SMALL SAMPLE PERFORMANCE AND CALIBRATION OF THE 
EMPIRICAL LIKELIHOOD METHOD

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS 
AND THE COMMITTEE ON GRADUATE STUDIES 
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS 
FOR THE DEGREE OF 
DOCTOR OF PHILOSOPHY

Sarah Emerson
September 2009
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

______________________________
(Art Owen) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

______________________________
(Joseph Romano)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

______________________________
(Guenther Walther)

Approved for the University Committee on Graduate Studies.
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 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 method in the vector mean setting have been proposed by a number of authors. We briefly explore a variety of such methods, commenting on the abilities of the various methods to address the calibration and convex hull challenges. In particular, a recent contribution suggests supplementing the observed dataset with an artificial data point, thereby eliminating the convex hull issue. We examine the performance 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 of the extra data approach, adding two balanced points rather than one, and changing the method used to determine the location of the artificial points. This modification demonstrates markedly improved calibration in
difficult examples, and also results in a small-sample connection between the modified empirical likelihood method and Hotelling’s T-square test. Varying the location of the added data points creates a continuum of tests, ranging from essentially the unmodified empirical likelihood method to a scaled Hotelling’s T-square test as the distance of the extra points from the sample mean increases. We investigate these consequences of the use and placement of artificial data points, exploring the accuracy of the chi-squared calibration and the power of the resulting test as the placement of the added points changes. Then we extend this method to multi-sample comparisons, where the new method demonstrates behavior comparable to that of multi-sample extensions of Hotelling’s T-square test with corrections, and has the added advantage that it may be tuned by varying the placement of the artificial points.
Acknowledgement

This thesis, as well as the remarkable and elegant body of work that precedes it, would not have been possible without the impressive and inspired efforts of Professor Art Owen. I have been extraordinarily privileged to have had Art as my advisor. The honor of the opportunity to work with the founder of empirical likelihood theory cannot be overstated, nor can I adequately express my gratitude for Art’s patient and indispensable guidance, both academic and general. Art devoted an incredible amount of time and attention to every detail that came up in our many discussions, invariably greatly expanding my perspective on the problem at hand. As a student in Art’s classes and working with him on this thesis, I have been inspired by Art’s work ethic and his dedication to teaching and communicating his understanding to others.

Professors Joseph Romano and Guenther Walther provided invaluable guidance and direction both as members of my reading committee and oral examination committee, and as instructors in classes and as mentors in general. I would like to thank Joe for his patient consideration of several tangential issues that I asked him about in the process of constructing this thesis, and for his careful and thoughtful review of this work. I am grateful to Guenther for contributing significant constructive comments at my oral examination and upon reading this thesis, and I would also like to
thank him for his valuable advice at several stages of my graduate career. Both Joe
and Guenther were exceptionally patient and generous with the tight time-frame at
the end of this process, and I am much obliged to them for their beneficence.

Professor Richard Olshen has taught me a great deal in our various interactions
over the past few years, and I am greatly indebted to him for his support and en-
couragement throughout my time as a student. I have been constantly impressed by
Richard’s ability to draw meaningful and illuminating connections between seemingly
distant topics. The many conversations that I had with Richard over the past two
years were key in shaping and directing both the progress of this work and my under-
standing of the field of statistics as a whole. I am also extremely grateful to Richard
for serving on my oral examination committee and for raising intriguing and probing
questions in the discussion that followed.

Professor Nancy Zhang was also very helpful and influential, serving on my oral
examination committee and contributing insightful questions, delving into areas of the
subject that I had not fully considered. Her questions and discussion were especially
useful in alerting me to areas of the work that warranted further investigation.

Finally, I wish to thank my wonderful family and friends for their incredible sup-
port and patience. My parents both served alternately as cheerleader, counselor,
colleague, editor, and refuge, and they were superb in every role. My sister Ruth
did everything from preparing me dinner during qualifying exams to submitting my
thesis for me, and a great deal more in between. My friends were available to talk
with, to help, and to distract me at all hours. My appreciation for their unwavering
encouragement and kindness is unbounded.
## Contents

Abstract iv  

Acknowledgement vi  

1 Introduction 1  

2 Basic Problem Setting and Hotelling’s T-square Test 4  
  2.1 Multivariate Hypothesis Testing 4  
  2.2 Hotelling’s T-square Statistic 6  

3 Empirical Likelihood Background 9  
  3.1 Empirical Likelihood Ratio Statistic 10  
  3.2 Asymptotic Results 12  
  3.3 Hotelling’s T-square Test and Euclidean Likelihood 15  
  3.4 Small-sample Challenges for Empirical Likelihood 16  

4 Calibration Methods for Empirical Likelihood Tests of a Multivariate Mean 22  
  4.1 F Calibration 23  
  4.2 Bootstrap Calibration 24
A Proof of Theorem 6.1.1

A.1 Step 1 ................................................................. 91
A.2 Step 2 ................................................................. 92
A.3 Step 3 ................................................................. 93
A.4 Step 4 ................................................................. 96
A.5 Step 5 ................................................................. 97

B Bootstrap Calibration Power Comparisons 99

C BEL Power Comparisons 108

C.1 Normal Data, \( d = 4, n = 10 \) ................................. 108
C.2 \( \chi^2 \) Data, \( d = 8, n = 40 \) ................................. 111
List of Tables

3.1 Probabilities of the true mean not belonging to the convex hull of a sample from asymmetric distributions ........................................ 18

4.1 Maximum confidence level for non-trivial AEL interval ............. 35

4.2 Comparison of small-sample calibration methods ..................... 36

5.1 Skewness and kurtosis of example distributions ..................... 49

7.1 Optimal values of scale factor $s^*(d, n)$ ................................. 70

7.2 Optimal values of scale factor $s^*(d)$ ................................ 71

8.1 Two-sample calibration comparisons: Scenario 1 .................... 84

8.2 Two-sample calibration comparisons: Scenario 2 .................... 85

8.3 Two-sample calibration comparisons: Scenario 3 .................... 85

8.4 Two-sample calibration comparisons: Scenario 1 with modified $s$ . . 85

8.5 Two-sample calibration comparisons: Scenario 2 with modified $s$ . . 85

8.6 Two-sample calibration comparisons: Scenario 3 with modified $s$ . . 86

B.1 Bootstrap calibration power comparisons: Normal data, $d = 2$, $n = 10$ 100

B.2 Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 2$, $n = 10$ . . 101

B.3 Bootstrap calibration power comparisons: Normal data, $d = 2$, $n = 40$ 102
B.4 Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 2$, $n = 40$ . . 103
B.5 Bootstrap calibration power comparisons: Normal data, $d = 4$, $n = 10$ 104
B.6 Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 4$, $n = 10$ . . 105
B.7 Bootstrap calibration power comparisons: Normal data, $d = 4$, $n = 40$ 106
B.8 Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 4$, $n = 40$ . . 107
List of Figures

3.1 Probabilities of the sample convex hull excluding the mean .................. 19
3.2 QQ and PP plots illustrating small-sample EL performance .................. 21

4.1 Illustration of quantities for defining extra point placement ............... 30
4.2 Placement of extra point for AEL method of Chen et al. (2008) .......... 31
4.3 QQ and PP plots illustrating small-sample AEL performance .............. 35

5.1 Placement of extra points for new calibration method ..................... 39
5.2 New calibration QQ plots for $d = 4$, $n = 10$ ................................. 51
5.3 New calibration PP plots for $d = 4$, $n = 10$ ................................. 52
5.4 New calibration QQ plots for $d = 4$, $n = 20$ ................................. 53
5.5 New calibration PP plots for $d = 4$, $n = 20$ ................................. 54
5.6 New calibration QQ plots for $d = 8$, $n = 20$ ................................. 55
5.7 New calibration PP plots for $d = 8$, $n = 20$ ................................. 56
5.8 New calibration QQ plots for $d = 8$, $n = 40$ ................................. 57
5.9 New calibration QQ plots for $d = 8$, $n = 40$ ................................. 58

6.1 Sample space ordering: Hotelling’s $T^2$ vs. BEL .......................... 65
6.2 Sample space ordering: EL vs. BEL ............................................. 66
Chapter 1

Introduction

Empirical likelihood methods, introduced by Owen (1988), provide nonparametric analogues 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 distribution, quantiles of a distribution, regression parameters, and linear contrasts in multi-sample problems.

In this thesis, 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 et al. (2008)) that in small samples or high dimensional problems, the asymptotic chi-squared calibration of the empirical likelihood ratio statistic produces a test that 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 of these adjustments to correct
the behavior of the empirical likelihood ratio test, and focus in particular on the 
method of Chen et al. (2008) which involves adding an artificial data point to the 
observed sample. This approach offers several key benefits in both ease of compu-
tation and accuracy. We explore the consequences of the recommended placement 
of the extra point, and we demonstrate a limitation of the method. We propose a 
modification of the data augmentation that provides improved calibration. Further-
more, we show that the new proposed method results in a continuum between the 
uncalibrated empirical likelihood method and Hotelling’s T-square test even in small 
samples. Simulation results demonstrate the effectiveness of the modified augmented 
empirical likelihood calibration.

We begin in Chapter 2 with a description of the basic setting for the inference 
problems that we will address throughout this thesis. We then introduce Hotelling’s 
T-square test, a parametric method that produces exact-level inference for Gaussian 
data and which boasts many other desirable qualities. Hotelling’s T-square test will be 
discussed frequently in the following chapters, as it has many interesting connections 
to the empirical likelihood approach.

In Chapter 3 we present the method of empirical likelihood, with particular focus 
on its application to tests for a vector mean parameter. The limiting asymptotic 
distribution of the resulting statistic, derived by Owen (1990), is discussed, and the 
relationship between empirical likelihood and other non-parametric likelihood-type 
methods is introduced. We investigate some of the challenges that empirical likelihood 
faces in small samples, where small here is considered relative to the dimension of 
the problem. This investigation indicates directions that may be taken to improve 
the small-sample performance of the empirical likelihood method.

Chapter 4 reviews a number of previously proposed methods of calibrating the
empirical likelihood method. We find that some of the commonly used methods have some significant drawbacks, and we develop these properties in some detail. For instance, the bootstrap calibration of empirical likelihood results in a noticeable loss of power in some settings, and thus the benefits that it offers for calibration are offset. Ideas from several of these calibration approaches are relevant to the method that we propose in Chapter 5, where we seek to remedy some of the weaknesses that we have identified. The calibration proposed in Chapter 5 is a modification of a calibration suggested by Chen et al. (2008), with several key differences which we discuss there. We show that our proposal satisfies the same asymptotic behavior as the original empirical likelihood statistic, and we illustrate the success of the calibration for our new method on a variety of simulated examples.

Chapter 6 develops a small sample connection between our calibrated empirical likelihood method and Hotelling’s T-square statistic, which occurs when a tuning parameter of the method is taken to the limit of $+\infty$. This connection has implications for the power of the resulting test, and is useful when we turn attention to multi-sample problems in Chapter 8. In Chapter 7 we discuss the choice of the tuning parameter, and describe simulations to obtain estimates of the optimal value of the parameter for different problem dimensions. Finally, in Chapter 8 we explore extensions of the calibration method to multi-sample problems, and we investigate the performance of the resulting statistic. We also comment on the development and use of Hotelling’s T-square statistic in this setting, and we show that the $t$-test for unequal variance and Hotelling’s T-square test for unequal variances can be derived as a sort of profile-likelihood test, and are closely related to the empirical likelihood ratio test as in the one-sample case.
Chapter 2

Basic Problem Setting and
Hotelling’s T-square Test

Here we introduce the basic setting of the questions that will be discussed throughout this thesis, and also describe the parametric approach given by Hotelling’s T-square test for multivariate normal data. In Chapter 8, we extend this question to the comparison of means from different samples, perhaps arising from different distributions.

2.1 Multivariate Hypothesis Testing

The subject of multivariate hypothesis testing has been studied by many authors over the past 80 years. Hotelling (1931) extended the single-sample univariate t-test to the multivariate setting, and robust versions of the resulting test have been developed by Utts and Hettmansperger (1980) and Tiku and Singh (1982). Roy (1958) developed a step-down procedure that may be used in place of Hotelling’s test, with different operating characteristics depending on the exact implementation of the step-down method. In the multi-sample case, a number of different methods have been proposed.
and studied. More detail on the multi-sample setting will be given in Chapter 8. The basic problem that we will begin with is that of testing hypotheses regarding the mean of a multivariate distribution $F_0$. This problem arises when multiple attributes are measured for each observation in a sample. Then each variable could be tested on its own, or the variables could be treated as a group. When treated individually, the overall type I error rate of the experiment will be increased, and must therefore be controlled using multiple hypothesis testing methods. Furthermore, the information contained in the dependence structure between different variables will not be used. In these cases, it is frequently desirable to simultaneously test hypotheses about all of the variables, taking into account the relationship of the variables to one another. The problem set-up is as follows: 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 $\mu_0 = E_{F_0}(X_i)$, that is, to test

$$H_0 : \mu_0 = \mu$$

(2.1)

where $\mu_0$ is the true mean vector of the distribution, and $\mu$ is a hypothesized value. Note that throughout this work we will be considering tests of $H_0$ versus alternatives of the form $H_A : \mu_0 \neq \mu$, the multivariate analogue of a univariate two-sided test. In the multivariate setting the parameter space is only partially ordered, and therefore it is unclear how one would extend the univariate notion of a one-sided test without collapsing the problem to a univariate linear combination of the multivariate hypothesis. While this is of course possible and potentially useful, it would obviate the use of the multivariate tests we consider here.

We now define the standardized setting that we will use throughout our discussion. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ denote the sample mean vector, and let $S_X$ denote the sample
covariance matrix:

$$S_X = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})^T.$$ 

We will assume throughout this work that $n > d$ and that $S_X$ is of full rank. Finally, let $A$ be an invertible matrix satisfying $AA^T = S$. Define the following standardized quantities:

$$Z_i = A^{-1} (X_i - \bar{X})$$

$$\eta = A^{-1} (\mu - \bar{X}).$$

The standardized random variables $Z_i$ clearly satisfy $\bar{Z} = 0$ and $S_Z = I$. We will use these standardized quantities to simplify notation in later sections. In general we will use the notation $S$, omitting the subscript, to refer to the matrix $S_X$.

### 2.2 Hotelling’s T-square Statistic

If we assume that $X_1, \ldots, X_n$ come from a multivariate Gaussian distribution $\mathcal{N}(\mu_0, \Sigma)$, the generalized likelihood ratio test of the hypothesis $H_0$ described in (2.1) is equivalent to (Seber, 1984) Hotelling’s T-square test (Hotelling, 1931), which rejects the null hypothesis for large values of the test statistic

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

This test is the multivariate analogue of the one-sample $t$-test, and it is easily seen that it reduces to the $t$-test in the univariate case where $d = 1$. In the case of Gaussian data, the statistic has a scaled $F_{d,n-d}$ distribution under the null hypothesis, 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
\[
\frac{n - d}{(n - 1)d} T^2(\mu) > F_{(d,n-d)}^{(1-\alpha)}.
\]
Seber (1984) also notes that Hotelling’s T-square test may be derived as a union-intersection test of the simultaneous univariate hypotheses \( H_{0\ell} : \ell^T \mu_0 = \ell^T \mu \) where \( \ell \) is a vector in \( \mathbb{R}^d \). In Chapter 3 we will develop another derivation of Hotelling’s T-square test based on maximizing a likelihood-type criterion similar to the approach of the empirical likelihood method.

Hotelling’s T-square test exhibits many desirable properties, including admissibility against several broad classes of alternatives (Stein, 1956; Kiefer and Schwartz, 1965), and invariance under the group of transformations defined by \( X \mapsto \tilde{X} = C X + v \), where \( C \) is a full-rank matrix of dimension \( d \times d \) (Lehmann and Romano, 2005). The hypothesis being tested using the transformed variables is then similarly transformed to \( H_0 : E(\tilde{X}_i) = \tilde{\mu} = C \mu + v \). In particular, in terms of standardized variables defined in the previous section, Hotelling’s T-square statistic simplifies to
\[
T^2(\mu) = n \eta^T \eta,
\]
a fact which will be employed in later chapters.

The multivariate central limit theorem (see, for instance, Anderson (2003, chap. 4)), along with a multivariate version of Slutsky’s theorem (see Timm (2002, chap. 2)), 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 or kurtotic distributions will of course require larger sample sizes to produce accurate inference using Hotelling’s T-square test. It has been shown that departures from symmetry, as measured by large skewness, have a greater effect on the validity of the univariate $t$-test than changes in kurtosis (Seber, 1984, chap. 3.3.3), and similar properties are surmised for the multivariate version given by Hotelling’s T-square test. Indeed, in our simulations in Chapter 5, we find that the accuracy of inference using Hotelling’s T-square is much more affected for highly skewed distributions than for highly kurtotic examples.

We have introduced Hotelling’s T-square test to serve as the standard to which we will compare other methods, and also because it will be shown to have some interesting connections to the empirical likelihood method, considered next, as well as to our adaptation of empirical likelihood in Chapter 5.
Chapter 3

Empirical Likelihood Background

The empirical likelihood method is constructed as a non-parametric extension of traditional likelihood-based approaches to hypothesis testing and confidence region construction. Empirical likelihood-type approaches were developed in work on survival analysis by Thomas and Grunkemeier (1975), and were formalized as a unified general method beginning with Owen (1988). The method provides a versatile approach that may be applied to perform inference for a wide variety of functionals of interest, and has been employed in a number of different areas of statistics. A brief examination of the current literature on empirical likelihood produces applications of empirical likelihood for problems including inference in missing data problems (Qin and Zhang, 2007); construction of confidence intervals for the area under ROC curves (Qin and Zhou, 2006); estimation of variogram model parameters (Nordman and Caragea, 2008); inference for GARCH models in time-series analysis (Chan and Ling, 2006); and many more varied topics. The flexibility of the empirical likelihood approach, as well as its relationship to many standard parametric procedures, make it a useful and interesting tool for many problems. Here we address the use
of empirical likelihood methods for testing hypotheses regarding vector-valued means of multivariate distributions, and later we will extend our results to multi-sample comparisons.

