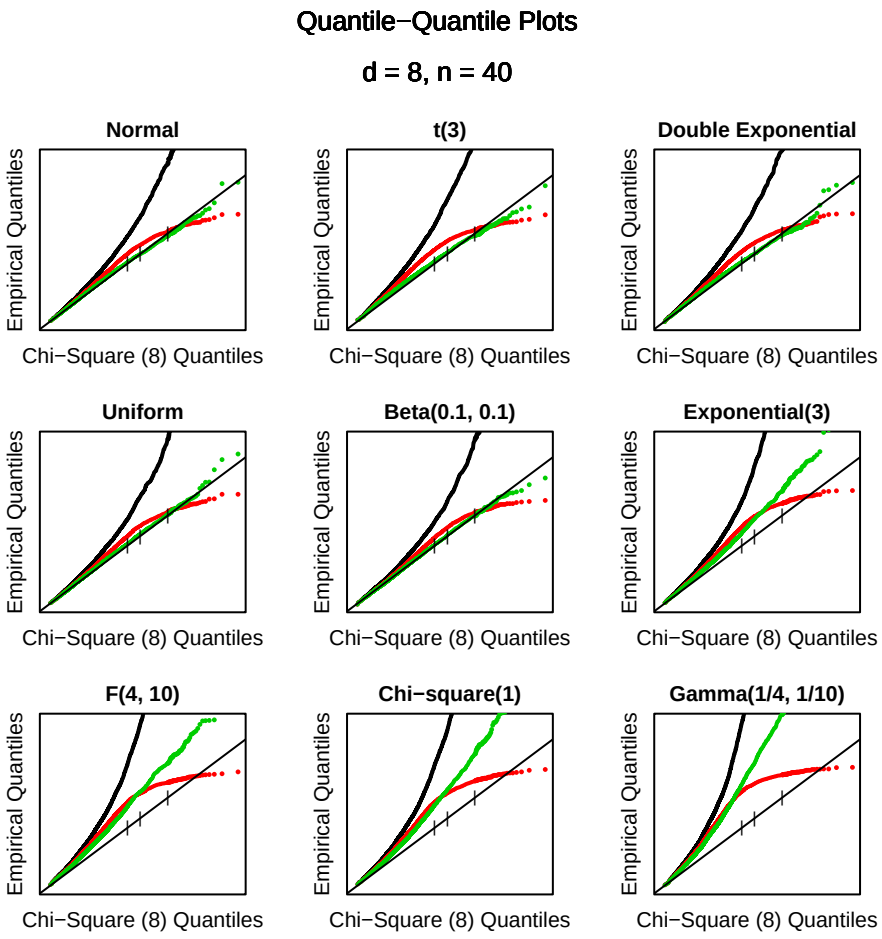


FIG 9. *Quantile-quantile plots for* $d = 8, n = 40$.

In additional simulations we have explored the effect of the scale factor $s$ on the resulting chi-square calibration of the BAEL statistic. We found that there is some variability in the value of $s^*(d)$ that produces the best $\chi^2_{(d)}$ calibration for a given dimension, but the range is fairly tight, from approximately 1.6 for $d = 2$ to 2.5 for $d = 30$. The optimal value $s^*(d)$ was chosen to be the value that gave the best overall fit to the $\chi^2_{(d)}$ distribution, as judged by the Kolmogorov-Smirnov statistic. The default value $s^*(d)$ warrants more detailed investigation, and will be explored further in later work.

We would like to investigate the potential of a Bartlett correction to improve the calibration in skewed samples. Since estimating the correction factor for a