### 3.1 Empirical Likelihood Ratio Statistic

The empirical likelihood method (EL), as proposed by Owen (1988), seeks to extend the parametric likelihood ratio hypothesis testing approach to a non-parametric setting. Given a parametric model $X_i \sim F(x; \theta)$ for a parameter $\theta \in \Omega$, a generalized likelihood ratio test of the null hypothesis $H_0 : \theta \in \omega_0$ rejects the null hypothesis for small values of $\Lambda$, where $\Lambda$ is the generalized likelihood ratio statistic given by

$$\Lambda = \frac{\max_{\theta \in \omega_0} \left[ \text{lik}(\theta) \right]}{\max_{\theta \in \Omega} \left[ \text{lik}(\theta) \right]}.$$

Here the likelihood functions are determined by the form of the parametric model. The empirical likelihood method replaces the parametric model with the set of all distributions supported on the observed data. That is, we restrict attention to distributions $F \ll F_n$, where $F_n$ is the empirical cumulative distribution function which places mass $\frac{1}{n}$ on each observed point. All such distributions $F$ satisfying $F \ll F_n$ therefore place mass $w_i(F) \geq 0$ on the observed points $X_i$, and satisfy $\sum_{i=1}^{n} w_i = 1$. For such distributions, the likelihood of a particular distribution $F$ is given by

$$\text{lik}(F) = \prod_{i=1}^{n} w_i(F)$$

and clearly the likelihood is maximized over all distributions $F \ll F_n$ when $w_i(F) = \frac{1}{n}$, that is, by $F_n$. Then an empirical likelihood ratio for testing $H_0 : \theta(F) \in \omega_0$ is defined
similarly to the parametric generalized likelihood ratio as

\[ R = \frac{\max_{F \in F_n, \theta(F) \in \omega_0} \left[ \text{lik}(F) \right]}{\max_{F \in F_n} \left[ \text{lik}(F) \right]} \]

\[ = \frac{\max_{F \in F_n, \theta(F) \in \omega_0} \left( \prod_{i=1}^{n} w_i(F) \right)}{\left( \frac{1}{n} \right)^n} \]

\[ = \max_{F \in F_n, \theta(F) \in \omega_0} \prod_{i=1}^{n} n w_i(F). \]

For testing a hypothesis \( H_0 : \mu_0(F) = \mu \) where \( \mu_0(F) \) is the expectation with respect to the distribution \( F \), we therefore have the empirical likelihood ratio

\[ R(\mu) = \max_{F \in F_n, \mu(F) = \mu} \prod_{i=1}^{n} n w_i(F) \]

\[ = \max_{w_1, \ldots, w_n} \left\{ \prod_{i=1}^{n} n w_i \mid \sum_{i=1}^{n} w_i X_i = \mu, \ sum_{i=1}^{n} w_i = 1 \right\}. \]

That is, the empirical likelihood ratio \( R(\mu) \) maximizes \( \prod_{i=1}^{n} n w_i \) subject to the constraints

\[ \sum_{i=1}^{n} w_i = 1 \]  \hspace{1cm} (3.1)

and

\[ \sum_{i=1}^{n} w_i X_i = \mu. \]  \hspace{1cm} (3.2)

The log empirical likelihood ratio statistic is then given by

\[ W(\mu) = -2 \log R(\mu). \]

When positive weights \( w_i \) satisfying the constraints (3.1) and (3.2) do not exist, the usual convention is to set \( R(\mu) = -\infty \), and thus \( W(\mu) = \infty \). 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. In the next section we describe a result regarding the asymptotic behavior of the statistic $W(\mu_0)$ where $\mu_0$ is the true mean of the data-generating distribution. This asymptotic result can then be used to generate approximate finite-sample hypothesis tests and confidence regions.

### 3.2 Asymptotic Results

For a distribution $F$ with mean $\mu_0$ and finite covariance matrix $\Sigma$ of rank $q > 0$, Owen (1990) shows that

$$W(\mu_0) \xrightarrow{d} \chi^2_q \text{ as } n \to \infty.$$ 

We will assume throughout that the covariance matrix $\Sigma$ is of full rank $d$, though the result holds more generally. The proof of this asymptotic result proceeds by showing that $W(\mu_0)$ converges in probability to Hotelling’s T-square statistic $T^2(\mu_0)$ as $n \to \infty$. We outline the proof here, as similar arguments will be used in later chapters to show equivalent asymptotic behavior for calibrated statistics. The full detailed proof may be found in Owen (2001).

Suppose that $w_1, \ldots, w_n$ are the weights that maximize $\prod_{i=1}^n nw_i$ subject to constraints (3.1) and (3.2). Then the log empirical likelihood ratio statistic $W(\mu)$ is given by

$$W(\mu) = -2 \log R(\mu) = -2 \sum_{i=1}^n \log(nw_i).$$

To identify the maximizing weights $w_1, \ldots, w_n$, we construct the Lagrangian function
CHAPTER 3. EMPIRICAL LIKELIHOOD BACKGROUND

\[ G(w_1, \ldots, w_n) \] to incorporate the constraints (3.1) and (3.2):

\[ G(w_1, \ldots, w_n) = \sum_{i=1}^{n} \log(nw_i) - n\lambda^T \left( \sum_{i=1}^{n} w_i(X_i - \mu) \right) + \gamma \left( \sum_{i=1}^{n} w_i - 1 \right), \]

with \( \lambda \in \mathbb{R}^d \) and \( \gamma \in \mathbb{R} \) the Lagrange multipliers. Differentiating with respect to \( w_i \), we have

\[ \frac{d}{dw_i} G = \frac{1}{w_i} - n\lambda^T(X_i - \mu) + \gamma \] \hspace{1cm} (3.3)

and then setting the derivative equal to zero and summing we find that

\[ 0 = \sum_{i=1}^{n} w_i \frac{d}{dw_i} G \]

\[ = \sum_{i=1}^{n} \left( 1 - n\lambda^T w_i(X_i - \mu) + \gamma w_i \right) \]

\[ = n + \gamma \]

and thus \( \gamma = -n \). Incorporating this result into (3.3) and setting equal to zero again gives

\[ w_i = \frac{1}{n \left( 1 + \lambda^T(X_i - \mu) \right)} \] \hspace{1cm} (3.4)

where the Lagrange multiplier \( \lambda \) satisfies

\[ \sum_{i=1}^{n} w_i(X_i - \mu) = \sum_{i=1}^{n} \frac{X_i - \mu}{n \left( 1 + \lambda^T(X_i - \mu) \right)} = 0. \]

The proof of the asymptotic behavior of \( \mathcal{W}(\mu_0) \) now proceeds by showing that weights \( w_i \) satisfying (3.1) and (3.2) exist almost surely as \( n \to \infty \), and then establishing a
bound for \(|\lambda|\) of the form \(|\lambda| = O_p(n^{-1/2})\). Then it is shown that

\[
\lambda = S^{-1}(\bar{X} - \mu_0) + o_p(n^{-1/2})
\]

where

\[
S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)(X_i - \mu_0)^T.
\]  

(3.5)

Finally, this expression for \(\lambda\) is substituted into a Taylor expansion of \(\mathcal{W}(\mu_0) = -2 \sum_{i=1}^{n} \log(nw_i)\) to give

\[
\mathcal{W}(\mu_0) = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0) + o_p(1)
\]  

(3.6)

and since \(S^{-1} \overset{p}{\to} \Sigma\), the multivariate central limit theorem applied to \(\bar{X} - \mu_0\) combined with Slutsky’s theorem gives

\[
\mathcal{W}(\mu_0) \overset{d}{\to} \chi^2_d
\]

as \(n \to \infty\).

With further moment assumptions on the distribution \(F\), we may refine the rate of convergence, but this is the basic asymptotic result. Using this result, we can base hypothesis tests on the log empirical likelihood ratio statistic using the \(\chi^2_d\) distribution as the reference distribution. That is, the test that rejects the hypothesis \(H_0 : \mu_0 = \mu\) if

\[
\mathcal{W}(\mu) > \chi^2_d(1 - \alpha)
\]

will have significance level converging to \(\alpha\) as \(n \to \infty\). Here \(\chi^2_d(1 - \alpha)\) is the \(1 - \alpha\) quantile of the \(\chi^2_d\) distribution.
3.3 Hotelling’s T-square Test and Euclidean Likelihood

As mentioned in Chapter 2, Hotelling’s T-square test may be derived in a manner very similar to the development of the empirical likelihood statistic. Baggerly (1998) explores the empirical likelihood method as a member of a larger family of goodness-of-fit tests based on Cressie-Read power-divergence statistics of the form

$$CR(\lambda) = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^{n} \left( (nw_i)^{-\lambda} - 1 \right).$$

A statistic for testing the hypothesis $H_0 : \mu_0 = \mu$ may then be obtained by maximizing over the weights $w_i$ subject to the constraints $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i X_i = \mu$, as in the case of empirical likelihood. Thus a test statistic is given by

$$C(\mu; \lambda) = \max_{w_1, \ldots, w_n} \left\{ \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^{n} \left( (nw_i)^{-\lambda} - 1 \right) \right\} \left| \sum_{i=1}^{n} w_i X_i = \mu, \sum_{i=1}^{n} w_i = 1 \right\}.$$

Baggerly (1998) shows that for all values of the parameter $\lambda$, the resulting statistic has an asymptotic chi-squared distribution under the null hypothesis, and thus the asymptotic behavior of the log empirical likelihood ratio statistic is a special case of this more general result. The empirical likelihood ratio statistic $R(\mu)$ is equivalent to $C(\mu; \lambda = 0)$, where $C(\mu; \lambda = 0)$ is taken to be the limit of $C(\mu; \lambda)$ as $\lambda \to 0$. Replacing $\lambda = 0$ with $\lambda = -2$, we get what Owen (2001) refers to as the Euclidean
log likelihood statistic, given by
\[
\ell_E(\mu) = \frac{1}{2} C(\mu; \lambda = -2)
\]
\[
= \frac{1}{2} \max_{w_1, \ldots, w_n} \left\{ -\sum_{i=1}^{n} (nw_i - 1)^2 \left| \sum_{i=1}^{n} w_i X_i = \mu, \sum_{i=1}^{n} w_i = 1 \right. \right\}.
\]

Owen (2001, chap. 3.15) shows that the Euclidean log likelihood statistic recovers Hotelling’s T-square statistic, up to a scalar multiple. That is,
\[
-2\ell_E(\mu) = \left( \frac{n}{n-1} \right) T^2(\mu).
\]

This relationship will be employed when we address multi-sample hypothesis testing in Chapter 8.

### 3.4 Small-sample Challenges for Empirical Likelihood

The asymptotic result of Section 3.2 above allows us to test hypotheses regarding the mean and to construct confidence intervals using the appropriate critical values arising from the chi-squared 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 \( p(n, d) \) 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
\[
p(n, d) = 2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}.
\]
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 \( W \sim \text{Binom}(n - 1, \frac{1}{2}) \). (Note: an isomorphism between the binomial coin-flipping problem and this convex hull problem has not been identified.)

Though the result of Wendel (1962) applies only to distributions that are symmetric through the origin, this result may be extended to arbitrary continuous distributions as a lower bound. That is, if we let \( p(n, d, F) \) be the probability that the convex hull of a sample of size \( n \) from the \( d \)-dimensional distribution \( F \) does not contain the mean of \( F \), then Tsao (2004a) postulates that \( p(n, d, F) \geq p(n, d) \) for all continuous distributions \( F \). Tsao (2004a) provides a proof for the cases \( d = 1 \) and 2, and conjectures that the result holds for all finite \( d \). In fact, an earlier paper by Wagner and Welzl (2001) proves this conjecture for all \( d \) as a continuous extension of the Upper Bound Theorem for the number of faces of convex polytopes with a fixed number of vertices. Wagner and Welzl (2001) further show that the bound is only achieved for distributions that are symmetric through the origin; that is, for absolutely continuous distributions \( F \) with density function \( f \) satisfying \( f(x) = f(-x) \).

The results of simulations to determine \( p(n, d, F) \) for the case \( d = 4, n = 10 \) are presented in Table 3.1 for several different distributions \( F \). In these simulations, the distributions \( F \) are all multivariate with independent margins distributed according to the Normal(0, 1) distribution; the Exponential(\( \lambda = 3 \)) distribution; the \( F_{(4,10)} \) distribution; the \( \chi^2_{(1)} \) distribution; and the Gamma(1/4, 1/10) distribution. A total of 10,000 samples were drawn from each of these multivariate distributions, and the proportion of samples that did not contain the mean in the convex hull was recorded. As demonstrated by the results in Table 3.1, the probability \( p(n, d, F) \) may greatly exceed the lower bound \( p(n, d) \) when \( F \) is highly asymmetric.
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. It is interesting to investigate the probabilities $p(n, d)$ as $d$ increases, for $n$ a fixed multiple of the dimension $d$, say $n = \alpha d$. As $d$ increases, it is easy to see from the properties of the binomial distribution that the probability $p(n, d)$ will go to one if $\alpha$ is less than 2; will be exactly 0.5 if $\alpha = 2$; and will go to zero if $\alpha$ is greater than 2. These probabilities are illustrated in Figure 3.1. Thus for symmetric distributions with large dimension $d$, it will be sufficient to have a sample size just over twice the dimension to avoid convex hull problems. However, for practical sizes of $d$ or highly asymmetric distributions, the convex hull constraint can create a significant problem in the application of the empirical likelihood method, even for sample sizes greater than twice $d$.

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 bounded, as we now demonstrate, and therefore the asymptotic chi-squared distribution for this term will clearly be inappropriate in small samples. Analogous

<table>
<thead>
<tr>
<th>Distribution $F$</th>
<th>$p(n, d, F)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal(0, 1)</td>
<td>0.2512</td>
</tr>
<tr>
<td>Exponential(3)</td>
<td>0.3540</td>
</tr>
<tr>
<td>$F_{(4,10)}$</td>
<td>0.3952</td>
</tr>
<tr>
<td>$\chi^2_{(1)}$</td>
<td>0.4435</td>
</tr>
<tr>
<td>Gamma(0.25, 0.1)</td>
<td>0.5774</td>
</tr>
</tbody>
</table>

*Table 3.1: Probabilities $p(n, d, F)$ that the convex hull of a sample of size $n = 10$ from multivariate distributions on $d = 4$ dimensions with margins distributed independently according to $F$ does not contain the mean of the distribution.*
Figure 3.1: The probability $p(n, d)$ that the convex hull of a sample of size $n = \alpha d$ from a symmetric distribution on $d$ dimensions will not contain the mean, as a function of $d$. Each line in this illustration corresponds to a different multiple $\alpha$. We see that for $\alpha > 2$, the probability $p(n = \alpha d, d)$ goes to zero as $d$ increases, and for $\alpha < 2$, the probability $p(n = \alpha d, d)$ goes to one as $d$ increases.
to the definition of $S$, recall that $S(\mu)$ is given by

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

as in (3.5). In the asymptotic expansion of the statistic $W(\mu_0)$, the first order term is

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

as described in (3.6), whereas Hotelling’s T-square statistic is given by

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

Manipulating the expressions for $S(\mu)$ and $S$, we find that $\tilde{T}^2(\mu)$ is related to Hotelling’s T-square statistic by

$$\tilde{T}^2(\mu) = \frac{nT^2(\mu)}{T^2(\mu) + n - 1}$$

(Owen, 2001), and so

$$\tilde{T}^2(\mu) \leq n.$$

It is difficult to quantify the effect of the deviation of this term from its chi-squared limit. 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 anti-conservative test in small samples. This is illustrated in the quantile-quantile and probability-probability plots shown in Figure 3.2, which are generated by simulating 5000 data sets consisting of 10 points from the multivariate Gaussian distribution in four dimensions, and then calculating
Figure 3.2: 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 distribution. 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.

the value of the EL statistic for the true mean $\mu_0 = \bar{0}$ for each data set. 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 is a sizable discrepancy between the nominal and actual type I error rates. The calibration methods presented in the next two chapters seek to remedy the small-sample behavior of the empirical likelihood method, thereby producing a more accurate and practical test even in the case of small samples or high dimensional data.
Chapter 4

Calibration Methods for Empirical Likelihood Tests of a Multivariate 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. These approaches may be divided into two groups: those that address the convex hull constraint, and those that do not. Of the following methods, only the last two belong to the former group, and so we will focus extra attention on these two approaches. Table 4.2 at the end of this chapter summarizes the calibrations investigated here, and their respective abilities to address the limitations of the empirical likelihood method in small samples.
4.1 F Calibration

The simplest and one of the first proposed calibrations for the empirical likelihood method is to use an appropriately scaled F distribution (Owen, 2001) in place of the usual $\chi^2$ reference distribution calibration. Owen (1991) shows that under the null hypothesis the log empirical likelihood ratio statistic may be approximated as

$$W(\mu_0) = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0) + o_p(1)$$

where $S^{-1} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)(X_i - \mu_0)^T$ as in (3.5). As discussed in Section 3.4, the first order term of $W(\mu_0)$ is closely related to Hotelling’s T-square statistic,

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

which has a scaled $F_{(d,n-d)}$ null distribution for Gaussian data, given by

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

as seen in Section 2.2. Therefore, it seems reasonable to hope that the distribution of $W(\mu_0)$ might also be close to the same scaled $F_{(d,n-d)}$ distribution. Note that as $n \to \infty$, the appropriately scaled $F_{(d,n-d)}$ distribution converges to the $\chi^2_{(d)}$ distribution. However, in many examples there is little improvement in the resulting calibration, and the convex hull issue is clearly not addressed.
4.2 Bootstrap Calibration

Owen (1988) proposes using a bootstrap calibration, which involves resampling from the original data set to get new data sets \( \{X_1^{(b)}, \ldots, X_n^{(b)}\} \). Then for each bootstrap sample, the empirical likelihood ratio statistic \( W^{(b)}(\bar{X}) \) 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 \( W(\mu) \) is then compared to the distribution of values \( W^{(b)}(\bar{X}), b = 1, \ldots, B \) 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 bootstrap 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 good, but it is quite computationally intensive. As with most bootstrap processes, the performance is improved with a higher number of bootstrap repetitions.

One major drawback of the bootstrap calibration is a pronounced loss of power for the resulting test. We performed simulations investigating the performance of the bootstrap calibration for a range of sample sizes, dimensions, and underlying distributions, and we found that the size-adjusted power of the bootstrap calibrated method tended to be noticeably lower than the size-adjusted power of the uncalibrated empirical likelihood method or Hotelling’s T-square test. We use the term \textit{size-adjusted} to indicate the process of first choosing the appropriate critical value \( c^*(\alpha, U(\mu_0)) \) for each statistic \( U \) to give an exact level \( \alpha \) test, and then calculating the size-adjusted power as the probability of the statistic \( U(\mu) \) exceeding \( c^*(\alpha, U) \) at the alternative \( \mu \). For instance, when \( d = 2 \) and \( n = 10 \) for standard Gaussian data, the uncalibrated empirical likelihood method is far from the asymptotic \( \chi^2(1) \) distribution, so it would be unfair and difficult to interpret if we compared power based on
the critical value $\chi^2_{(1)}(1 - \alpha) = 5.99$ for $\alpha = 0.05$. Instead, we use the critical value $c^*(0.05, W(\mu_0)) \approx 17.97$, which is approximated based on simulations under the null hypothesis.

In the simulations presented here, we consider multivariate distributions with independent margins distributed according to the Normal or $\chi^2_{(1)}$ distributions, for dimensions $d = 2$ and $4$, with sample sizes $n = 10$ and $40$. For each combination of underlying distribution, dimension $d$, and sample size $n$, we generated 5000 samples. Then for each sample we computed $W(\mu)$, the bootstrap $p$-value $p_B(\mu)$, and Hotelling’s T-square statistic $T^2(\mu)$ at a range of values of $\mu$ including the true value $\mu_0$. We estimated the critical value $c^*(0.05, U(\mu_0))$ for each statistic, and used these estimated critical values to approximate the power. The estimate of the power was therefore the number of samples for which the statistic $U(\mu)$ exceeds the appropriate critical value. The results of these simulations are presented in Tables B.1 - B.8 in Appendix B. The decrease in power using the bootstrap calibration is larger for the $\chi^2_{(1)}$ data than for the normal data. For normal data and the larger sample size $n = 40$, there is very little difference between the power of the bootstrap calibrated empirical likelihood method and the power of the uncalibrated EL statistic and Hotelling’s T-square test. However, even with the larger sample size, the bootstrap calibration for $\chi^2_{(1)}$ data suffers up to a roughly 11% loss in power for the scenarios and alternatives considered here.

Hall and Titterington (1989) discuss the fact that Monte Carlo tests must suffer some loss of power due to the randomness inherent in the test, and Davidson and MacKinnon (2000) explore further the effect of the number of bootstrap samples on the power of the resulting bootstrap procedure. They show that for the setting of a pivotal statistic in a parametric bootstrap setting, the power of the bootstrap
procedure increases with the number of bootstrap samples $B$. It seems reasonable to expect that this same behavior might be observed in the bootstrap empirical likelihood setting, that is, that with more bootstrap samples, the power loss would be at least partially reversed. However, as noted earlier, performing bootstrap calibration for the empirical likelihood method is very computationally intensive, and thus increasing the value of $B$ may not be a desirable option.

4.3 Signed Root and Bartlett Correction

DiCiccio and Romano (1989) explore improving the performance of the empirical likelihood method using a mean adjustment to the signed root of the empirical likelihood ratio statistic. In the univariate setting, they show that this adjustment reduces the one-sided coverage error. For multivariate tests they develop an approximation that is asymptotically equivalent to the mean adjustment, but one-sided coverage is not a well-defined quantity for the multivariate problem, and the overall order of the two-sided coverage error is the same as for the unadjusted empirical likelihood. This adjustment does not address the convex hull constraint.

In a similar approach involving expansions of the empirical likelihood statistic, DiCiccio et al. (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. Using the Bartlett correction, a level $\alpha$ test is obtained by
rejecting the hypothesis $H_0 : \mu_0 = \mu$ when

$$W(\mu) > \left(1 - \frac{a}{n}\right)^{-1} \chi^2_{(d)}(1 - \alpha),$$

or equivalently, when

$$\left(1 - \frac{a}{n}\right) W(\mu) > \chi^2_{(d)}(1 - \alpha).$$

The same correction factor $a$ is used for all values of $\alpha$. Unless the underlying distribution is known, the ideal value of $a$ is unknown, but in many cases an estimate $\hat{a}$ may be obtained from the sample. For inference about a vector mean, the Bartlett correction factor $a$ is given by (Owen, 2001)

$$a = \frac{5}{3} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} (\mu_{jkl})^2 - 2 \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} \mu_{jkl} \mu_{kk} + \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} \mu_{jkk},$$

where $\mu_{jkl} = E(X^j X^k X^l)$ and $\mu_{jklm} = E(X^j X^k X^l X^m)$, letting $X^j$ denote the $j^{th}$ component of the random vector $X$. Therefore, the Bartlett correction factor may be estimated as

$$\hat{a} = \frac{5}{3} \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} (\hat{\mu}_{jkl})^2 - 2 \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} \hat{\mu}_{jkl} \hat{\mu}_{kk} + \frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} \hat{\mu}_{jkk},$$

estimating the values of $\mu_{jkl}$ and $\mu_{jklm}$ by

$$\hat{\mu}_{jkl} = \frac{1}{n} \sum_{i=1}^{n} X^j_i X^k_i X^l_i$$

and

$$\hat{\mu}_{jklm} = \frac{1}{n} \sum_{i=1}^{n} X^j_i X^k_i X^l_i X^m_i$$
respectively.

The Bartlett correction with an estimated correction factor performs well when the sample size is large enough or in small dimensions, but clearly scaling the reference distribution 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.

4.4 Exact Normal Calibration of Tsao

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. We suggest a similar idea for tuning the calibration of our proposed method in Chapter 7, as the distribution of the statistic under the normal distribution seems to do quite well at approximating the behavior for most symmetric distributions.

4.5 Penalized Empirical Likelihood Calibration

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 re-weighted sample distribution and the hypothesized mean \( \mu \). The penalized empirical likelihood ratio proposed by Bartolucci (2007) depends on a penalty parameter
CHAPTER 4. CALIBRATION METHODS

$h$, and is given by

\[
R^\dagger(\mu, h) = \max_{w_1, \ldots, w_n} \left\{ \left( \prod_{i=1}^n nw_i \right) e^{-n\delta(\nu-\mu)/2h^2} \left| \sum_{i=1}^n w_i x_i = \nu, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right. \right\}
\]

where $\delta(\nu-\mu)$ is a measure of the distance between the re-weighted mean $\nu$ and the hypothesized mean $\mu$, given by $\delta(\nu-\mu) = (\nu - \mu)^T S^{-1} (\nu - \mu)$. Here, as usual, $S$ is the sample covariance matrix. Then it is shown that

\[
W^\dagger(\mu, h) = -2 \log R^\dagger(\mu, h)
\]

has an asymptotic $\chi^2_d$ distribution as $n \to \infty$ for values of the penalty parameter $h$ that satisfy $h = O(n^{-1/2})$.

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.

4.6 Adjusted Empirical Likelihood

Finally, Chen et al. (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
Figure 4.1: Illustration of the quantities $v^*$, $r^*$, and $u^*$ used to define the placement of the extra point for the AEL method of Chen et al. (2008)

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 et al. 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.
Chen et al. (2008) propose adding an additional point to the sample as follows. Define the following quantities:

\[ v^* = \bar{X} - \mu, \]

\[ r^* = \|v^*\| = \|\bar{X} - \mu\|, \quad \text{and} \]

\[ u^* = \frac{v^*}{r^*} = \frac{\bar{X} - \mu}{\|\bar{X} - \mu\|}, \]

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 \( v^* \). These values are illustrated in Figure 4.1. In terms of these variables, for the hypothesis test described in Section 2.1, the
extra point $X_{n+1}$ that Chen et al. suggest is

$$X_{n+1} = \mu - a_n (\bar{X} - \mu)$$

$$= \mu - a_n v^*$$

$$= \mu - a_n r^* u^*,$$

where $a_n$ is a positive constant that may depend on the sample size $n$. Figure 4.2 demonstrates the placement of the extra point $X_{n+1}$. With this extra point, the resulting adjusted log empirical likelihood ratio statistic is

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

where

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

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 properties 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-squared calibration very inappropriate. The following Proposition describes this phenomenon.

**Proposition 4.6.1.** With an extra point placed as proposed in Chen et al. (2008) at $X_{n+1} = \mu - a_n (\bar{X} - \mu)$, the statistic $W^*(\mu) = -2 \log \mathcal{R}^*(\mu)$ is bounded above. The
CHAPTER 4. CALIBRATION METHODS

bound is given by

\[
W^*(\mu) \leq -2 \left[ n \log \left( \frac{(n+1)a_n}{n(a_n+1)} \right) + \log \left( \frac{n+1}{a_n+1} \right) \right].
\]

Proof. We show that weights \( \tilde{w}_i \) given by

\[
\tilde{w}_i = \frac{a_n}{n(a_n+1)} \quad \text{for } i = 1, \ldots, n
\]

\[
\tilde{w}_{n+1} = \frac{1}{a_n + 1}
\]

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

\[
\sum_{i=1}^{n+1} \tilde{w}_i X_i = \sum_{i=1}^{n} \tilde{w}_i X_i + \tilde{w}_{n+1} X_{n+1}
\]

\[
= \sum_{i=1}^{n} \frac{a_n}{n(a_n+1)} X_i + \frac{1}{a_n + 1} (\mu - a_n(X - \mu))
\]

\[
= \frac{a_n}{a_n + 1} X - \frac{a_n}{a_n + 1} X + \frac{1}{a_n + 1} \mu + \frac{a_n}{a_n + 1} \mu
\]

\[
= \mu.
\]

Then since clearly \( \sum_{i=1}^{n+1} \tilde{w}_i = 1 \), we therefore have

\[
\mathcal{R}^*(\mu) = \max \left\{ \prod_{i=1}^{n+1} (n+1)w_i \mid \sum_{i=1}^{n+1} w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^{n+1} w_i = 1 \right\}
\]

\[
\geq \prod_{i=1}^{n+1} (n+1)\tilde{w}_i.
\]
So taking logarithms and multiplying by $-2$ we find that:

$$W^*(\mu) \leq -2 \sum_{i=1}^{n+1} \log [(n + 1)\tilde{w}_i]$$

$$= -2n \log \left(\frac{(n + 1)a_n}{(a_n + 1)n}\right) - 2 \log \left(\frac{n + 1}{a_n + 1}\right).$$

This result clearly indicates the poor performance of the chi-squared 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 cause chi-squared calibrated $1 - \alpha$ confidence intervals to include the entire parameter space. Table 4.1 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 there are 10 observations in $d$ dimensions. For small values of $d$, the bound will not cause much of a problem, but even for $d = 3$, a 95% confidence region based on the $\chi^2_{(3)}$ reference distribution will be all of $\mathbb{R}^d$. Predictably, as $d$ increases this issue becomes more pronounced. Figure 4.3 illustrates the bound phenomenon for 10 points in 4 dimensions, and also demonstrates sub-optimal calibration even for values of $\alpha$ for which the boundedness of the statistic is not an issue.
### Table 4.1: Maximum possible confidence level for a non-trivial chi-squared calibrated confidence interval using the AEL method of Chen et al. (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 $B(n, a_n) = 7.334$.

<table>
<thead>
<tr>
<th>Dimension $d$</th>
<th>$P\left(\chi^2_{(d)} \leq B(n, a_n)\right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.993</td>
</tr>
<tr>
<td>2</td>
<td>0.974</td>
</tr>
<tr>
<td>3</td>
<td>0.938</td>
</tr>
<tr>
<td>4</td>
<td>0.881</td>
</tr>
<tr>
<td>5</td>
<td>0.803</td>
</tr>
<tr>
<td>6</td>
<td>0.709</td>
</tr>
<tr>
<td>7</td>
<td>0.605</td>
</tr>
<tr>
<td>8</td>
<td>0.499</td>
</tr>
<tr>
<td>9</td>
<td>0.398</td>
</tr>
</tbody>
</table>

**Figure 4.3:** 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 distribution. 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.
### Table 4.2: Comparisons of the small-sample properties of the calibration methods discussed in Chapter 4.

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.

<table>
<thead>
<tr>
<th>Calibration Method</th>
<th>Escape Convex Hull</th>
<th>Small-sample Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-calibration</td>
<td>No</td>
<td>Somewhat</td>
</tr>
<tr>
<td>Bootstrap calibration</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Bartlett correction</td>
<td>No</td>
<td>Somewhat</td>
</tr>
<tr>
<td>Tsao (2001) calibration</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Tsao (2004) calibration</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Bartolucci (2007) calibration</td>
<td>Yes</td>
<td>Somewhat</td>
</tr>
<tr>
<td>Chen, et al. (2008) calibration</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Chapter 5

New Calibration Approach: Balanced Augmented Empirical Likelihood

5.1 Modification of Extra Point Method

Inspired by the approach of the AEL method, we propose augmenting the sample with artificial data to address the challenges mentioned in Chapter 3. 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 et al., we propose adding two points to preserve the mean of the augmented data at $\bar{X}$. We also modify
the placement of the points. Recall the quantities

\[ v^* = \bar{X} - \mu, \]
\[ r^* = \|v^*\| = \|X - \mu\|, \quad \text{and} \]
\[ u^* = \frac{v^*}{r^*} = \frac{\bar{X} - \mu}{\|X - \mu\|} \]

from Section 4.6. In terms of these quantities, the placement of the new points is given by

\[ X_{n+1} = \mu - sc_{u^*}u^* \]
\[ X_{n+2} = 2\bar{X} - \mu + sc_{u^*}u^* \]

where \( c_{u^*} = \left(u^*T S^{-1} u^* \right)^{-1/2} \). 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. Figure 5.1 compares the extra point placement of the AEL method with the extra points placement described here.

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 later chapters. 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
Figure 5.1: Comparison of the placement of the extra points for the new calibration method BEL to the placement of the extra point for the AEL method of Chen et al. (2008).
the original dataset, and compute $\tilde{W}(\mu) = -2 \log(\tilde{R}(\mu))$ where

$$\tilde{R}(\mu) = \max \left\{ \prod_{i=1}^{n+2} (n + 2)w_i \left| \sum_{i=1}^{n+2} w_iX_i = \mu, w_i \geq 0, \sum_{i=1}^{n+2} w_i = 1 \right. \right\}.$$ 

We will refer to this statistic and method as the balanced empirical likelihood method (BEL) throughout the paper, to distinguish it from the unadjusted empirical likelihood statistic (EL) and the adjusted empirical likelihood statistic (AEL) of Chen et al. (2008). We show in the next section that the BEL statistic has the same asymptotic distribution as the EL statistic for any fixed value of $s$. Other desirable properties of the EL statistic are retained as well, as addressed in the following proposition.

**Proposition 5.1.1.** Placing the points according to (5.1) preserves the invariance property of the empirical likelihood method under transformations of the form $X \mapsto \tilde{X} = CX$, where $C$ is a full-rank matrix of dimension $d \times d$.

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

$$\tilde{u} = \frac{\tilde{X} - \tilde{\mu}}{\|\tilde{X} - \tilde{\mu}\|} = \frac{C (X - \mu)}{\|CX - C\mu\|},$$

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

$$\tilde{c}_a = \left( \tilde{u}^T (CSC^T)^{-1} \tilde{u} \right)^{-1/2} = \|C \tilde{X} - C\mu\| \left[ (\tilde{X} - \mu)^T S^{-1} (\tilde{X} - \mu) \right]^{-1/2}.$$
Thus we have

\[
\tilde{c}_u\tilde{u} = \|C\bar{X} - C\mu\| \left[ (\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) \right]^{-1/2} \frac{C (\bar{X} - \mu)}{\|C\bar{X} - C\mu\|}
\]
\[
= C \|\bar{X} - \mu\| \left[ (\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) \right]^{-1/2} \frac{\bar{X} - \mu}{\|\bar{X} - \mu\|}
\]
\[
= C \left[ u^T S^{-1} u^* \right]^{-1/2} u^*.
\]

Finally, when we place \(\bar{X}_{n+1}\) and \(\bar{X}_{n+2}\) based on the transformed data, we get

\[
\bar{X}_{n+1} = \bar{\mu} - s\tilde{c}_u\tilde{u} = C\mu - sC \left[ u^T S^{-1} u^* \right]^{-1/2} u^*
\]
\[
= CX_{n+1}
\]
\[
\bar{X}_{n+2} = 2\bar{\mu} + s\tilde{c}_u\tilde{u} = 2CX - C\mu + sC \left[ u^T S^{-1} u^* \right]^{-1/2} u^*
\]
\[
= CX_{n+2}.
\]

So 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 \(W^*(\mu)\) 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 Chapter 6.

In the next chapter we will address the effect 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. In Chapter 7, we explore in more detail the optimal value of $s$ to give the best approximation to a $\chi^2_{(d)}$ distribution.

5.2 Asymptotics of Balanced Augmented Empirical Likelihood

For a fixed value of the scale factor $s$, the BEL statistic $\tilde{W}(\mu_0)$ has the same asymptotic distribution as the uncalibrated log empirical likelihood ratio statistic $W(\mu_0)$. That is, as $n \to \infty$, for fixed $s$,

$$\tilde{W}(\mu_0) \xrightarrow{d} \chi^2_{(d)}.$$
In fact, we have the stronger result that as $n \to \infty$

$$\tilde{W}(\mu_0) - W(\mu_0) \overset{p}{\to} 0.$$  

That this should be the case is not surprising; as the sample size increases, two points placed at a distance $O(1)$ from the mean should have diminishing effects on the overall statistic. We now present the formal justification for this claim. First, we note that

$$X_{n+1} - \mu_0 = sc_{u^*} u^*$$

and

$$\begin{align*}
(X_{n+2} - \mu_0) &= 2(\bar{X} - \mu) - sc_{u^*} u^* \\
&= (2r^* - sc_{u^*})u^*.  
\end{align*}$$

Then by the weak law of large numbers, $r = \bar{X} - \mu \overset{p}{\to} 0$ as $n \to \infty$, and so $\|X_{n+1} - \mu_0\| = O_p(1)$ and $\|X_{n+2} - \mu_0\| = O_p(1)$. Furthermore, $c_{u^*} = (u^*TS^{-1}u^*)^{-1/2} = O_p(1)$ since for any unit vector $\theta$, we have $\gamma_1^{-1} + o_p(1) \leq \theta^T S^{-1} \theta \leq \gamma_d^{-1} + o_p(1)$ where $\gamma_1 \geq \gamma_d > 0$ are respectively the largest and smallest eigenvalues of $\Sigma = \text{Var}(X_i)$. These bounds result from the stochastic convergence of $S$ to $\Sigma$ as $n \to \infty$.

Now we explore the asymptotic behavior of the augmented sample covariance matrix estimate. Let $S$ be the estimate of the covariance matrix based on the original sample given by

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)(X_i - \mu_0)^T,$$
and let $\tilde{S}$ denote the same quantity computed for the augmented sample:

$$\tilde{S} = \frac{1}{n+2} \sum_{i=1}^{n+2} (X_i - \mu_0)(X_i - \mu_0)^T.$$ 

Then we have the following lemma:

**Lemma 5.2.1.** For fixed $s$, as $n \to \infty$,

$$\tilde{S} - S \xrightarrow{p} 0.$$

**Proof.** We expand the expression for $\tilde{S}$ as

$$\tilde{S} = \frac{1}{n+2} \left( \sum_{i=1}^{n} (X_i - \mu_0)(X_i - \mu_0)^T \right) + \frac{1}{n+2} ((X_{n+1} - \mu_0)(X_{n+1} - \mu_0)^T + (X_{n+2} - \mu_0)(X_{n+2} - \mu_0)^T)$$

$$= \frac{1}{n+2} \left( nS + s^2 c_u^2 u^* u^{*T} + (2r^* + sc_u^*)^2 u^* u^{*T} \right)$$

$$= \frac{n}{n+2} S + \frac{4r^*^2 + 4r^*sc_u^* + 2s^2 c_u^2}{n+2} u^* u^{*T}.$$ 

So we have

$$\tilde{S} - S = \frac{n - (n + 2)}{n+2} S + \frac{4r^*^2 + 4r^*sc_u^* + 2s^2 c_u^2}{n+2} u^* u^{*T}$$

$$= -\frac{2}{n+2} S + \frac{4r^*^2 + 4r^*sc_u^* + 2s^2 c_u^2}{n+2} u^* u^{*T}. \tag{5.2}$$

Then since $S$ converges in probability to $\Sigma$ as $n \to \infty$, the first term of (5.2) clearly converges in probability to the zero matrix. The second term of (5.2) is a matrix with all entries of order $O_p(1)O(n^{-1}) = O_p(n^{-1})$ since $s = O(1)$, $r^* = o_p(1)$, and
Using the facts presented above and the result of Lemma 5.2.1, we may follow the structure of the proof given in Owen (2001, chap. 11.2) to prove the following theorem:

**Theorem 5.2.1.** Let \( X_1, \ldots, X_n \) be independent observations distributed according to a common distribution function \( F \) with mean \( \mu_0 \) and finite covariance matrix \( \Sigma \). With extra points placed according to (5.1), the BEL statistic \( \tilde{W}(\mu_0) \) converges in distribution to \( \chi^2(d) \) as \( n \rightarrow \infty \). Furthermore, \( \tilde{W}(\mu_0) - W(\mu_0) \overset{p}{\rightarrow} 0 \).