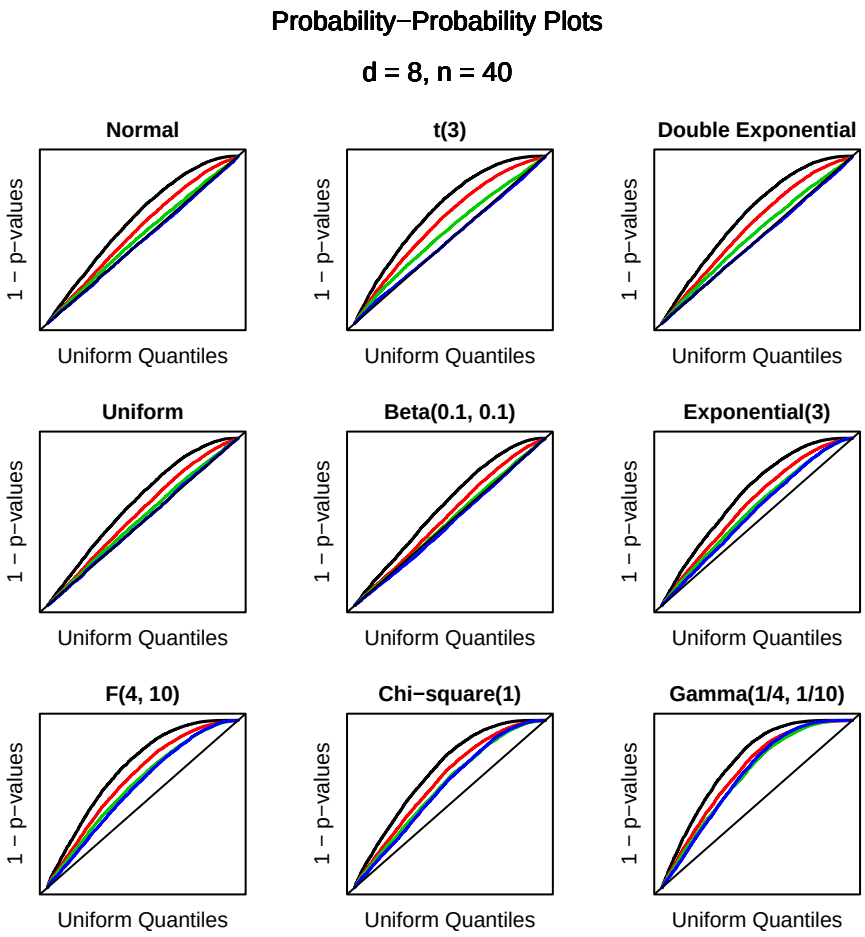


FIG 10. *Probability-Probability plots for* $d = 8, n = 40$.

Bartlett correction involves estimating fourth moments, it will be a challenge in small samples and high dimensions, but it does appear that there may be significant gains possible. The linearity of the quantile-quantile plots in the skewed distributions indicates that perhaps the skewness just scales the chi-square distribution of the augmented empirical likelihood statistic, but does not otherwise significantly alter it. This certainly warrants further exploration and theoretical justification.

Concurrent work by Liu and Chen (2009) has explored the use of two additional data points in another context. They kindly shared a preprint of their article with us as we were finishing work on our approach. Liu and Chen (2009)

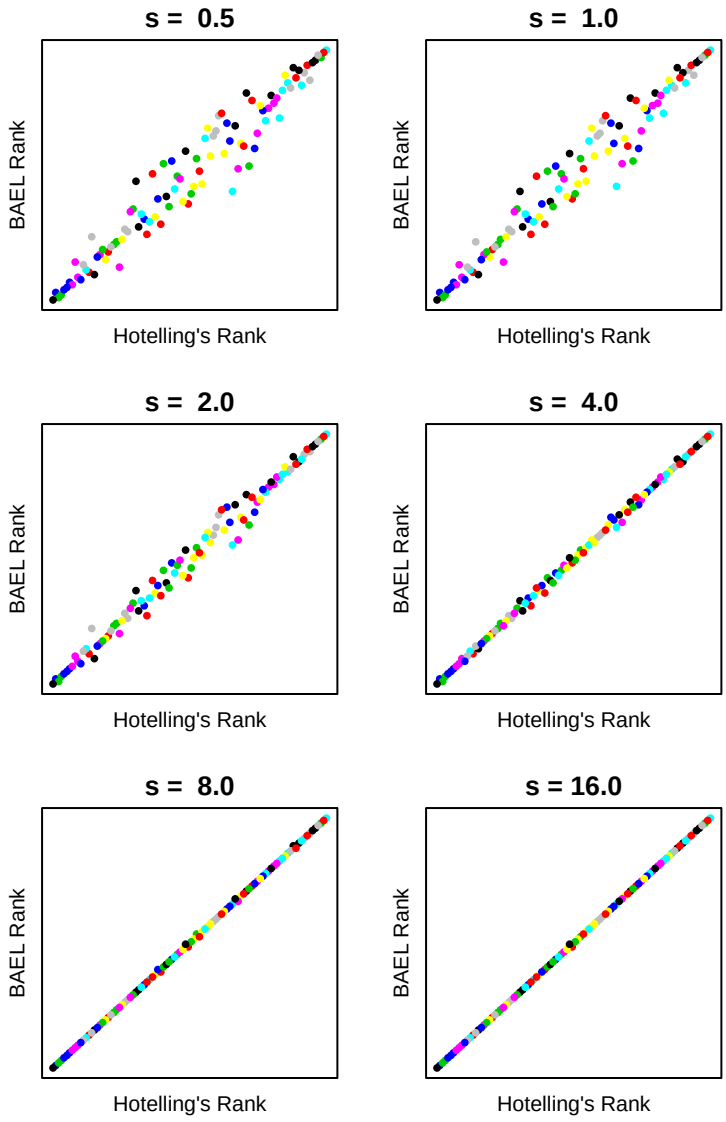


FIG 11. *Comparing the ranking of 100 samples according to Hotelling's T-square statistic (x-axis) vs. the BAEL statistic (y-axis) as s increases from 0.5 to 16.*

use different criteria for determining the placement of the extra points, and they investigate a connection between their resulting method and the Bartlett correction for empirical likelihood.

We have not addressed the power of the resulting test in this work, but we have made preliminary investigations into the effect of our modification on the

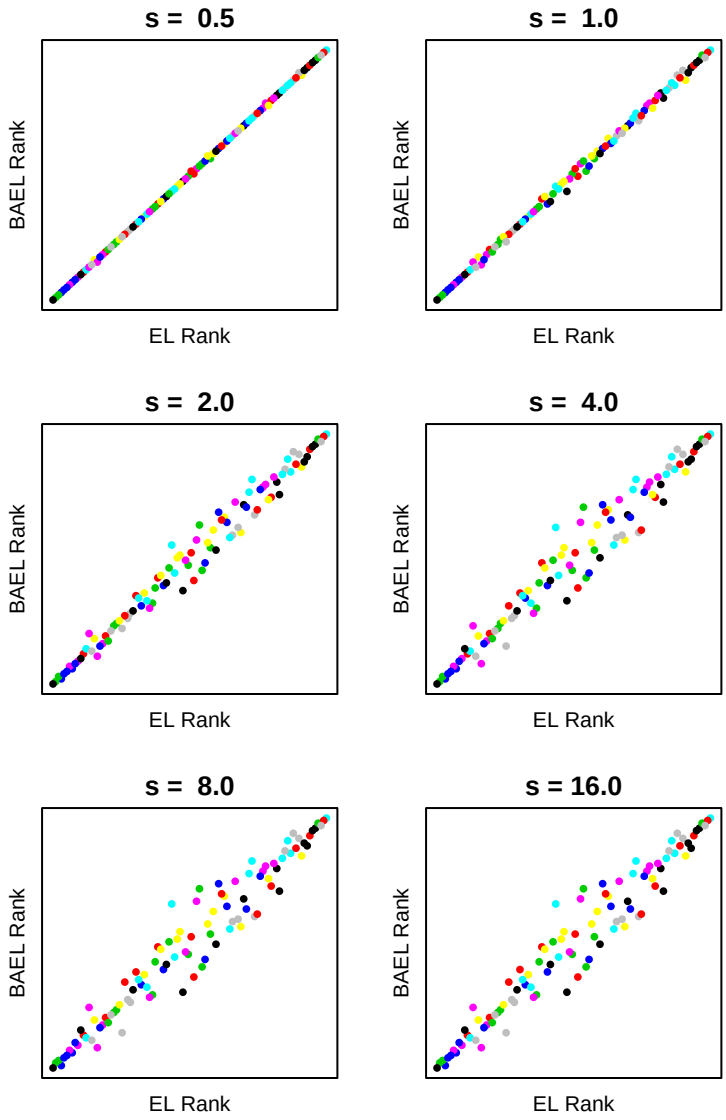


FIG 12. *Comparing the ranking of the 100 samples from Figure 11 according to the EL statistic (x-axis) vs. the BAEL statistic (y-axis) as s increases from 0.5 to 16.*

power of the competing tests. As might be expected, we have found that the power of BAEL is between that of the ordinary empirical likelihood and the power of Hotelling's T-square test. The relationship described in Theorem 4.2 explains this behavior on a heuristic level, and also indicates that as $s$ increases, the power curve of the augmented empirical likelihood test will more closely

resemble the power curve from Hotelling's T-square. For most of the examples and alternatives that we explored, the power of the ordinary empirical likelihood and the power of Hotelling's test were very close.