**Proof.** We will use the result of Lemma 11.2 of Owen (2001) which states that for independent random variables \( Y_i \) distributed according to a common distribution function with \( \mathbb{E}(Y_i^2) < \infty \), \( \max_{i=1,\ldots,n} |Y_i| = o(n^{1/2}) \).

Recalling the construction of the weights \( w_i \) in terms of the Lagrange multiplier \( \lambda \) given by (3.4) in Section 3.2, for the augmented sample we have

\[
w_i = \left( \frac{1}{n + 2} \right) \frac{1}{1 + \tilde{\lambda}^T (X_i - \mu_0)},
\]

where the Lagrange multiplier \( \tilde{\lambda} \) satisfies

\[
\sum_{i=1}^{n+2} \frac{X_i - \mu_0}{1 + \tilde{\lambda}^T (X_i - \mu_0)} = 0. \tag{5.3}
\]

Write \( \hat{\lambda} = \|\tilde{\lambda}\|\theta \) where \( \theta \) is a \( d \)-dimensional unit vector. We begin by bounding the
magnitude $\|\hat{\lambda}\|$. Employing the identity $\frac{1}{1+x} = 1 - \frac{x}{1+x}$ we have

$$0 = \frac{1}{n+2} \theta^T \sum_{i=1}^{n+2} \frac{X_i - \mu_0}{1 + \hat{\lambda}^T (X_i - \mu_0)}$$

$$= \frac{1}{n+2} \sum_{i=1}^{n+2} \theta^T (X_i - \mu_0) - \frac{1}{n+2} \sum_{i=1}^{n+2} \frac{\theta^T (X_i - \mu_0) \hat{\lambda}^T (X_i - \mu_0)}{1 + \hat{\lambda}^T (X_i - \mu_0)}.$$ 

Using the fact that $\frac{1}{n+2} \sum_{i=1}^{n+2} X_i = \bar{X}$ and rearranging, we therefore find that

$$\theta^T (\bar{X} - \mu_0) = \|\hat{\lambda}\| \theta^T \tilde{S} \theta$$  \hspace{1cm} (5.4)

where

$$\tilde{S} = \frac{1}{n+2} \sum_{i=1}^{n+2} \frac{(X_i - \mu_0)(X_i - \mu_0)^T}{1 + \hat{\lambda}^T (X_i - \mu_0)}.$$ 

Then for $\tilde{S} = \frac{1}{n+2} \sum_{i=1}^{n+2} (X_i - \mu_0)(X_i - \mu_0)^T$ as above, and $Y_i = \hat{\lambda}^T (X_i - \mu_0)$ we have

$$\|\hat{\lambda}\| \theta^T \tilde{S} \theta \leq \|\hat{\lambda}\| \theta^T \tilde{S} \theta (1 + \max Y_i)$$

$$\leq \|\hat{\lambda}\| \theta^T \tilde{S} \theta (1 + \|\hat{\lambda}\| \max \|X_i - \mu_0\|)$$

$$= \theta^T (\bar{X} - \mu_0) (1 + \|\hat{\lambda}\| \max \|X_i - \mu_0\|),$$

where the last equality follows from equation (5.4). Rearranging this inequality gives

$$\|\hat{\lambda}\| \left( \theta^T \tilde{S} \theta - (\max \|X_i - \mu_0\|) \theta^T (\bar{X} - \mu_0) \right) \leq \theta^T (\bar{X} - \mu_0).$$

Now using the fact that $\tilde{S}$ converges in probability to $\Sigma$, we have $\gamma_1 + o_p(1) \geq \theta^T \tilde{S} \theta \geq \gamma_d + o_p(1)$, where $\gamma_1 \geq \gamma_d > 0$ are the largest and smallest eigenvalues, respectively, of $\Sigma$. Further, $\max_{1 \leq i \leq n} \|X_i - \mu_0\| = o_p(n^{1/2})$ as given by Lemma 11.2 of Owen (2001) mentioned above, combined with the fact that $\|X_{n+1} - \mu_0\| = O(1)$ and $\|X_{n+2} - \mu_0\| =
Finally, by the multivariate central limit theorem, \( \theta^T(\bar{X} - \mu_0) = O_p(n^{-1/2}) \), and so we find

\[
\|\tilde{\lambda}\| = O_p(n^{-1/2}),
\]

and therefore \( \max \tilde{\lambda}^T(X_i - \mu_0) = O_p(n^{-1/2})o(n^{1/2}) = o_p(1) \). We can rewrite (5.3) using the identity \( \frac{1}{1+x} = 1 - x + \frac{x^2}{1-x} \) as

\[
0 = \frac{1}{n + 2} \sum_{i=1}^{n+2} (X_i - \mu_0) \left( 1 - \tilde{\lambda}^T(X_i - \mu_0) + \frac{[\tilde{\lambda}^T(X_i - \mu_0)]^2}{1 + \tilde{\lambda}^T(X_i - \mu_0)} \right)
\]

\[
= \bar{X} - \mu_0 - \tilde{S}\lambda + \frac{1}{n + 2} \sum_{i=1}^{n+2} \frac{(X_i - \mu_0)[\tilde{\lambda}^T(X_i - \mu_0)]^2}{1 + \tilde{\lambda}^T(X_i - \mu_0)}.
\]

The last term above is bounded in norm by

\[
\frac{1}{n + 2} \sum_{i=1}^{n+2} \|X_i - \mu_0\|^3 \|	ilde{\lambda}\|^2 |1 + \tilde{\lambda}^T(X_i - \mu_0)|^{-1}
\]

\[
\leq (\max \|X_i - \mu_0\|) \frac{1}{n + 2} \sum_{i=1}^{n+2} \|X_i - \mu_0\|^2 \|	ilde{\lambda}\|^2 |1 + \tilde{\lambda}^T(X_i - \mu_0)|^{-1}
\]

\[
= o_p(n^{1/2})O_p(1)O_p(n^{-1})O_p(1)
\]

\[
= o_p(n^{1/2}),
\]

so \( \tilde{\lambda} = \tilde{S}^{-1}(\bar{X} - \mu_0) + o_p(n^{-1/2}) \). Finally, using the Taylor series expansion for \( \log(x) \)
about 1, we get

\[ \tilde{W}(\mu_0) = -2 \sum_{i=1}^{n+2} \log [(n + 2)w_i] \]

\[ = 2 \sum_{i=1}^{n+2} \log(1 + \tilde{\lambda}^T(X_i - \mu_0)) \]

\[ = 2 \sum_{i=1}^{n+2} \tilde{\lambda}^T(X_i - \mu_0) - \sum_{i=1}^{n+2} \left[ \tilde{\lambda}^T(X_i - \mu_0) \right]^2 + o_p(1) \]

\[ = 2(n + 2)\tilde{\lambda}^T(\bar{X} - \mu_0) - (n + 2)\tilde{\lambda}^T\tilde{S}\tilde{\lambda} + o_p(1) \]

\[ = (n + 2)(\bar{X} - \mu_0)^T\tilde{S}^{-1}(\bar{X} - \mu_0) + o_p(1). \]

Then since \( W(\mu_0) = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0) + o_p(1), \) \( \tilde{S} - S \xrightarrow{p} 0, \) and \( \frac{n}{n+2} \rightarrow 1 \) as \( n \rightarrow \infty, \) it follows that \( \tilde{W}(\mu_0) - W(\mu_0) \xrightarrow{p} 0. \)

\section{5.3 Improvement in Calibration}

To compare the calibration of EL, AEL, and BEL, we performed numerical comparisons based on simulated datasets for a variety of settings. We considered four combinations of sample size and dimension: \((d, n) = (4, 10), (4, 20), (8, 20),\) and \((8, 40).\) For each combination, we simulated data sets 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 5.1. We compared the chi-squared calibrations of the original empirical likelihood method, the adjusted empirical likelihood method of Chen et al., and the new adjusted method by creating quantile-quantile plots of the log empirical likelihood ratio statistics versus the appropriate chi-squared distribution. Figures 5.2 – 5.9 show the
resulting improvement in chi-squared calibration using our BEL method. We also plotted the p-values resulting from the chi-squared 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 et al.; and the green lines or points are the results of our BEL 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 Chapter 7.

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 achieved distinct improvement over the other two versions of empirical likelihood. We have also essentially matched the calibration performance of Hotelling’s T-square test. 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 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-squared distribution are quite non-linear.
Figure 5.2: 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 BEL 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.
CHAPTER 5. NEW CALIBRATION

Figure 5.3: Probability-Probability plots for \( d = 4, n = 10 \), for the same scenarios as illustrated in Figure 5.2. 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 BEL statistic (green). Hotelling’s T-square \( 1 - p \)-values are also included on this plot (blue).
Figure 5.4: Quantile-quantile plots for $d = 4, n = 20$. 
Figure 5.5: Probability-Probability plots for $d = 4, n = 20$. 
Figure 5.6: Quantile-quantile plots for $d = 8, n = 20$. 
Probability–Probability Plots

\( d = 8, n = 20 \)

- Normal
- \( t(3) \)
- Double Exponential
- Uniform
- Beta(0.1, 0.1)
- Exponential(3)
- \( F(4, 10) \)
- Chi-square(1)
- Gamma(1/4, 1/10)

Figure 5.7: Probability-Probability plots for \( d = 8, n = 20 \).
Figure 5.8: Quantile-quantile plots for $d = 8, n = 40$. 
Figure 5.9: Probability-Probability plots for $d = 8, n = 40$. 
Chapter 6

Small Sample Connection to Hotelling’s T-square Test

Here we develop a result connecting the BEL statistic to Hotelling’s T-square statistic in finite samples. This connection arises from the limiting behavior of $\tilde{W}(\mu)$ as the scale factor $s$ increases. For a fixed sample of size $n$, as the distance of the two extra points from $\mu$ increases, it is easy to see that the resulting statistic $\tilde{W}(\mu)$ tends to zero since weights very close to $w_i = \frac{1}{n+2}$ will satisfy $\sum_{i=1}^{n+2} w_i X_i = \mu$. However, the rate at which $\tilde{W}(\mu)$ converges to zero as $s \to \infty$ and the effect on the relative ordering of $\tilde{W}(\mu_1)$ and $\tilde{W}(\mu_2)$ for hypotheses $\mu_1$ and $\mu_2$ are questions of interest. Knowing the behavior of the statistic for different values of $s$ may help us choose an appropriate value, or calibrate the statistic for a range of different scale factors. The main result of this chapter deals with both of the questions mentioned above.
6.1 Limiting Behavior as $s$ Increases

To reduce notation, we will work with the standardized versions of the data and the hypothesized mean as described in Section 2.1, so

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

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

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

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

$$
Z_{n+1} = \eta - su \quad \text{and} \quad Z_{n+2} = -\eta + su. \quad (6.1)
$$

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

**Theorem 6.1.1.** For a fixed sample of size $n$

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

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

\[
\sum_{i=1}^{n+2} \frac{1}{(n+2)} \frac{Z_i - \eta}{1 + \lambda^T(Z_i - \eta)} = 0 \tag{6.2}
\]

so then the weights that maximize \( \tilde{R}(\eta) \) 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 (6.2), and employing 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 $\frac{2ns^2}{(n+2)r} \widetilde{W}(\mu) \to T^2$.

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 $x \to 1$.

These five steps are presented in full detail in Sections A.1 - A.5 of the Appendix A.

We mentioned in Section 3.2 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.

### 6.2 Sample Space Ordering

Next we explored the degree to which our new calibration deviates from the ordinary empirical likelihood method and agrees 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; that is, if one test is a monotone transformation of the other. Otherwise, if the tests produce different orderings of possible samples, they may make different decisions on the same data set even when both tests are perfectly calibrated. For instance, the $t$-test 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-squared 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 that 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 BEL statistic will become equivalent to Hotelling’s T-square statistic, but we would like to explore the extent to which this is true for small values of $s$. To do this, we generated 100 data sets, each consisting of 40 observations from a standard multivariate Gaussian distribution in 8 dimensions. For each data set, we computed Hotelling’s T-square statistic $T^2(\mu_0)$, the EL statistic $W(\mu_0)$, and the BEL statistic $\tilde{W}(\mu_0)$. 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 6.1 plots the ranking of the samples according to the BEL 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 6.2, except now the $x$-axis is the rank according to the EL statistic.

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

### 6.3 Power Implications

Given the relationship between Hotelling’s T-square test and BEL presented above, it follows that as $s$ increases the size-adjusted power of BEL at a fixed alternative $\mu$ will converge to the size-adjusted power of Hotelling’s T-square test. To explore the effect of $s$ on the power of the resulting BEL test, we perform simulations to compare the power of BEL for different values of $s$ to the power of EL, AEL, and Hotelling’s test. As we did for the Bootstrap calibration power comparisons in Section 4.2, we first calibrate each test to have exact level $\alpha$ in order to obtain meaningful comparisons.

The results of our power simulations are presented in Appendix C. Section C.1 displays the power for multivariate Gaussian data with dimension $d = 4$ and samples of size $n = 10$. We generated 1000 samples under the null hypothesis and under shift alternatives specified by the parameter $\mu$. For each sample, we calculated the EL, AEL, and Hotelling’s T-square statistics, as well as the BEL statistic using three different values of $s$: 1.0, 1.9, and 2.5. Then for each statistic we found the appropriate critical value under the null distribution to give an exact level $\alpha$ test for $\alpha = 0.1, 0.05, 0.02,$ and 0.01. To compare the power of the different statistics, we calculated under each alternative the proportion of samples for which a given test statistic exceeded its critical value. Section C.2 gives results for similar simulations performed for samples of size $n = 40$ generated with independent $\chi^2_{(1)}$ margins in $d = 8$ dimensions.
Figure 6.1: Comparing the ranking of 100 samples according to Hotelling’s $T$-square statistic (x-axis) vs. the BEL statistic (y-axis) as $s$ increases from 0.5 to 16.
Figure 6.2: Comparing the ranking of the 100 samples from Figure 6.1 according to the EL statistic (x-axis) vs. the BEL statistic (y-axis) as s increases from 0.5 to 16.
We see that within each of these two simulation settings these 6 tests have remarkably similar power performance. As expected, the $s$ parameter serves to tune the power curves from that of the EL statistic to that of Hotelling’s T-square statistic. In cases where EL has higher power, lower values of $s$ result in better power for BEL, and similarly when Hotelling’s T-square test has higher power, higher values of $s$ are better. Choosing a mid-range value of $s$ might therefore allow us to achieve a higher minimum power against all alternatives. This minimax-like behavior would need to be established more formally, but heuristically it would seem reasonable.

These size-adjusted power comparisons allow us to explore the performance of each statistic in terms of its ordering of the sample space. Such results should of course not be taken to represent the actual power of the test as it would be generally performed, where the null distribution of the statistics in finite samples are typically unavailable.
Chapter 7

Choice of Scale Factor $s$

In additional simulations we have explored the effect of the scale factor $s$ on the resulting chi-squared calibration of the BEL 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 1.7 for $d = 2$ to 2.5 for $d = 30$. We could of course derive or approximate the distribution of the test statistic $\tilde{W}$ for different values of $s$ to accommodate a range of sample space orderings. This could be done through simulation, using, for instance, multivariate Gaussian data as proposed by Tsao (2004b) for the original log empirical likelihood ratio statistic. However, given the ease of use offered by the chi-squared calibration, it seems worthwhile to try to identify appropriate scale factors to obtain the best chi-squared calibration.

7.1 Estimation of Optimal Scale Factor $s$

To estimate the value of the optimal value of the scale factor $s$ for a given dimension $d$, we performed simulations on multivariate Gaussian data for $d$ ranging from 1 to 20, and with sample sizes given by $n = \alpha d$ for $\alpha = 1.5, 2, 2.5, 3, 4, 5, 7.5, 10, 25$. For each
simulated data set, we computed the BEL statistic for values of $s$ ranging from 0.1 to 3.0 in steps of size 0.1. We also calculated Hotelling’s T-square statistic, the ordinary EL statistic, and the AEL statistic for comparison purposes. For each combination of $n$ and $d$ we selected the scale factor $s^*(d, n)$ that gave the best overall fit to the $\chi^2(d)$ distribution as measured by the Kolmogorov-Smirnov statistic. The resulting scale factors are displayed in Table 7.1. To combine the results across different values of $n$ for the same dimension $d$, we chose the value of $s$ that resulted in the smallest sum of the Kolmogorov-Smirnov discrepancies across the sample-sizes considered. These optimal values $s^*(d)$ are presented in Table 7.2. As $n$ increases, the value of $s$ has a smaller effect on the resulting statistic and a wider range of values will give similar calibration accuracy. For small values of $n$, a small change in $s$ can more significantly affect the quality of the chi-square calibration.
Table 7.1: The values $s^*(d, n)$ that minimize the Kolmogorov-Smirnov statistic comparing BEL to the reference $\chi^2_{(d)}$ distribution, for various combinations of $d$ and $n$, where $n = \alpha d$. 