The connection to Hotelling's T-square test may prove to be especially interesting in the multi-sample setting. This result has potential implications beyond the one-sample mean setting, where it is largely of theoretical interest. In multi-sample settings, this relationship, combined with the generality of the empirical likelihood method, might be useful in extending Hotelling's test to scenarios where it currently does not apply, such as unequal group sizes with different variances. The use of the empirical likelihood-Hotelling's T-square continuum could enable us to produce tests with the accuracy of Hotelling's T-square, but with the flexibility and relaxed assumptions of the empirical likelihood framework. Similar extensions may also be made to regression problems.

## 7. Acknowledgements

This work was supported by grants DMS-0906056 and DMS-0604939 from the U.S. National Science Foundation.

## Appendix A: Proof of Theorem 4.2

Recall that $Z_i$ are the standardized variables $Z_i = \mathbf{A}^{-1}\left(X_i - \bar{X}\right)$, leading to the standardized versions of the sample mean $\bar{Z} = 0$, and the hypothesized mean $\eta = \mathbf{A}^{-1}\left(\mu - \bar{X}\right)$. We have defined the following quantities:

$$v = \bar{Z} - \eta = -\eta, \qquad r = \|v\| = \|\eta\|, \qquad \text{and} \qquad u = \frac{v}{r} = \frac{-\eta}{\|\eta\|}.$$

Note that, by (5), $Z_{n+1} - \eta = -su$ and $Z_{n+2} - \eta = -2\eta + su = (2r + s)u$. In the following, $n$ is fixed and all limits and $O(\cdot)$ and $o(\cdot)$ notations are to be interpreted as $s \to \infty$.

### *A.1. Step 1*

Since $\lambda$ satisfies

$$0 = \sum_{i=1}^{n} w_i(Z_i - \eta) + w_{n+1}(-su) + w_{n+2}(2r + s)u,$$

we have

$$\sum_{i=1}^{n} w_i(Z_i - \eta) + w_{n+2}2ru = (w_{n+1} - w_{n+2})\, su.$$

Dividing both sides by $s$, and multiplying on the left by $u^T$ gives

$$\frac{1}{s}\left(\sum_{i=1}^{n} w_i u^T(Z_i - \eta) + w_{n+2}2r\right) = w_{n+1} - w_{n+2}.$$

Now because $\eta$ is inside the convex hull of the augmented sample, $0 < w_i < 1$ and

$$\left\|u^T(Z_i - \eta)\right\| \leq \max_{i=1,\ldots,n} \left\{(Z_i - \eta)^T(Z_i - \eta)\right\} = O(1),$$

we have $w_i u^T(Z_i - \eta) = O(1)$. Similarly, $w_{n+2}2r = O(1)$, and therefore,

$$(n+2)\left(w_{n+1} - w_{n+2}\right) = \frac{1}{1 - s\lambda^T u} - \frac{1}{1 + (2r+s)\lambda^T u} = O(s^{-1}). \tag{7}$$

Thus since $1 - s\lambda^T u > 1 \Rightarrow 1 + (2r+s)\lambda^T u < 1$ and vice versa, we must have

$$\lambda^T u = o(s^{-1}). \tag{8}$$

### *A.2. Step 2*

Since $0 < w_i < 1$ for $i = 1, \ldots, n+2$, we have that $1 + \lambda^T(Z_i - \eta) > 0$ which implies $\lambda^T(Z_i - \eta) > -1$ for all $i$. Then using the fact that

$$\sum_{i=1}^{n+2} \lambda^T(Z_i - \eta) = \lambda^T(n+2)(-\eta) = (n+2)r\lambda^T u$$

and the bound given by (8), we conclude that

$$\max_{i=1,\ldots,n+2} \left\{1 + \lambda^T(Z_i - \eta)\right\} \leq 1 + (n+2)r\lambda^T u + (n+1) = O(1). \tag{9}$$

Now we employ the identity

$$\frac{1}{1+x} = 1 - \frac{x}{1+x} \tag{10}$$

to get

$$\begin{aligned} 0 &= \sum_{i=1}^{n+2} \frac{Z_i - \eta}{1 + \lambda^T(Z_i - \eta)} \\ &= \sum_{i=1}^{n+2}(Z_i - \eta) - \sum_{i=1}^{n+2} \frac{(Z_i - \eta)\left(\lambda^T(Z_i - \eta)\right)}{1 + \lambda^T(Z_i - \eta)}. \end{aligned}$$

Letting $\lambda = \|\lambda\|\,\theta$, rearranging the above equality, and multiplying both sides by $\lambda^T$, we have

$$\sum_{i=1}^{n+2} \lambda^T(Z_i - \eta) = \|\lambda\|^2 \sum_{i=1}^{n+2} \frac{\theta^T(Z_i - \eta)\left(\theta^T(Z_i - \eta)\right)}{1 + \lambda^T(Z_i - \eta)},$$

which gives $r\lambda^T u = \|\lambda\|^2 \theta^T \widetilde{\mathbf{S}} \theta$ where

$$\widetilde{\mathbf{S}} = \sum_{i=1}^{n+2} \frac{(Z_i - \eta)(Z_i - \eta)^T}{1 + \lambda^T(Z_i - \eta)}.$$

Then letting $\mathbf{S}^* = \sum_{i=1}^{n+2}(Z_i-\eta)(Z_i-\eta)^T$ and substituting in the bound (9) on $\lambda^T(Z_i-\eta)$ from above, we have

$$\begin{aligned}\|\lambda\|^2\,\theta^T\mathbf{S}^*\theta &\le \|\lambda\|^2\,\theta^T\widetilde{\mathbf{S}}\theta\left[\max_{i=1,\ldots,n+2}\{1+\lambda^T(Z_i-\eta)\}\right]\\ &= r\lambda^T u\left[\max_{i=1,\ldots,n+2}\{1+\lambda^T(Z_i-\eta)\}\right]\\ &= o(s^{-1})O(1).\end{aligned}$$

Furthermore, $\theta^T\mathbf{S}^*\theta \ge l_d$ where $l_d$ is the smallest eigenvalue of the matrix $\sum_{i=1}^{n}(Z_i-\eta)(Z_i-\eta)^T$, and thus $\left(\theta^T\mathbf{S}^*\theta\right)^{-1} \le l_d^{-1} = O(1)$, so $\|\lambda\|^2 = O(1)\times o(s^{-1})O(1)$. Therefore,

$$\|\lambda\| = o(s^{-1/2}). \tag{11}$$

### *A.3. Step 3*

Let $(Z_i-\eta) = f_i u + r_i$ where $f_i = (Z_i-\eta)^T u$ and $r_i = (Z_i-\eta) - f_i u$ so $r_i^T u = 0$ for all $i = 1,\ldots,n+2$. Note that

$$r_{n+1} = r_{n+2} = 0 \tag{12}$$

since both $(Z_{n+1}-\eta)$ and $(Z_{n+2}-\eta)$ are multiples of $u$. The remaining $r_i$, for $i = 1,\ldots,n$ satisfy

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n}(Z_i-\eta) - \left[(Z_i-\eta)^T u\right]u = 0. \tag{13}$$

Also, we have

$$\begin{aligned}\sum_{i=1}^{n} f_i r_i &= \sum_{i=1}^{n} f_i\left[(Z_i-\eta) - f_i u\right]\\ &= \sum_{i=1}^{n}(Z_i-\eta)(Z_i-\eta)^T u - \sum_{i=1}^{n} f_i^2 u\\ &= \left[(n-1)\mathbf{I}_d + n\eta\eta^T\right]u - \left(\sum_{i=1}^{n} f_i^2\right)u\\ &= \left[(n-1) + nr^2 - \sum_{i=1}^{n} f_i^2\right]u.\end{aligned}$$

But since $r_i^T u = 0$ for all $i$, the only way this equality can hold is if both sides are 0, so

$$\sum_{i=1}^{n} f_i r_i = 0. \tag{14}$$

Similarly, we can rewrite the original constraint for $\lambda$ as

$$\begin{aligned} 0 &= \sum_{i=1}^{n+2} \frac{(Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)} \\ &= \sum_{i=1}^{n+2} \frac{f_i u}{1 + \lambda^T (Z_i - \eta)} + \sum_{i=1}^{n+2} \frac{r_i}{1 + \lambda^T (Z_i - \eta)} \end{aligned}$$

so that, using (12),

$$\sum_{i=1}^{n+2} \frac{r_i}{1 + \lambda^T (Z_i - \eta)} = \sum_{i=1}^{n} \frac{r_i}{1 + \lambda^T (Z_i - \eta)} = 0.$$