<table>
<thead>
<tr>
<th>$d$</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7.5</th>
<th>10</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.0</td>
<td>2.1</td>
<td>1.8</td>
<td>1.7</td>
<td>1.7</td>
<td>1.6</td>
<td>1.7</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.0</td>
<td>1.8</td>
<td>1.8</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.6</td>
<td>2.0</td>
<td>1.9</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.3</td>
<td>2.0</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.5</td>
<td>2.0</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>2.0</td>
<td>1.9</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.4</td>
<td>2.1</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.5</td>
<td>2.1</td>
<td>2.1</td>
<td>2.0</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.4</td>
<td>2.2</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.5</td>
<td>2.2</td>
<td>2.1</td>
<td>2.1</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2.5</td>
<td>2.2</td>
<td>2.1</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2.5</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.5</td>
<td>2.3</td>
<td>2.2</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2.6</td>
<td>2.3</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td>2.4</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.6</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.6</td>
<td>2.4</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2.6</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2.7</td>
<td>2.4</td>
<td>2.3</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td>2.5</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2.6</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td>2.6</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2.7</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.5</td>
<td>2.6</td>
<td>2.6</td>
<td>2.3</td>
<td></td>
</tr>
</tbody>
</table>
Table 7.2: The values $s^*(d)$ that minimize the average Kolmogorov-Smirnov statistic comparing BEL to the reference $\chi^2(d)$ distribution across all sample sizes considered for a particular $d$.
Chapter 8

Extensions to Multi-sample Problems

We now extend the problem under consideration to one of comparing means across several different samples. This setting is the more common use of multivariate inference methods, as it arises whenever two or more groups are to be compared on the basis of more than one attribute. Examples of such comparisons can be found in many different applications and fields. For instance, Helbig (1991) compares direction vectors for bird migrations; Bartlett et al. (2000) compares measures of competence in eight different areas for graduates of two different nursing programs; and Tailby et al. (2005) explores the magnitude and orientation of neuronal dendrites in control and treatment groups of rats. In each of these studies, the response of interest is multivariate, and so a multivariate inference method is used to test the hypothesis that the group means differ.

The general set-up used throughout this chapter is as follows: we have $k$ samples of observations which we will designate as $X_{ij}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where
the index $i$ indicates the sample to which the observation belongs, so $X_{ij}$ is the $j^{th}$ observation from sample $i$. We will assume that within each sample $i$, the observations $X_{ij}$ are independent with common distribution function $F_i$. The standard goal that we will address for this setting is to test the hypothesis that the means $\mu_i$ of the distributions $F_i$ are equal. More complicated hypotheses regarding the relationship of the $\mu_i$ could of course be formed as well, but this basic hypothesis of equality forms the vast majority of all multi-sample hypothesis testing. We will therefore seek methods to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k.$$  

We begin by considering the case of $k = 2$ samples, which in the case of Gaussian data is referred to as the Behrens-Fisher problem. Some parametric approaches to this problem are described below. We use these parametric approaches as standards to which we compare the calibrated empirical likelihood methods developed later in this chapter.

### 8.1 Multivariate Behrens-Fisher Problem

The problem of testing equality of means for two Gaussian samples with perhaps unequal covariance matrices is generally referred to as the Behrens-Fisher problem. It has been addressed in the univariate and multivariate setting by many authors over the years. For the univariate case, Welch (1938) proposed an approximate solution based on comparing the $t$-test for unequal variances to a $t_{(r)}$ distribution where the degrees of freedom value $r$ is estimated from the sample. This approach has been extended to the multivariate case in different ways by James (1954) and Yao.
(1965). Both the James (1954) and Yao (1965) procedures compute what may be termed Hotelling’s two-sample T-square statistic for unequal variances, analogous to the univariate $t$-test for unequal variances. Scheffe (1943) suggests an exact alternate method for the univariate case of the Behrens-Fisher problem involving essentially forming random differences between the two groups and performing a one-sample test on the random differences, and Bennett (1950) extends this method to the multivariate case. In a later paper, Scheffe (1970) retracts this solution as it will give different results depending on the (presumed arbitrary) ordering of the observations in the two samples. Subrahmaniam and Subrahmaniam (1973) compare the methods of Bennett (1950), James (1954), and Yao (1965) on the basis of attained significance level and power, concluding that while the method of Bennett (1950) protects the significance level the best, it does so at the cost of a loss of power. The procedures of James (1954) and Yao (1965) are found to be roughly comparable in terms of achieved significance level and power. Dudewicz et al. (2007) explores exact solutions to the Behrens-Fisher problem that require being able to obtain a random additional amount of data in both samples. These solutions seem less practical for the settings in which such questions generally arise, so we will not consider them further.

### 8.2 Hotelling’s T-square Test for Unequal Variances

As mentioned in Section 8.1 above, a multivariate extension of the $t$-test for unequal variances for testing the hypothesis specified in (8.1) for two samples rejects the null
CHAPTER 8. MULTI-SAMPLE PROBLEMS

hypothesis for large values of

\[ T^2(\mu_1 = \mu_2) = (\bar{X}_1 - \bar{X}_2)^T \left( \frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\bar{X}_1 - \bar{X}_2) \]

where \( \bar{X}_i \) and \( S_i \) denote the sample mean and sample covariance matrix, respectively, of sample \( i \). This test is often referred to as Hotelling’s T-square test for unequal variances, and in the case \( d = 1 \) it clearly reduces to the \( t \)-test for unequal variances. Unlike the one-sample Hotelling’s T-square test, the test statistic \( T^2(\mu_1 = \mu_2) \) does not have an exact scaled \( F \) distribution under the null hypothesis even for Gaussian data. As mentioned in the previous section, approximations to the distribution of \( T^2(\mu_1 = \mu_2) \) have been proposed as in the univariate case.

In the next section we will describe the empirical likelihood approach to the general case of the hypothesis (8.1), and we will explore the analogous approach using the Euclidean log-likelihood function which was introduced in Chapter 3. As in the univariate case, the Euclidean log-likelihood method produces a statistic that is very closely related to Hotelling’s T-square statistic, and we demonstrate that therefore Hotelling’s T-square test for unequal variances can be derived by minimizing

\[ \tilde{T}^2(\mu) = T^2_1(\mu) + T^2_2(\mu) \]

over \( \mu \), where \( T^2_i(\mu) \) is the one-sample Hotelling’s T-square statistic for testing the hypothesis that the mean of the \( i^{th} \) sample is \( \mu \). That is, the \( t \)-test and Hotelling’s T-square test for unequal variances can be thought of heuristically as profile likelihood ratio statistics.
8.3 Empirical Likelihood for $k$ Samples

Owen (1991) develops the use of empirical likelihood for the hypothesis described in (8.1). The empirical likelihood ratio statistic for the more general hypothesis $H_0 : \mu_1 = \mu_1^0, \ldots, \mu_k = \mu_k^0$ is given by

$$
\mathcal{R}(\mu_1^0, \ldots, \mu_k^0) = \max_{w_{ij}} \left\{ \prod_{i=1}^{k} \prod_{j=1}^{n_i} n_i w_{ij} \mid w_{ij} \geq 0, \sum_{j=1}^{n_i} w_{ij} = 1, \sum_{j=1}^{n_i} w_{ij} X_{ij} = \mu_i^0 \text{ for } i = 1, \ldots, k \right\}.
$$

Thus the log empirical likelihood ratio statistic for this general hypothesis is

$$
W(\mu_1^0, \ldots, \mu_k^0) = \sum_{i=1}^{k} W_i(\mu_i^0) \quad (8.2)
$$

where $W_i(\mu_i^0) = -2 \log \mathcal{R}_i(\mu_i^0)$ is the log empirical likelihood ratio statistic for testing that the mean of sample $i$ is $\mu_i^0$. To test the null hypothesis (8.1) of equality among the means $\mu_1, \ldots, \mu_k$, the profile log empirical likelihood ratio statistic given by

$$
W(H_0) = \min_{\mu} W(\mu, \ldots, \mu) \quad (8.3)
$$

is used, and Owen (2001) gives a heuristic argument justifying a limiting $\chi^2_d (k-1)$ distribution for the statistic $W(H_0)$ as $\min_i n_i \to \infty$. The usual moment assumptions must also hold.

As described in Chapter 3, in a manner analogous to the empirical log-likelihood function, the Euclidean log-likelihood function may be used to measure the distance between an observed sample $X_1, \ldots, X_n$ and a hypothesized mean $\mu$ of the underlying
distribution. Recall that the Euclidean log-likelihood function is given by

$$\ell_E = \sum_{i=1}^{n} (nw_i - 1)^2$$

for weights $w_1, \ldots, w_n$ satisfying

(i) $\sum_{i=1}^{n} w_i = 1$

(ii) $\sum_{i=1}^{n} w_i X_i = \mu$.

A test statistic is obtained by maximizing $\ell_E$ and taking

$$U(\mu) = -2 \max_{\{w_1, \ldots, w_n\}} \left\{ \sum_{i=1}^{n} (nw_i - 1)^2 \left| \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i X_i = \mu \right. \right\}.$$ 

Let $\mathbf{S}$ be the sample covariance matrix, as usual, given by

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

and let $\tilde{\mathbf{S}}$ similarly be defined as

$$\tilde{\mathbf{S}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})^T = \frac{n-1}{n} \mathbf{S}.$$
Then as mentioned in Chapter 3, it is easily shown (Owen, 2001) that

\[ U(\mu) = n (\bar{X} - \mu)^T \tilde{S}^{-1} (\bar{X} - \mu) \]
\[ = n (\bar{X} - \mu)^T \left( \frac{n - 1}{n - \bar{S}} \right)^{-1} (\bar{X} - \mu) \]
\[ = \left( \frac{n}{n - 1} \right) \frac{n}{n} (\bar{X} - \mu)^T \tilde{S}^{-1} (\bar{X} - \mu) \]
\[ = \left( \frac{n}{n - 1} \right) T^2(\mu) \]

where \( T^2(\mu) \) is Hotelling’s T-square statistic. This approach of maximizing the Euclidean log-likelihood may be extended to the multi-sample setting, where the data now come from \( k \) different samples, with observations \( X_{ij}, j = 1, \ldots, n_i \) from sample \( i \), for \( i = 1, \ldots, k \). To test the hypothesis \( H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \), we similarly define a test statistic

\[ U_{H_0} = -2 \max_{\{w_{ij}\}, \mu} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (n_i w_{ij} - 1)^2 \left| \sum_{j=1}^{n_i} w_{ij} X_{ij} = \mu \right. \right. \text{for } i = 1, \ldots, k \} \]

This may be equivalently expressed as

\[ U_{H_0} = \min_{\mu} \left\{ U_1(\mu) + U_2(\mu) + \ldots + U_k(\mu) \right\} \]
\[ = \min_{\mu} \left\{ \sum_{i=1}^{k} U_i(\mu) \right\} \]

where \( U_i(\mu) \) is the statistic for the test of the hypothesis \( H_{0}^{(i)} : \mu_i = \mu \) based on the observations from the \( i^{th} \) sample. That is,

\[ U_i = n_i (\bar{X}_i - \mu)^T \tilde{S}_i^{-1} (\bar{X}_i - \mu) \]
with $\bar{X}_i$ and $\bar{S}_i$ the sample mean vector and covariance matrix, respectively, for sample $i$. Define $G(\mu) = U_1(\mu) + U_2(\mu) + \ldots + U_k(\mu)$, and let $\mu^*$ be the value of $\mu$ that minimizes $G(\mu)$, so $U_{H_0} = G(\mu^*)$. Then $\mu^*$ is found to satisfy

$$\left. \frac{d}{d\mu} G(\mu) \right|_{\mu^*} = 0$$

and since

$$\frac{d}{d\mu} U_i(\mu) = n_i\bar{S}_i^{-1}(\bar{X}_i - \mu)$$

we have that $\mu^*$ satisfies

$$\sum_{i=1}^{k} n_i\bar{S}_i^{-1}(\bar{X}_i - \mu^*) = 0.$$

Thus we find that $\mu^*$ is given by

$$\mu^* = \left( \sum_{i=1}^{k} n_i\bar{S}_i^{-1} \right)^{-1} \sum_{i=1}^{k} n_i\bar{S}_i^{-1}\bar{X}_i. \quad (8.4)$$

Letting $\tilde{V}_i = \frac{1}{n_i}\bar{S}_i$ and $A = \sum_{i=1}^{k} \tilde{V}_i^{-1}$, this becomes

$$\mu^* = A^{-1} \sum_{i=1}^{k} \tilde{V}_i^{-1}\bar{X}_i,$$
and therefore we get

\[ U_{H_0} = \sum_{i=1}^{k} U_i(\mu^*) \]

\[ = \sum_{i=1}^{k} n_i (\bar{X}_i - \mu^*)^T \tilde{S}_i^{-1} (\bar{X}_i - \mu^*) \]

\[ = \sum_{i=1}^{k} \left( \bar{X}_i^T \tilde{V}_i^{-1} \bar{X}_i - 2 \bar{X}_i^T \tilde{V}_i^{-1} \mu^* + \mu^* \bar{X}_i^T \tilde{V}_i^{-1} \mu^* \right) \]

\[ = \sum_{i=1}^{k} \left( \bar{X}_i^T \tilde{V}_i^{-1} \bar{X}_i - 2 \left( \sum_{i=1}^{k} \bar{X}_i^T \tilde{V}_i^{-1} \right) A^{-1} \left( \sum_{i=1}^{k} \tilde{V}_i^{-1} \bar{X}_i \right) \right. \]

\[ + \left. \left( \sum_{i=1}^{k} \tilde{V}_i^{-1} \bar{X}_i \right)^T A^{-1} A A^{-1} \left( \sum_{i=1}^{k} \tilde{V}_i^{-1} \bar{X}_i \right) \right) \]

\[ = \sum_{i=1}^{k} \bar{X}_i^T \tilde{V}_i^{-1} \bar{X}_i - \left( \sum_{i=1}^{k} \tilde{V}_i^{-1} \bar{X}_i \right)^T A^{-1} \left( \sum_{i=1}^{k} \tilde{V}_i^{-1} \bar{X}_i \right) . \]

This is the statistic developed in James (1954) for tests of the hypothesis (8.1), except that the statistic proposed there uses the unbiased estimates \( S \) in place of \( \tilde{S} \). In the two-sample case, the resulting statistic may be alternately expressed as

\[ U_{H_0} = \bar{X}_1^T \tilde{V}_1^{-1} \bar{X}_1 + \bar{X}_2^T \tilde{V}_2^{-1} \bar{X}_2 \]

\[ - \bar{X}_1^T \tilde{V}_1^{-1} A^{-1} \left( \tilde{V}_1^{-1} \bar{X}_1 + \tilde{V}_2^{-1} \bar{X}_2 \right) \]

\[ - \bar{X}_2^T \tilde{V}_2^{-1} A^{-1} \left( \tilde{V}_1^{-1} \bar{X}_1 + \tilde{V}_2^{-1} \bar{X}_2 \right) \]

\[ = \bar{X}_1^T \tilde{V}_1^{-1} \left[ \left( I - A^{-1} \tilde{V}_1^{-1} \right) \bar{X}_1 - A^{-1} \tilde{V}_2^{-1} \bar{X}_2 \right] \]

\[ + \bar{X}_2^T \tilde{V}_2^{-1} \left[ \left( I - A^{-1} \tilde{V}_2^{-1} \right) \bar{X}_2 - A^{-1} \tilde{V}_1^{-1} \bar{X}_1 \right] . \]
This can be simplified using the fact that

\[ A = \tilde{V}_1^{-1} + \tilde{V}_2^{-1} \]

and so \( A^{-1}\tilde{V}_1^{-1} = I - A^{-1}\tilde{V}_2^{-1} \), and similarly \( A^{-1}\tilde{V}_2^{-1} = I - A^{-1}\tilde{V}_1^{-1} \), giving

\[ U_{H_0} = \bar{X}_1^T\tilde{V}_1^{-1} \left[ A A^{-1}\tilde{V}_1^{-1}\bar{X}_1 - A^{-1}\tilde{V}_2^{-1}\bar{X}_1 \right] \]
\[ + \bar{X}_2^T\tilde{V}_2^{-1} \left[ A^{-1}\tilde{V}_1^{-1}\bar{X}_2 - A^{-1}\tilde{V}_1^{-1}\bar{X}_2 \right] \]
\[ = \bar{X}_1^T\tilde{V}_1^{-1} A^{-1}\tilde{V}_2^{-1} (\bar{X}_1 - \bar{X}_2) - \bar{X}_2^T\tilde{V}_2^{-1} A^{-1}\tilde{V}_1^{-1} (\bar{X}_1 - \bar{X}_2) . \]

Finally, we note that

\[ \tilde{V}_1^{-1} A^{-1}\tilde{V}_2^{-1} = \left( \tilde{V}_2 A \tilde{V}_1 \right)^{-1} \]
\[ = \left[ \tilde{V}_2 \left( \tilde{V}_1^{-1} + \tilde{V}_2^{-1} \right) \tilde{V}_1 \right]^{-1} \]
\[ = \left( \tilde{V}_2 + \tilde{V}_1 \right)^{-1} \]

and by similar reasoning \( \tilde{V}_2^{-1} A^{-1}\tilde{V}_1^{-1} = \left( \tilde{V}_1 + \tilde{V}_2 \right)^{-1} \). Therefore, we have that

\[ U_{H_0} = (\bar{X}_1 - \bar{X}_2)^T \left( \tilde{V}_1 + \tilde{V}_2 \right)^{-1} (\bar{X}_1 - \bar{X}_2) . \]

Replacing \( \tilde{V}_i \) with \( V_i = \frac{1}{n_i} S_i \) in the above arguments would produce the two-sample Hotelling’s T-square statistic for unequal variances. This means that, as mentioned in Section 8.2 above, Hotelling’s T-square test for unequal variances arises from minimizing

\[ \tilde{T}^2(\mu) = T_1^2(\mu) + T_2^2(\mu) \]
over $\mu$.

### 8.4 Balanced Empirical Likelihood for Multi-sample Problems

To apply the balanced empirical likelihood calibration to the multi-sample problem, we make use of (8.2) and (8.3) to define the statistic

$$\tilde{W}(H_0) = \min_{\mu} \sum_{i=1}^{k} \tilde{W}_i(\mu)$$

(8.5)

where $\tilde{W}_i(\mu) = -2\log R_i(\mu)$ is the BEL statistic for testing that the mean of sample $i$ is $\mu$. Thus extra points are added to each sample as in (5.1), dependent upon the value of $\mu$ being tested. The optimizing required in (8.5) clearly involves a search over $\mu$, which could be quite time-consuming, but we circumvent that obstacle by using the optimal value $\mu^*$ defined above by (8.4) for the Euclidean log likelihood. Asymptotically, this modified approach will be the same as if we had actually searched for the optimal $\mu$, and it drastically reduces the computation time required to use empirical likelihood in this setting.

We have only explored using the same value of the scale factor $s$ in adding the extra points to each sample, but there is no requirement that this restriction be maintained in general. The asymptotic behavior as $\min_i n_i \to \infty$ will be the same regardless of the fixed scale factors $s_1, \ldots, s_k$ used to calculate $\tilde{W}_1(\mu), \ldots, \tilde{W}_k(\mu)$ respectively. We have considered several small examples using this calibrated empirical likelihood method, and have found in general that it performs comparably to the parametric approaches of James (1954) and Yao (1965). The results of our preliminary investigations are
detailed in the next section.

8.5 Multi-sample Calibration Results

To assess the accuracy of the BEL calibration for two-sample tests, we performed three sets of simulations. The scenarios for the three sets of simulations are described below.

1. Both groups multivariate Gaussian with mean zero, with \( d = 4 \), \( n_1 = 10 \), \( n_2 = 15 \). The covariance matrices of sample 1 and sample 2 are given by \( \Sigma_1 \) and \( \Sigma_2 \):

\[
\Sigma_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

and

\[
\Sigma_2 = \begin{bmatrix}
1 & 1.8 & 2.4 & 2.8 \\
1.8 & 4 & 5.4 & 6.4 \\
2.4 & 5.4 & 9 & 10.8 \\
2.8 & 6.4 & 10.8 & 16
\end{bmatrix}
\]

2. Both groups multivariate Gaussian with mean zero, with \( d = 4 \), \( n_1 = 15 \), \( n_2 = 10 \). The covariance matrices of sample 1 and sample 2 are the same as in scenario 1.

3. Group 1 has independent \( \chi^2_{(1)} - 1 \) margins, and Group 2 has independent \( N(0, 1) \) margins, with \( d = 8 \), \( n_1 = 16 \), \( n_2 = 32 \). Thus both distributions have mean equal to the origin.
For each scenario, we generated 5000 data sets. We calculated $p$-values according to the method of James (1954), according to the method of Yao (1965), and using the BEL statistic with a value of $s$ equal to the optimal values presented in Table 7.2 for the single-sample case. Then we assessed how many of the resulting $p$-values were less than the nominal significance level $\alpha$, for $\alpha = 0.1$, 0.05, 0.02, and 0.01. The values obtained are presented in Tables 8.1 – 8.3. The BEL statistic produces a more conservative test than either of the other two, and so we also try reducing the value of $s$ slightly for each scenario, thereby producing a test more similar to EL. The results with $s = s^*(d) - 0.2$ are given in Tables 8.4 – 8.6. We find that with the reduced $s$, the BEL method very closely approximates the degrees of freedom approximation proposal of Yao (1965). Thus BEL appears to offer a potentially quite useful alternative to current approaches to the multi-sample mean comparison problem. The ability to tune $s$ in this problem is an attractive feature, as it allows adjustment of the resulting calibration. However, in order to use the method it will be necessary to set a value of $s$, so more work on evaluating the best choice of the scale factor $s$ in the multi-sample case is certainly warranted.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.117</td>
<td>0.093</td>
<td>0.075</td>
</tr>
<tr>
<td>0.05</td>
<td>0.065</td>
<td>0.047</td>
<td>0.035</td>
</tr>
<tr>
<td>0.02</td>
<td>0.031</td>
<td>0.019</td>
<td>0.011</td>
</tr>
<tr>
<td>0.01</td>
<td>0.015</td>
<td>0.008</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table 8.1: Scenario 1: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with $s = s^*(d)$. 
Table 8.2: Scenario 2: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with \( s = s^*(d) \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.113</td>
<td>0.092</td>
<td>0.074</td>
</tr>
<tr>
<td>0.05</td>
<td>0.062</td>
<td>0.047</td>
<td>0.033</td>
</tr>
<tr>
<td>0.02</td>
<td>0.027</td>
<td>0.020</td>
<td>0.014</td>
</tr>
<tr>
<td>0.01</td>
<td>0.019</td>
<td>0.010</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 8.3: Scenario 3: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with \( s = s^*(d) \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.273</td>
<td>0.271</td>
<td>0.207</td>
</tr>
<tr>
<td>0.05</td>
<td>0.185</td>
<td>0.182</td>
<td>0.126</td>
</tr>
<tr>
<td>0.02</td>
<td>0.114</td>
<td>0.109</td>
<td>0.061</td>
</tr>
<tr>
<td>0.01</td>
<td>0.075</td>
<td>0.074</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 8.4: Scenario 1: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with \( s = s^*(d) - 0.2 \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.129</td>
<td>0.104</td>
<td>0.118</td>
</tr>
<tr>
<td>0.05</td>
<td>0.072</td>
<td>0.057</td>
<td>0.062</td>
</tr>
<tr>
<td>0.02</td>
<td>0.035</td>
<td>0.024</td>
<td>0.027</td>
</tr>
<tr>
<td>0.01</td>
<td>0.018</td>
<td>0.010</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table 8.5: Scenario 2: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with \( s = s^*(d) - 0.2 \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.130</td>
<td>0.106</td>
<td>0.119</td>
</tr>
<tr>
<td>0.05</td>
<td>0.070</td>
<td>0.052</td>
<td>0.058</td>
</tr>
<tr>
<td>0.02</td>
<td>0.033</td>
<td>0.021</td>
<td>0.023</td>
</tr>
<tr>
<td>0.01</td>
<td>0.018</td>
<td>0.013</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Table 8.6: Scenario 3: Comparison of achieved significance level for the two-sample testing methods of James (1954) and Yao (1965) and the BEL method with $s = s^*(d) - 0.2$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>James</th>
<th>Yao</th>
<th>BEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.273</td>
<td>0.272</td>
<td>0.260</td>
</tr>
<tr>
<td>0.05</td>
<td>0.188</td>
<td>0.188</td>
<td>0.171</td>
</tr>
<tr>
<td>0.02</td>
<td>0.121</td>
<td>0.118</td>
<td>0.100</td>
</tr>
<tr>
<td>0.01</td>
<td>0.083</td>
<td>0.083</td>
<td>0.064</td>
</tr>
</tbody>
</table>
Chapter 9

Discussion

We have introduced and explored many of the properties of a new augmented data empirical likelihood calibration. It performs remarkably well in difficult problems with quite small sample sizes, and produces a versatile family of tests that would 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.

Based on the extensive simulations done in this work, we have found little difference between the performance of empirical likelihood and Hotelling’s T-square statistic for inference about a vector mean. Taking into account the simplicity of computation enjoyed by Hotelling’s T-square, and the attainment of the exact level for Gaussian data, it would seem reasonable to use Hotelling’s rather than empirical likelihood in most cases involving a test of a single-sample vector mean. Since the justification for using both tests is based on the central limit theorem in the case of non-Gaussian data, there is little reason to expect that one should be better than the other in a small-sample setting.
However, in the multi-sample case the comparison becomes more interesting. The approach of empirical likelihood can be used to derive Hotelling’s T-square test for unequal variances in the two-sample case, and to give analogous tests in multi-sample cases with the number of samples $k > 2$. The exact distribution of these tests are no longer specified even for Gaussian data, and so one advantage of Hotelling’s is eliminated. In the multi-sample setting we developed a test with competitive coverage error and roughly equivalent power by extending our one-sample results. As in the one-sample case, the resulting test is asymptotically equivalent to the Hotelling’s T-square extensions.

We are interested in investigating the potential of a Bartlett-type correction to improve the calibration of the balanced empirical likelihood test in skewed samples. Since estimating the correction factor for a 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 near-linearity of the quantile-quantile plots in the skewed distributions indicates that perhaps the skewness just scales the chi-squared distribution of the augmented empirical likelihood statistic, but does not otherwise significantly alter it. This certainly warrants further exploration and theoretical justification.

Finally, we would like to comment on the trade-off between accurate calibration and power that has been a recurring theme throughout these investigations, though not specifically highlighted in this work. Clearly, an exact level $\alpha$ test may be obtained by choosing a random variable $U \sim U(0,1)$, and rejecting $H_0$ whenever $U \geq 1 - \alpha$. However, this test has power $\alpha$ at every alternative. On the other hand, a test that orders the outcome space according to a near ideal ordering for many distributions may suffer from inaccurate level in most cases, but result in very good power. As
should be expected, we found in our explorations that whenever we had an idea that
was achieving unusually good calibration across a range of underlying distributions,
the power of the method in question would be severely diminished as compared to
its competitors. This is of course a well-known and well-studied phenomenon, but
nevertheless managed to surprise and disappoint each time it occurred. Hence the
reminder here that we cannot expect to obtain a test for the mean that achieves perfect
calibration across the space of all distributions that satisfy the moment conditions
unless we essentially ignore the observed data. Bahadur and Savage (1956) discuss
this issue in greater depth, proving that there is no “effective” test of a hypothesis
regarding the mean in a nonparametric setting, and therefore no non-trivial confidence
intervals can be accurate for this problem.
Appendix A

Proof of Theorem 6.1.1

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

$$v = \bar{Z} - \eta = -\eta$$
$$r = ||v|| = ||\eta||$$
$$u = \frac{v}{r} = \frac{-\eta}{||\eta||}.$$ 

Note that, by (6.1), $Z_{n+1} - \eta = -su$ and $Z_{n+2} - \eta = -2\eta + su = (2r + s)u$. In the following, all limits and $O(\cdot)$ and $o(\cdot)$ notations are to be interpreted as $s \to \infty$. 

90
A.1 Step 1

Since $\lambda$ satisfies

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

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

we have

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

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

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

Now since $0 < w_i < 1$ and

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

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

$$(n + 2) (w_{n+1} - w_{n+2}) = \frac{1}{1 - s\lambda^Tu} - \frac{1}{1 + (2r + s)\lambda^Tu} = O(s^{-1}).$$  \hfill (A.2)

Then since $1 - (1 - s\lambda^Tu)$ and $1 - (1 + (2r + s)\lambda^Tu)$ are of opposite signs, we clearly must have $s\lambda^Tu \to 0$ as $s \to \infty$. Thus we have

$$\lambda^Tu = o(s^{-1}).$$  \hfill (A.3)
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
\[
\sum_{i=1}^{n+2} \lambda^T(Z_i - \eta) = \lambda^T(n + 2)(-\eta) = (n + 2)r\lambda^Tu
\]
and the bound given by (A.3), we conclude that
\[
\max_{i=1,\ldots,n+2} \left\{ 1 + \lambda^T(Z_i - \eta) \right\} \leq 1 + (n + 2)r\lambda^Tu + (n + 1) = O(1). \tag{A.4}
\]

Now we employ the identity
\[
\frac{1}{1 + x} = \frac{1 + x - x}{1 + x} = 1 - \frac{x}{1 + x} \tag{A.5}
\]
to get
\[
0 = \sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) \frac{Z_i - \eta}{1 + \lambda^T(Z_i - \eta)}
= \sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) (Z_i - \eta) - \sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) \frac{(Z_i - \eta)(\lambda^T(Z_i - \eta))}{1 + \lambda^T(Z_i - \eta)}.
\]

Letting $\lambda = \|\lambda\|\theta$, rearranging the above equality, and multiplying both sides by $\lambda^T$, we have
\[
\sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) \lambda^T(Z_i - \eta) = \|\lambda\|^2 \sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) \theta^T(Z_i - \eta) \frac{\theta^T(Z_i - \eta)}{1 + \lambda^T(Z_i - \eta)},
\]
which gives $r\lambda^T u = \|\lambda\|^2 \theta^T \tilde{S}\theta$ where

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

Then letting $S^* = \sum_{i=1}^{n+2} \left( \frac{1}{n+2} \right) (Z_i - \eta)(Z_i - \eta)^T$ and substituting in the bound (A.4) on $\lambda^T(Z_i - \eta)$ from above, we have

$$
\|\lambda\|^2 \theta^T S^* \theta \leq \|\lambda\|^2 \theta^T \tilde{S}\theta \left[ \max_{i=1, \ldots, n+2} \{1 + \lambda^T(Z_i - \eta)\} \right] = r\lambda^T u \left[ \max_{i=1, \ldots, n+2} \{1 + \lambda^T(Z_i - \eta)\} \right] = o(s^{-1})O(1).
$$

Furthermore, $\theta^T S^* \theta \geq l_d$ where $l_d$ is the smallest eigenvalue of the matrix $\sum_{i=1}^{n}(Z_i - \eta)(Z_i - \eta)^T$. Thus $(\theta^T S^* \theta)^{-1} \leq l_d^{-1} = O(1)$, so $\|\lambda\|^2 = O(1) o(s^{-1}) O(1)$ and therefore

$$
\|\lambda\| = o(s^{-1/2}). \tag{A.6}
$$

### 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

$$
r_{n+1} = r_{n+2} = 0 \tag{A.7}
$$
since both \((Z_{n+1} - \eta)\) and \((Z_{n+2} - \eta)\) are multiples of \(u\). The remaining \(r_i, i = 1, \ldots, n\) satisfy

\[
\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (Z_i - \eta) - [(Z_i - \eta)^T u] u = 0. \tag{A.8}
\]

Also, we have

\[
\sum_{i=1}^{n} f_i r_i = \sum_{i=1}^{n} f_i [(Z_i - \eta) - f_i u] \\
= \sum_{i=1}^{n} (Z_i - \eta)(Z_i - \eta)^T u - \sum_{i=1}^{n} f_i^2 u \\
= [(n - 1)I_d + n\eta\eta^T] u - \left(\sum_{i=1}^{n} f_i^2\right) u \\
= \left[(n - 1) + nr^2 - \sum_{i=1}^{n} f_i^2\right] u.
\]

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

\[
\sum_{i=1}^{n} f_i r_i = 0. \tag{A.9}
\]

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

\[
0 = \sum_{i=1}^{n+2} \frac{(Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)} \\
= \sum_{i=1}^{n+2} \frac{f_i u}{1 + \lambda^T (Z_i - \eta)} + \sum_{i=1}^{n+2} \frac{r_i}{1 + \lambda^T (Z_i - \eta)}
\]

so that, using (A.7),

\[
\sum_{i=1}^{n+2} \frac{r_i}{1 + \lambda^T (Z_i - \eta)} = \sum_{i=1}^{n} \frac{r_i}{1 + \lambda^T (Z_i - \eta)} = 0.
\]
Then using identity (A.5) and the equality given by (A.8), we have

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

\[
= \sum_{i=1}^{n} \frac{r_i \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)}
\]

\[
= \sum_{i=1}^{n} \frac{f_i \theta^T r_i \lambda^T u}{1 + \lambda^T (Z_i - \eta)} + \sum_{i=1}^{n} \frac{\theta^T r_i \lambda^T r_i}{1 + \lambda^T (Z_i - \eta)}
\]

where in the last equality we have used the substitution \((Z_i - \eta) = f_i u + r_i\) and multiplied on the left by \(\theta^T\). Dividing by \(\|\lambda\|\) and employing identity (A.5) again, we have

\[
0 = \sum_{i=1}^{n} \frac{f_i \theta^T r_i \theta^T u}{1 + \lambda^T (Z_i - \eta)} + \sum_{i=1}^{n} \frac{(\theta^T r_i)^2}{1 + \lambda^T (Z_i - \eta)}
\]

\[
= \sum_{i=1}^{n} \frac{f_i \theta^T r_i \theta^T u}{1 + \lambda^T (Z_i - \eta)} - \sum_{i=1}^{n} \frac{f_i \theta^T r_i \theta^T u \lambda^T (Z_i - \eta)}{1 + \lambda^T (Z_i - \eta)}
\]

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

The first term of the last equality is 0 by (A.9), and the second and fourth terms are both \(o(s^{-1/2})\) by (A.6) 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

$$\theta^T u \to 1 \quad (A.10)$$

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

$$\|\lambda\| = o(s^{-1}). \quad (A.11)$$

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

$$\frac{1}{1 + x} = \frac{(1 + x)(1 - x) + x^2}{1 + x} = 1 - x + \frac{x^2}{1 + x} \quad (A.12)$$

to give, using (A.11),

$$0 = \sum_{i=1}^{n+2} (Z_i - \eta) - \left( \sum_{i=1}^{n} (Z_i - \eta) \lambda^T (Z_i - \eta) + s^2 u \lambda^T u + (s + 2r)^2 u \lambda^T u \right)$$

$$+ \left( \sum_{i=1}^{n} \frac{(Z_i - \eta) [\lambda^T (Z_i - \eta)]^2}{1 + \lambda^T (Z_i - \eta)} - \frac{s^3 u (\lambda^T u)^2}{1 - s \lambda^T u} + \frac{(s + 2r)^3 u (\lambda^T u)^2}{1 + (s + 2r) \lambda^T u} \right)$$

$$= (n + 2)ru - \left( o(s^{-1}) + 2s^2 u \lambda^T u + o(1) \right)$$

$$+ \left( o(s^{-2}) - s^3 u (\lambda^T u)^2 \left[ (n + 2)(w_{n+2} - w_{n+1}) \right] + o(1) \right) .$$
APPENDIX A. PROOF OF THEOREM 6.1.1

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

$$s^2 \lambda^T u \rightarrow \frac{(n + 2)r}{2} \quad (A.13)$$

and since $\lambda^T u = \|\lambda\| \theta^T u$, by (A.10) we conclude

$$\|\lambda\| = O(s^{-2}). \quad (A.14)$$

A.5 Step 5

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

$$- \log ((n + 2)w_i) = \log (1 + \lambda^T(Z_i - \eta))$$

$$= \lambda^T(Z_i - \eta) - \frac{1}{2} (\lambda^T(Z_i - \eta))^2 + \frac{1}{3} (\lambda^T(Z_i - \eta))^3 - d_i \quad (A.15)$$

where $\|d_i\| = O(s^{-4})$ from (A.14) and the boundedness of the other terms in the expansion. Using the representation (A.15) in the expression $\tilde{W}(\eta) = -2 \sum_{i=1}^{n+2} \log ((n + 2)w_i)$, we have

$$\tilde{W}(\eta) = 2 \left[ \sum_{i=1}^{n+2} \lambda^T(Z_i - \eta) - \frac{1}{2} \sum_{i=1}^{n+2} (\lambda^T(Z_i - \eta))^2 + \frac{1}{3} \sum_{i=1}^{n+2} (\lambda^T(Z_i - \eta))^3 - \sum_{i=1}^{n+2} d_i \right]$$

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

$$\left. + \frac{1}{3} \left( \sum_{i=1}^{n} (\lambda^T(Z_i - \eta))^3 - s^3(\lambda^T u)^3 + (s + 2r)^3(\lambda^T u)^3 \right) - O(s^{-4}) \right].$$
Multiplying both sides of this equality by $s^2$ and employing (A.14) gives

$$s^2 \tilde{W}(\eta) = 2 \left[ (n + 2)rs^2\lambda^T u - \frac{1}{2} (O(s^{-2}) + 2s^4(\lambda^T u)^2 + O(s^{-1})) + \frac{1}{3} (O(s^{-4}) + O(s^{-2}) + O(s^{-3}) + O(s^{-4})) - O(s^{-2}) \right]$$

$$= 2 \left[ (n + 2)rs^2\lambda^T u - s^4(\lambda^T u)^2 + O(s^{-1}) \right].$$

Substituting in the limiting expression (A.13) for $s^2\lambda^T u$, we have

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

which simplifies to

$$\frac{2ns^2}{(n + 2)^2} \tilde{W}(\eta) \rightarrow nr^2. \quad (A.16)$$

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.
Appendix B

Bootstrap Calibration Power Comparisons
APPENDIX B. BOOTSTRAP CALIBRATION POWER COMPARISONS