Then using identity (10) twice, and the equality given by (13), we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} r_i - \sum_{i=1}^{n} \frac{r_i \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)} \\ &= \sum_{i=1}^{n} \frac{r_i \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)} \\ &= \sum_{i=1}^{n} \frac{f_i \theta^T r_i \theta^T u}{1 + \lambda^T (Z_i - \eta)} + \sum_{i=1}^{n} \frac{(\theta^T r_i)^2}{1 + \lambda^T (Z_i - \eta)} \\ &= \sum_{i=1}^{n} f_i \theta^T r_i \theta^T u - \sum_{i=1}^{n} \frac{f_i \theta^T r_i \theta^T u \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)} \\ &\qquad + \sum_{i=1}^{n} (\theta^T r_i)^2 - \sum_{i=1}^{n} \frac{(\theta^T r_i)^2 \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)}. \end{aligned}$$

The first term of the last equality is 0 by (14), and the second and fourth terms are both $o(s^{-1/2})$ by (11) because each includes a $\|\lambda\|$ factor and everything else in both terms is bounded. Thus we have

$$\sum_{i=1}^{n} (\theta^T r_i)^2 = o(s^{-1/2})$$

so $\theta^T r_i = o(s^{-1/4})$ for all $i$, and therefore

$$\theta^T u \to 1 \tag{15}$$

because $\theta$ is a unit vector, and we have shown that for any other vector $w$ such that $u^T w = 0$ we have $\theta^T w \to 0$. Then since $\lambda^T u = \|\lambda\| \theta^T u = o(s^{-1})$, and $\theta^T u \to 1$, we may conclude

$$\|\lambda\| = o(s^{-1}). \tag{16}$$

### A.4. Step 4

We once again use the fact that

$$\sum_{i=1}^{n+2} \frac{Z_i - \eta}{1 + \lambda^T (Z_i - \eta)} = 0$$

together with the identity

$$\frac{1}{1+x} = 1 - x + \frac{x^2}{1+x} \tag{17}$$

to give, using (16),

$$\begin{aligned} 0 &= \sum_{i=1}^{n+2} (Z_i - \eta) - \left( \sum_{i=1}^{n} (Z_i - \eta)\lambda^T (Z_i - \eta) + s^2 u\lambda^T u + (s+2r)^2 u\lambda^T u \right) \\ &\quad + \left( \sum_{i=1}^{n} \frac{(Z_i - \eta)\left[\lambda^T (Z_i - \eta)\right]^2}{1 + \lambda^T (Z_i - \eta)} - \frac{s^3 u(\lambda^T u)^2}{1 - s\lambda^T u} + \frac{(s+2r)^3 u(\lambda^T u)^2}{1 + (s+2r)\lambda^T u} \right) \\ &= (n+2)ru - \left( o(s^{-1}) + 2s^2 u\lambda^T u + o(1) \right) \\ &\quad + \left( o(s^{-2}) - s^3 u(\lambda^T u)^2 \left[ (n+2)(w_{n+2} - w_{n+1}) \right] + o(1) \right). \end{aligned}$$

In the last line, the term $s^3 u(\lambda^T u)^2 \left[ (n+2)(w_{n+2} - w_{n+1}) \right]$ is of order $o(s^3) \times o(s^{-2}) O(s^{-1}) = o(1)$, using (7) and (8). Thus we get $0 = (n+2)ru - 2s^2(\lambda^T u)u + o(1)$, giving

$$s^2 \lambda^T u \to \frac{(n+2)r}{2} \tag{18}$$

and since $\lambda^T u = \|\lambda\| \, \theta^T u$, by (15) we conclude

$$\|\lambda\| = O(s^{-2}). \tag{19}$$

### A.5. Step 5

Finally, we use the Taylor series expansion for $\log(1+x)$ about 1 to write

$$\begin{aligned} -\log\left((n+2)w_i\right) &= \log\left(1 + \lambda^T (Z_i - \eta)\right) \\ &= \lambda^T (Z_i - \eta) - \frac{1}{2}\left(\lambda^T (Z_i - \eta)\right)^2 + \frac{1}{3}\left(\lambda^T (Z_i - \eta)\right)^3 - d_i \end{aligned} \tag{20}$$