Normal margins, $d = 2$, $n = 10$

<table>
<thead>
<tr>
<th>$|\mu|$</th>
<th>$\theta = (1, 1)$</th>
<th>$\theta = (1, 0)$</th>
<th>$|\mu|$</th>
<th>$\theta = (1, 1)$</th>
<th>$\theta = (1, 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.053</td>
<td>0.052</td>
<td>0.053</td>
<td>0.048</td>
<td>0.050</td>
</tr>
<tr>
<td>0.10</td>
<td>0.058</td>
<td>0.057</td>
<td>0.054</td>
<td>0.050</td>
<td>0.053</td>
</tr>
<tr>
<td>0.15</td>
<td>0.063</td>
<td>0.064</td>
<td>0.060</td>
<td>0.052</td>
<td>0.059</td>
</tr>
<tr>
<td>0.20</td>
<td>0.070</td>
<td>0.072</td>
<td>0.068</td>
<td>0.060</td>
<td>0.066</td>
</tr>
<tr>
<td>0.25</td>
<td>0.079</td>
<td>0.083</td>
<td>0.082</td>
<td>0.072</td>
<td>0.075</td>
</tr>
<tr>
<td>0.30</td>
<td>0.091</td>
<td>0.096</td>
<td>0.095</td>
<td>0.087</td>
<td>0.089</td>
</tr>
<tr>
<td>0.35</td>
<td>0.107</td>
<td>0.113</td>
<td>0.110</td>
<td>0.105</td>
<td>0.108</td>
</tr>
<tr>
<td>0.40</td>
<td>0.122</td>
<td>0.131</td>
<td>0.129</td>
<td>0.124</td>
<td>0.127</td>
</tr>
<tr>
<td>0.45</td>
<td>0.143</td>
<td>0.153</td>
<td>0.153</td>
<td>0.143</td>
<td>0.147</td>
</tr>
<tr>
<td>0.50</td>
<td>0.167</td>
<td>0.177</td>
<td>0.187</td>
<td>0.168</td>
<td>0.172</td>
</tr>
<tr>
<td>0.55</td>
<td>0.199</td>
<td>0.205</td>
<td>0.228</td>
<td>0.199</td>
<td>0.205</td>
</tr>
<tr>
<td>0.60</td>
<td>0.228</td>
<td>0.240</td>
<td>0.263</td>
<td>0.230</td>
<td>0.239</td>
</tr>
<tr>
<td>0.65</td>
<td>0.263</td>
<td>0.280</td>
<td>0.300</td>
<td>0.260</td>
<td>0.274</td>
</tr>
<tr>
<td>0.70</td>
<td>0.300</td>
<td>0.319</td>
<td>0.342</td>
<td>0.299</td>
<td>0.315</td>
</tr>
<tr>
<td>0.75</td>
<td>0.338</td>
<td>0.357</td>
<td>0.390</td>
<td>0.332</td>
<td>0.357</td>
</tr>
<tr>
<td>0.80</td>
<td>0.381</td>
<td>0.396</td>
<td>0.434</td>
<td>0.372</td>
<td>0.390</td>
</tr>
<tr>
<td>0.85</td>
<td>0.416</td>
<td>0.443</td>
<td>0.484</td>
<td>0.410</td>
<td>0.435</td>
</tr>
<tr>
<td>0.90</td>
<td>0.462</td>
<td>0.484</td>
<td>0.527</td>
<td>0.450</td>
<td>0.478</td>
</tr>
<tr>
<td>0.95</td>
<td>0.499</td>
<td>0.529</td>
<td>0.572</td>
<td>0.488</td>
<td>0.517</td>
</tr>
<tr>
<td>1.00</td>
<td>0.537</td>
<td>0.569</td>
<td>0.622</td>
<td>0.528</td>
<td>0.563</td>
</tr>
<tr>
<td>1.05</td>
<td>0.576</td>
<td>0.606</td>
<td>0.668</td>
<td>0.570</td>
<td>0.605</td>
</tr>
<tr>
<td>1.10</td>
<td>0.611</td>
<td>0.645</td>
<td>0.709</td>
<td>0.609</td>
<td>0.644</td>
</tr>
<tr>
<td>1.15</td>
<td>0.650</td>
<td>0.683</td>
<td>0.751</td>
<td>0.648</td>
<td>0.682</td>
</tr>
<tr>
<td>1.20</td>
<td>0.683</td>
<td>0.720</td>
<td>0.787</td>
<td>0.681</td>
<td>0.721</td>
</tr>
<tr>
<td>1.25</td>
<td>0.715</td>
<td>0.753</td>
<td>0.820</td>
<td>0.715</td>
<td>0.759</td>
</tr>
<tr>
<td>1.30</td>
<td>0.744</td>
<td>0.785</td>
<td>0.848</td>
<td>0.748</td>
<td>0.793</td>
</tr>
<tr>
<td>1.35</td>
<td>0.772</td>
<td>0.813</td>
<td>0.875</td>
<td>0.780</td>
<td>0.818</td>
</tr>
<tr>
<td>1.40</td>
<td>0.799</td>
<td>0.841</td>
<td>0.896</td>
<td>0.807</td>
<td>0.846</td>
</tr>
<tr>
<td>1.45</td>
<td>0.822</td>
<td>0.863</td>
<td>0.917</td>
<td>0.834</td>
<td>0.873</td>
</tr>
<tr>
<td>1.50</td>
<td>0.846</td>
<td>0.885</td>
<td>0.932</td>
<td>0.851</td>
<td>0.892</td>
</tr>
</tbody>
</table>

Table B.1: Power comparisons for the bootstrap empirical likelihood calibration versus the uncalibrated empirical likelihood method and Hotelling’s $T$-square test for multivariate data with independent Normal margins in $d = 2$ dimensions, for a sample of size $n = 10$. Here $\theta$ is the direction and $\|\mu\|$ is the norm of the alternative $\mu$ at which the corresponding powers are computed.
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
$\|\mu\|$ & Bootstrap & ELM & Hotelling & Bootstrap & ELM & Hotelling \\
\hline
0.05 & 0.055 & 0.057 & 0.057 & 0.054 & 0.056 & 0.056 \\
0.10 & 0.061 & 0.066 & 0.064 & 0.058 & 0.060 & 0.061 \\
0.15 & 0.068 & 0.073 & 0.071 & 0.064 & 0.068 & 0.067 \\
0.20 & 0.077 & 0.079 & 0.080 & 0.070 & 0.073 & 0.073 \\
0.25 & 0.083 & 0.089 & 0.088 & 0.074 & 0.080 & 0.079 \\
0.30 & 0.091 & 0.098 & 0.098 & 0.082 & 0.087 & 0.086 \\
0.35 & 0.100 & 0.109 & 0.108 & 0.089 & 0.094 & 0.094 \\
0.40 & 0.111 & 0.118 & 0.118 & 0.095 & 0.103 & 0.104 \\
0.45 & 0.119 & 0.129 & 0.129 & 0.102 & 0.111 & 0.113 \\
0.50 & 0.128 & 0.140 & 0.138 & 0.109 & 0.122 & 0.122 \\
0.55 & 0.138 & 0.153 & 0.153 & 0.118 & 0.132 & 0.133 \\
0.60 & 0.150 & 0.165 & 0.166 & 0.128 & 0.142 & 0.144 \\
0.65 & 0.160 & 0.178 & 0.176 & 0.135 & 0.149 & 0.155 \\
0.70 & 0.171 & 0.191 & 0.189 & 0.144 & 0.159 & 0.164 \\
0.75 & 0.182 & 0.204 & 0.202 & 0.153 & 0.171 & 0.174 \\
0.80 & 0.193 & 0.216 & 0.220 & 0.160 & 0.181 & 0.186 \\
0.85 & 0.204 & 0.229 & 0.234 & 0.170 & 0.190 & 0.199 \\
0.90 & 0.217 & 0.243 & 0.248 & 0.179 & 0.201 & 0.209 \\
0.95 & 0.227 & 0.258 & 0.261 & 0.192 & 0.213 & 0.220 \\
1.00 & 0.237 & 0.277 & 0.278 & 0.200 & 0.223 & 0.236 \\
1.05 & 0.247 & 0.293 & 0.292 & 0.209 & 0.236 & 0.248 \\
1.10 & 0.259 & 0.308 & 0.308 & 0.221 & 0.250 & 0.262 \\
1.15 & 0.273 & 0.325 & 0.323 & 0.232 & 0.262 & 0.277 \\
1.20 & 0.285 & 0.340 & 0.343 & 0.241 & 0.273 & 0.289 \\
1.25 & 0.300 & 0.360 & 0.364 & 0.252 & 0.285 & 0.303 \\
1.30 & 0.312 & 0.377 & 0.382 & 0.260 & 0.300 & 0.315 \\
1.35 & 0.323 & 0.392 & 0.398 & 0.269 & 0.314 & 0.329 \\
1.40 & 0.337 & 0.409 & 0.415 & 0.279 & 0.328 & 0.342 \\
1.45 & 0.350 & 0.427 & 0.430 & 0.291 & 0.343 & 0.357 \\
1.50 & 0.362 & 0.443 & 0.446 & 0.301 & 0.356 & 0.371 \\
\hline
\end{tabular}
\caption{Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 2$, $n = 10$}
\end{table}
APPENDIX B. BOOTSTRAP CALIBRATION POWER COMPARISONS

Normal margins, $d = 2$, $n = 40$

| $||\mu||$ | $\theta = (1,1)$ | $\theta = (1,0)$ |
|----------|----------------|----------------|
|          | Bootstrap      | ELM            | Hotelling |
| 0.05     | 0.057          | 0.057          | 0.058     |
| 0.10     | 0.079          | 0.077          | 0.079     |
| 0.15     | 0.120          | 0.119          | 0.119     |
| 0.20     | 0.180          | 0.178          | 0.179     |
| 0.25     | 0.263          | 0.259          | 0.264     |
| 0.30     | 0.356          | 0.351          | 0.363     |
| 0.35     | 0.462          | 0.457          | 0.470     |
| 0.40     | 0.580          | 0.577          | 0.586     |
| 0.45     | 0.690          | 0.688          | 0.693     |
| 0.50     | 0.783          | 0.780          | 0.788     |
| 0.55     | 0.853          | 0.852          | 0.858     |
| 0.60     | 0.911          | 0.910          | 0.913     |
| 0.65     | 0.949          | 0.952          | 0.956     |
| 0.70     | 0.975          | 0.974          | 0.977     |
| 0.75     | 0.989          | 0.989          | 0.989     |
| 0.80     | 0.994          | 0.993          | 0.994     |
| 0.85     | 0.997          | 0.997          | 0.997     |
| 0.90     | 0.999          | 0.999          | 0.999     |
| 0.95     | 1.000          | 1.000          | 0.999     |
| 1.00     | 1.000          | 1.000          | 1.000     |
| 1.05     | 1.000          | 1.000          | 1.000     |
| 1.10     | 1.000          | 1.000          | 1.000     |
| 1.15     | 1.000          | 1.000          | 1.000     |
| 1.20     | 1.000          | 1.000          | 1.000     |
| 1.25     | 1.000          | 1.000          | 1.000     |
| 1.30     | 1.000          | 1.000          | 1.000     |
| 1.35     | 1.000          | 1.000          | 1.000     |
| 1.40     | 1.000          | 1.000          | 1.000     |
| 1.45     | 1.000          | 1.000          | 1.000     |
| 1.50     | 1.000          | 1.000          | 1.000     |

Table B.3: Bootstrap calibration power comparisons: Normal data, $d = 2$, $n = 40$
$\chi^2_{(1)}$ margins, $d = 2$, $n = 40$

<table>
<thead>
<tr>
<th>$\parallel \mu \parallel$</th>
<th>$\theta = (1, 1)$</th>
<th>$\theta = (1, 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bootstrap</td>
<td>ELM</td>
</tr>
<tr>
<td>0.05</td>
<td>0.059</td>
<td>0.061</td>
</tr>
<tr>
<td>0.10</td>
<td>0.073</td>
<td>0.075</td>
</tr>
<tr>
<td>0.15</td>
<td>0.091</td>
<td>0.095</td>
</tr>
<tr>
<td>0.20</td>
<td>0.116</td>
<td>0.122</td>
</tr>
<tr>
<td>0.25</td>
<td>0.149</td>
<td>0.155</td>
</tr>
<tr>
<td>0.30</td>
<td>0.183</td>
<td>0.194</td>
</tr>
<tr>
<td>0.35</td>
<td>0.225</td>
<td>0.237</td>
</tr>
<tr>
<td>0.40</td>
<td>0.269</td>
<td>0.287</td>
</tr>
<tr>
<td>0.45</td>
<td>0.315</td>
<td>0.336</td>
</tr>
<tr>
<td>0.50</td>
<td>0.365</td>
<td>0.390</td>
</tr>
<tr>
<td>0.55</td>
<td>0.413</td>
<td>0.438</td>
</tr>
<tr>
<td>0.60</td>
<td>0.459</td>
<td>0.488</td>
</tr>
<tr>
<td>0.65</td>
<td>0.507</td>
<td>0.539</td>
</tr>
<tr>
<td>0.70</td>
<td>0.551</td>
<td>0.589</td>
</tr>
<tr>
<td>0.75</td>
<td>0.599</td>
<td>0.636</td>
</tr>
<tr>
<td>0.80</td>
<td>0.646</td>
<td>0.689</td>
</tr>
<tr>
<td>0.85</td>
<td>0.694</td>
<td>0.735</td>
</tr>
<tr>
<td>0.90</td>
<td>0.731</td>
<td>0.778</td>
</tr>
<tr>
<td>0.95</td>
<td>0.767</td>
<td>0.814</td>
</tr>
<tr>
<td>1.00</td>
<td>0.801</td>
<td>0.845</td>
</tr>
<tr>
<td>1.05</td>
<td>0.829</td>
<td>0.868</td>
</tr>
<tr>
<td>1.10</td>
<td>0.849</td>
<td>0.892</td>
</tr>
<tr>
<td>1.15</td>
<td>0.872</td>
<td>0.911</td>
</tr>
<tr>
<td>1.20</td>
<td>0.891</td>
<td>0.929</td>
</tr>
<tr>
<td>1.25</td>
<td>0.906</td>
<td>0.941</td>
</tr>
<tr>
<td>1.30</td>
<td>0.921</td>
<td>0.951</td>
</tr>
<tr>
<td>1.35</td>
<td>0.931</td>
<td>0.961</td>
</tr>
<tr>
<td>1.40</td>
<td>0.941</td>
<td>0.970</td>
</tr>
<tr>
<td>1.45</td>
<td>0.948</td>
<td>0.976</td>
</tr>
<tr>
<td>1.50</td>
<td>0.956</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Table B.4: Bootstrap calibration power comparisons: $\chi^2_{(1)}$ data, $d = 2$, $n = 40$
Normal margins, $d = 4$, $n = 10$

$$\theta = (1,1,1,1)$$

$\theta = (1,0,0,0)$

<table>
<thead>
<tr>
<th>$\parallel \mu \parallel$</th>
<th>Bootstrap ELM Hotelling</th>
<th>Bootstrap ELM Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.052 0.049 0.050</td>
<td>0.054 0.051 0.051</td>
</tr>
<tr>
<td>0.10</td>
<td>0.054 0.052 0.052</td>
<td>0.056 0.054 0.053</td>
</tr>
<tr>
<td>0.15</td>
<td>0.057 0.054 0.054</td>
<td>0.059 0.057 0.057</td>
</tr>
<tr>
<td>0.20</td>
<td>0.059 0.059 0.060</td>
<td>0.063 0.061 0.062</td>
</tr>
<tr>
<td>0.25</td>
<td>0.066 0.066 0.066</td>
<td>0.068 0.066 0.068</td>
</tr>
<tr>
<td>0.30</td>
<td>0.074 0.074 0.074</td>
<td>0.077 0.074 0.074</td>
</tr>
<tr>
<td>0.35</td>
<td>0.083 0.081 0.083</td>
<td>0.082 0.081 0.080</td>
</tr>
<tr>
<td>0.40</td>
<td>0.090 0.092 0.092</td>
<td>0.093 0.091 0.091</td>
</tr>
<tr>
<td>0.45</td>
<td>0.100 0.104 0.104</td>
<td>0.105 0.102 0.102</td>
</tr>
<tr>
<td>0.50</td>
<td>0.113 0.119 0.119</td>
<td>0.115 0.115 0.114</td>
</tr>
<tr>
<td>0.55</td>
<td>0.127 0.133 0.136</td>
<td>0.127 0.131 0.130</td>
</tr>
<tr>
<td>0.60</td>
<td>0.144 0.154 0.152</td>
<td>0.143 0.151 0.149</td>
</tr>
<tr>
<td>0.65</td>
<td>0.161 0.177 0.173</td>
<td>0.161 0.174 0.172</td>
</tr>
<tr>
<td>0.70</td>
<td>0.179 0.196 0.202</td>
<td>0.178 0.195 0.197</td>
</tr>
<tr>
<td>0.75</td>
<td>0.201 0.222 0.226</td>
<td>0.201 0.221 0.222</td>
</tr>
<tr>
<td>0.80</td>
<td>0.228 0.249 0.250</td>
<td>0.229 0.249 0.250</td>
</tr>
<tr>
<td>0.85</td>
<td>0.252 0.274 0.278</td>
<td>0.257 0.272 0.283</td>
</tr>
<tr>
<td>0.90</td>
<td>0.279 0.300 0.307</td>
<td>0.285 0.304 0.313</td>
</tr>
<tr>
<td>0.95</td>
<td>0.308 0.328 0.338</td>
<td>0.315 0.336 0.345</td>
</tr>
<tr>
<td>1.00</td>
<td>0.337 0.363 0.374</td>
<td>0.349 0.370 0.383</td>
</tr>
<tr>
<td>1.05</td>
<td>0.365 0.399 0.409</td>
<td>0.379 0.409 0.419</td>
</tr>
<tr>
<td>1.10</td>
<td>0.397 0.435 0.446</td>
<td>0.414 0.440 0.455</td>
</tr>
<tr>
<td>1.15</td>
<td>0.430 0.469 0.484</td>
<td>0.442 0.474 0.489</td>
</tr>
<tr>
<td>1.20</td>
<td>0.466 0.504 0.521</td>
<td>0.473 0.507 0.525</td>
</tr>
<tr>
<td>1.25</td>
<td>0.500 0.541 0.559</td>
<td>0.505 0.542 0.561</td>
</tr>
<tr>
<td>1.30</td>
<td>0.531 0.579 0.596</td>
<td>0.536 0.577 0.594</td>
</tr>
<tr>
<td>1.35</td>
<td>0.565 0.613 0.626</td>
<td>0.563 0.612 0.633</td>
</tr>
<tr>
<td>1.40</td>
<td>0.595 0.647 0.664</td>
<td>0.593 0.650 0.666</td>
</tr>
<tr>
<td>1.45</td>
<td>0.626 0.676 0.699</td>
<td>0.631 0.682 0.701</td>
</tr>
<tr>
<td>1.50</td>
<td>0.656 0.712 0.733</td>
<td>0.661 0.713 0.738</td>
</tr>
</tbody>
</table>