where $\|d_i\| = O(s^{-4})$ from (19) and the boundedness of the other terms in the expansion. Using the representation (20) in the expression

$$\widetilde{W}(\eta) = -2 \sum_{i=1}^{n+2} \log\left((n+2)w_i\right), \tag{21}$$

we have

$$\begin{aligned}\widetilde{W}(\eta) &= 2\left[\sum_{i=1}^{n+2}\lambda^T(Z_i-\eta) - \frac{1}{2}\sum_{i=1}^{n+2}\left(\lambda^T(Z_i-\eta)\right)^2 + \frac{1}{3}\sum_{i=1}^{n+2}\left(\lambda^T(Z_i-\eta)\right)^3 - \sum_{i=1}^{n+2} d_i\right] \\ &= 2\left[(n+2)r\lambda^T u - \frac{1}{2}\left(\sum_{i=1}^{n}\left(\lambda^T(Z_i-\eta)\right)^2 + s^2(\lambda^T u)^2 + (s+2r)^2(\lambda^T u)^2\right)\right. \\ &\quad \left. + \frac{1}{3}\left(\sum_{i=1}^{n}\left(\lambda^T(Z_i-\eta)\right)^3 - s^3(\lambda^T u)^3 + (s+2r)^3(\lambda^T u)^3\right) - O(s^{-4})\right].\end{aligned}$$

Multiplying both sides of this equality by $s^2$ and employing (19) gives

$$\begin{aligned}s^2\widetilde{W}(\eta) &= 2\left[(n+2)rs^2\lambda^T u - \frac{1}{2}\left(O(s^{-2}) + 2s^4(\lambda^T u)^2 + O(s^{-1})\right)\right. \\ &\quad \left. + \frac{1}{3}\left(O(s^{-4}) + O(s^{-2}) + O(s^{-3}) + O(s^{-4})\right) - O(s^{-2})\right] \\ &= 2\left[(n+2)rs^2\lambda^T u - s^4(\lambda^T u)^2 + O(s^{-1})\right].\end{aligned}$$

Substituting in the limiting expression (18) for $s^2\lambda^T u$, we have

$$s^2\widetilde{W}(\eta) \to 2\left[\frac{(n+2)^2r^2}{2} - \frac{(n+2)^2r^2}{4}\right]$$

which simplifies to

$$\frac{2ns^2}{(n+2)^2}\widetilde{W}(\eta) \to nr^2. \tag{22}$$

Then, since in this standardized setting Hotelling's T-square statistic is given by

$$T^2 = n\eta^T\eta = n(-ru)^T(-ru) = nr^2,$$

this completes the proof.

## References


Bartolucci, F. (2007). A penalized version of the empirical likelihood ratio for the population mean. *Statistics & Probability Letters* **77** 104–110. MR2339024

Chen, J., Variyath, A. M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics* **17** 426:443. MR2439967

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *The Annals of Statistics* **19** 1053–1061. MR1105861

Hotelling, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics* **2** 360–378.

Kiefer, J. and Schwartz, R. (1965). Admissible Bayes character of $T^2-$, $R^2-$, and other fully invariant tests for classical multivariate normal problems. *The Annals of Mathematical Statistics* **34** 747–770. MR0175245

Liu, Y. and Chen, J. (2009). Adjusted empirical likelihood with high-order precision. *Annals of Statistics* (to appear).

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18** 90–120. MR1041387

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall/CRC, New York.

Stein, C. (1956). The admissibility of Hotelling's $T^2$-test. *The Annals of Mathematical Statistics* **27** 616–623. MR0080413

Tsao, M. (2001). A small sample calibration method for the empirical likelihood ratio. *Statistics & Probability Letters* **54** 41–45. MR1857869

Tsao, M. (2004a). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *The Annals of Statistics* **32** 1215–1221. MR2065203

Tsao, M. (2004b). A new method of calibration for the empirical loglikelihood ratio. *Statistics & Probability Letters* **68** 305–314. MR2083899

Wendel, J. G. (1962). A problem in geometric probability. *Mathematica Scandinavica* **11** 109–111. MR0146858