*Table B.5: Bootstrap calibration power comparisons: Normal data, d = 4, n = 10*
\[ \chi^2_{(1)} \text{ margins, } d = 4, n = 10 \]

<table>
<thead>
<tr>
<th>( |\mu| )</th>
<th>( \theta = (1, 1, 1, 1) )</th>
<th>( \theta = (1, 0, 0, 0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bootstrap</td>
<td>ELM</td>
</tr>
<tr>
<td>0.05</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td>0.10</td>
<td>0.058</td>
<td>0.062</td>
</tr>
<tr>
<td>0.15</td>
<td>0.064</td>
<td>0.069</td>
</tr>
<tr>
<td>0.20</td>
<td>0.070</td>
<td>0.076</td>
</tr>
<tr>
<td>0.25</td>
<td>0.077</td>
<td>0.082</td>
</tr>
<tr>
<td>0.30</td>
<td>0.084</td>
<td>0.089</td>
</tr>
<tr>
<td>0.35</td>
<td>0.091</td>
<td>0.097</td>
</tr>
<tr>
<td>0.40</td>
<td>0.097</td>
<td>0.105</td>
</tr>
<tr>
<td>0.45</td>
<td>0.106</td>
<td>0.114</td>
</tr>
<tr>
<td>0.50</td>
<td>0.114</td>
<td>0.121</td>
</tr>
<tr>
<td>0.55</td>
<td>0.119</td>
<td>0.129</td>
</tr>
<tr>
<td>0.60</td>
<td>0.126</td>
<td>0.137</td>
</tr>
<tr>
<td>0.65</td>
<td>0.134</td>
<td>0.146</td>
</tr>
<tr>
<td>0.70</td>
<td>0.144</td>
<td>0.154</td>
</tr>
<tr>
<td>0.75</td>
<td>0.151</td>
<td>0.163</td>
</tr>
<tr>
<td>0.80</td>
<td>0.157</td>
<td>0.172</td>
</tr>
<tr>
<td>0.85</td>
<td>0.163</td>
<td>0.181</td>
</tr>
<tr>
<td>0.90</td>
<td>0.173</td>
<td>0.192</td>
</tr>
<tr>
<td>0.95</td>
<td>0.183</td>
<td>0.202</td>
</tr>
<tr>
<td>1.00</td>
<td>0.191</td>
<td>0.213</td>
</tr>
<tr>
<td>1.05</td>
<td>0.198</td>
<td>0.222</td>
</tr>
<tr>
<td>1.10</td>
<td>0.206</td>
<td>0.236</td>
</tr>
<tr>
<td>1.15</td>
<td>0.216</td>
<td>0.248</td>
</tr>
<tr>
<td>1.20</td>
<td>0.228</td>
<td>0.261</td>
</tr>
<tr>
<td>1.25</td>
<td>0.239</td>
<td>0.274</td>
</tr>
<tr>
<td>1.30</td>
<td>0.248</td>
<td>0.285</td>
</tr>
<tr>
<td>1.35</td>
<td>0.259</td>
<td>0.297</td>
</tr>
<tr>
<td>1.40</td>
<td>0.269</td>
<td>0.309</td>
</tr>
<tr>
<td>1.45</td>
<td>0.278</td>
<td>0.320</td>
</tr>
<tr>
<td>1.50</td>
<td>0.287</td>
<td>0.335</td>
</tr>
</tbody>
</table>

Table B.6: Bootstrap calibration power comparisons: \( \chi^2_{(1)} \) data, \( d = 4, n = 10 \)
Normal margins, $d = 4$, $n = 40$

<table>
<thead>
<tr>
<th>$|\mu|$</th>
<th>Bootstrap</th>
<th>ELM</th>
<th>Hotelling</th>
<th>Bootstrap</th>
<th>ELM</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.056</td>
<td>0.056</td>
<td>0.053</td>
<td>0.055</td>
<td>0.054</td>
<td>0.053</td>
</tr>
<tr>
<td>0.10</td>
<td>0.067</td>
<td>0.066</td>
<td>0.065</td>
<td>0.067</td>
<td>0.065</td>
<td>0.063</td>
</tr>
<tr>
<td>0.15</td>
<td>0.091</td>
<td>0.088</td>
<td>0.088</td>
<td>0.091</td>
<td>0.088</td>
<td>0.092</td>
</tr>
<tr>
<td>0.20</td>
<td>0.129</td>
<td>0.123</td>
<td>0.127</td>
<td>0.123</td>
<td>0.121</td>
<td>0.124</td>
</tr>
<tr>
<td>0.25</td>
<td>0.183</td>
<td>0.180</td>
<td>0.182</td>
<td>0.178</td>
<td>0.177</td>
<td>0.174</td>
</tr>
<tr>
<td>0.30</td>
<td>0.249</td>
<td>0.244</td>
<td>0.253</td>
<td>0.244</td>
<td>0.241</td>
<td>0.247</td>
</tr>
<tr>
<td>0.35</td>
<td>0.336</td>
<td>0.333</td>
<td>0.337</td>
<td>0.332</td>
<td>0.327</td>
<td>0.333</td>
</tr>
<tr>
<td>0.40</td>
<td>0.431</td>
<td>0.427</td>
<td>0.438</td>
<td>0.423</td>
<td>0.419</td>
<td>0.433</td>
</tr>
<tr>
<td>0.45</td>
<td>0.533</td>
<td>0.528</td>
<td>0.540</td>
<td>0.530</td>
<td>0.530</td>
<td>0.543</td>
</tr>
<tr>
<td>0.50</td>
<td>0.632</td>
<td>0.630</td>
<td>0.651</td>
<td>0.630</td>
<td>0.628</td>
<td>0.641</td>
</tr>
<tr>
<td>0.55</td>
<td>0.732</td>
<td>0.730</td>
<td>0.745</td>
<td>0.721</td>
<td>0.720</td>
<td>0.736</td>
</tr>
<tr>
<td>0.60</td>
<td>0.811</td>
<td>0.811</td>
<td>0.825</td>
<td>0.800</td>
<td>0.800</td>
<td>0.817</td>
</tr>
<tr>
<td>0.65</td>
<td>0.875</td>
<td>0.876</td>
<td>0.888</td>
<td>0.869</td>
<td>0.869</td>
<td>0.878</td>
</tr>
<tr>
<td>0.70</td>
<td>0.921</td>
<td>0.919</td>
<td>0.931</td>
<td>0.917</td>
<td>0.917</td>
<td>0.928</td>
</tr>
<tr>
<td>0.75</td>
<td>0.950</td>
<td>0.951</td>
<td>0.959</td>
<td>0.950</td>
<td>0.951</td>
<td>0.958</td>
</tr>
<tr>
<td>0.80</td>
<td>0.973</td>
<td>0.974</td>
<td>0.979</td>
<td>0.971</td>
<td>0.972</td>
<td>0.978</td>
</tr>
<tr>
<td>0.85</td>
<td>0.986</td>
<td>0.987</td>
<td>0.990</td>
<td>0.985</td>
<td>0.986</td>
<td>0.987</td>
</tr>
<tr>
<td>0.90</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td>0.95</td>
<td>0.996</td>
<td>0.996</td>
<td>0.998</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
</tr>
<tr>
<td>1.00</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>0.998</td>
<td>0.999</td>
</tr>
<tr>
<td>1.05</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.10</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.15</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.20</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.25</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.30</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.35</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.40</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.45</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.50</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table B.7: Bootstrap calibration power comparisons: Normal data, $d = 4$, $n = 40$
APPENDIX B. BOOTSTRAP CALIBRATION POWER COMPARISONS

\[ \chi^2(1) \] margins, \( d = 4, \ n = 40 \)

| \|\mu\| | Bootstrap ELM Hotelling | \( \theta = (1,1,1,1) \) | Bootstrap ELM Hotelling | \( \theta = (1,0,0,0) \) |
|-----|--------|-----------------|--------|-----------------|--------|
| 0.05 | 0.061  | 0.060 0.061 | 0.059  | 0.057 0.056 |
| 0.10 | 0.073  | 0.071 0.078 | 0.066  | 0.066 0.067 |
| 0.15 | 0.090  | 0.087 0.099 | 0.078  | 0.077 0.079 |
| 0.20 | 0.115  | 0.110 0.125 | 0.091  | 0.091 0.101 |
| 0.25 | 0.141  | 0.137 0.153 | 0.111  | 0.108 0.121 |
| 0.30 | 0.174  | 0.169 0.185 | 0.133  | 0.127 0.143 |
| 0.35 | 0.204  | 0.204 0.221 | 0.153  | 0.150 0.165 |
| 0.40 | 0.236  | 0.236 0.263 | 0.184  | 0.177 0.200 |
| 0.45 | 0.277  | 0.277 0.308 | 0.211  | 0.208 0.235 |
| 0.50 | 0.321  | 0.323 0.358 | 0.242  | 0.238 0.271 |
| 0.55 | 0.362  | 0.374 0.409 | 0.271  | 0.272 0.313 |
| 0.60 | 0.415  | 0.425 0.462 | 0.300  | 0.307 0.352 |
| 0.65 | 0.460  | 0.476 0.512 | 0.335  | 0.344 0.397 |
| 0.70 | 0.503  | 0.524 0.565 | 0.370  | 0.379 0.444 |
| 0.75 | 0.551  | 0.577 0.616 | 0.400  | 0.414 0.490 |
| 0.80 | 0.596  | 0.622 0.663 | 0.435  | 0.452 0.535 |
| 0.85 | 0.645  | 0.671 0.710 | 0.479  | 0.490 0.580 |
| 0.90 | 0.688  | 0.715 0.759 | 0.515  | 0.529 0.622 |
| 0.95 | 0.726  | 0.757 0.798 | 0.548  | 0.572 0.656 |
| 1.00 | 0.761  | 0.795 0.834 | 0.583  | 0.603 0.690 |
| 1.05 | 0.793  | 0.830 0.864 | 0.610  | 0.637 0.727 |
| 1.10 | 0.820  | 0.865 0.894 | 0.642  | 0.668 0.759 |
| 1.15 | 0.848  | 0.888 0.914 | 0.670  | 0.699 0.787 |
| 1.20 | 0.874  | 0.907 0.929 | 0.701  | 0.730 0.814 |
| 1.25 | 0.893  | 0.924 0.942 | 0.727  | 0.759 0.835 |
| 1.30 | 0.908  | 0.941 0.955 | 0.751  | 0.785 0.859 |
| 1.35 | 0.924  | 0.951 0.964 | 0.774  | 0.806 0.877 |
| 1.40 | 0.937  | 0.960 0.973 | 0.797  | 0.825 0.895 |
| 1.45 | 0.947  | 0.968 0.980 | 0.815  | 0.841 0.910 |
| 1.50 | 0.956  | 0.976 0.984 | 0.832  | 0.857 0.922 |

Table B.8: Bootstrap calibration power comparisons: \( \chi^2(1) \) data, \( d = 4, \ n = 40 \)
Appendix C

BEL Power Comparisons

C.1 Normal Data, $d = 4$, $n = 10$

The following tables compare the power at the alternative $\mu$ listed above each table for the ordinary empirical likelihood method (EL); the adjusted empirical likelihood method of Chen et al. (2008) (AEL); the balanced empirical likelihood method for three different values of $s$ (BEL(1.0), BEL(1.9), and BEL(2.5)); and Hotelling’s T-square test (Hotelling). The data for these examples come from a multivariate Gaussian distribution. Only one alternative direction is considered, as all of these tests are invariant under rotations and the distribution has elliptical contours.

<table>
<thead>
<tr>
<th>$\mu = (0.15, 0, 0, 0)$</th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.091</td>
<td>0.087</td>
<td>0.089</td>
<td>0.092</td>
<td>0.093</td>
<td>0.092</td>
</tr>
<tr>
<td>0.05</td>
<td>0.035</td>
<td>0.039</td>
<td>0.042</td>
<td>0.041</td>
<td>0.041</td>
<td>0.040</td>
</tr>
<tr>
<td>0.02</td>
<td>0.015</td>
<td>0.016</td>
<td>0.015</td>
<td>0.014</td>
<td>0.014</td>
<td>0.013</td>
</tr>
<tr>
<td>0.01</td>
<td>0.006</td>
<td>0.007</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
</tbody>
</table>
### APPENDIX C. BEL POWER COMPARISONS

#### $\mu = (0.3, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.128</td>
<td>0.120</td>
<td>0.125</td>
<td>0.130</td>
<td>0.134</td>
<td>0.130</td>
</tr>
<tr>
<td>0.05</td>
<td>0.060</td>
<td>0.061</td>
<td>0.065</td>
<td>0.062</td>
<td>0.060</td>
<td>0.060</td>
</tr>
<tr>
<td>0.02</td>
<td>0.023</td>
<td>0.021</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
<td>0.020</td>
</tr>
<tr>
<td>0.01</td>
<td>0.013</td>
<td>0.013</td>
<td>0.012</td>
<td>0.012</td>
<td>0.012</td>
<td>0.011</td>
</tr>
</tbody>
</table>

#### $\mu = (0.45, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.206</td>
<td>0.202</td>
<td>0.212</td>
<td>0.211</td>
<td>0.212</td>
<td>0.207</td>
</tr>
<tr>
<td>0.05</td>
<td>0.092</td>
<td>0.101</td>
<td>0.104</td>
<td>0.099</td>
<td>0.099</td>
<td>0.096</td>
</tr>
<tr>
<td>0.02</td>
<td>0.052</td>
<td>0.051</td>
<td>0.048</td>
<td>0.046</td>
<td>0.046</td>
<td>0.045</td>
</tr>
<tr>
<td>0.01</td>
<td>0.028</td>
<td>0.027</td>
<td>0.027</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
</tbody>
</table>

#### $\mu = (0.6, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.269</td>
<td>0.248</td>
<td>0.259</td>
<td>0.263</td>
<td>0.265</td>
<td>0.261</td>
</tr>
<tr>
<td>0.05</td>
<td>0.134</td>
<td>0.138</td>
<td>0.145</td>
<td>0.142</td>
<td>0.142</td>
<td>0.141</td>
</tr>
<tr>
<td>0.02</td>
<td>0.080</td>
<td>0.080</td>
<td>0.076</td>
<td>0.075</td>
<td>0.074</td>
<td>0.074</td>
</tr>
<tr>
<td>0.01</td>
<td>0.048</td>
<td>0.046</td>
<td>0.043</td>
<td>0.042</td>
<td>0.042</td>
<td>0.042</td>
</tr>
</tbody>
</table>

#### $\mu = (0.75, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.355</td>
<td>0.355</td>
<td>0.380</td>
<td>0.390</td>
<td>0.391</td>
<td>0.386</td>
</tr>
<tr>
<td>0.05</td>
<td>0.208</td>
<td>0.215</td>
<td>0.223</td>
<td>0.222</td>
<td>0.221</td>
<td>0.218</td>
</tr>
<tr>
<td>0.02</td>
<td>0.107</td>
<td>0.113</td>
<td>0.103</td>
<td>0.102</td>
<td>0.101</td>
<td>0.099</td>
</tr>
<tr>
<td>0.01</td>
<td>0.056</td>
<td>0.053</td>
<td>0.048</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
</tr>
</tbody>
</table>

#### $\mu = (0.9, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.439</td>
<td>0.438</td>
<td>0.458</td>
<td>0.460</td>
<td>0.461</td>
<td>0.455</td>
</tr>
<tr>
<td>0.05</td>
<td>0.271</td>
<td>0.284</td>
<td>0.294</td>
<td>0.289</td>
<td>0.285</td>
<td>0.285</td>
</tr>
<tr>
<td>0.02</td>
<td>0.164</td>
<td>0.165</td>
<td>0.153</td>
<td>0.148</td>
<td>0.148</td>
<td>0.148</td>
</tr>
<tr>
<td>0.01</td>
<td>0.090</td>
<td>0.092</td>
<td>0.088</td>
<td>0.085</td>
<td>0.085</td>
<td>0.084</td>
</tr>
</tbody>
</table>
### APPENDIX C. BEL POWER COMPARISONS

\[ \mu = (1.05, 0, 0, 0) \]

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.560</td>
<td>0.565</td>
<td>0.588</td>
<td>0.596</td>
<td>0.599</td>
<td>0.593</td>
</tr>
<tr>
<td>0.05</td>
<td>0.382</td>
<td>0.401</td>
<td>0.416</td>
<td>0.410</td>
<td>0.406</td>
<td>0.406</td>
</tr>
<tr>
<td>0.02</td>
<td>0.245</td>
<td>0.250</td>
<td>0.236</td>
<td>0.229</td>
<td>0.229</td>
<td>0.227</td>
</tr>
<tr>
<td>0.01</td>
<td>0.148</td>
<td>0.148</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
</tr>
</tbody>
</table>

\[ \mu = (1.2, 0, 0, 0) \]

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.669</td>
<td>0.658</td>
<td>0.683</td>
<td>0.691</td>
<td>0.693</td>
<td>0.687</td>
</tr>
<tr>
<td>0.05</td>
<td>0.440</td>
<td>0.460</td>
<td>0.475</td>
<td>0.471</td>
<td>0.471</td>
<td>0.472</td>
</tr>
<tr>
<td>0.02</td>
<td>0.293</td>
<td>0.299</td>
<td>0.292</td>
<td>0.285</td>
<td>0.284</td>
<td>0.282</td>
</tr>
<tr>
<td>0.01</td>
<td>0.207</td>
<td>0.204</td>
<td>0.192</td>
<td>0.187</td>
<td>0.187</td>
<td>0.187</td>
</tr>
</tbody>
</table>

\[ \mu = (1.35, 0, 0, 0) \]

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.772</td>
<td>0.774</td>
<td>0.794</td>
<td>0.798</td>
<td>0.798</td>
<td>0.795</td>
</tr>
<tr>
<td>0.05</td>
<td>0.583</td>
<td>0.609</td>
<td>0.624</td>
<td>0.617</td>
<td>0.615</td>
<td>0.616</td>
</tr>
<tr>
<td>0.02</td>
<td>0.426</td>
<td>0.434</td>
<td>0.413</td>
<td>0.405</td>
<td>0.405</td>
<td>0.402</td>
</tr>
<tr>
<td>0.01</td>
<td>0.309</td>
<td>0.309</td>
<td>0.297</td>
<td>0.295</td>
<td>0.293</td>
<td>0.291</td>
</tr>
</tbody>
</table>

\[ \mu = (1.5, 0, 0, 0) \]

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.865</td>
<td>0.863</td>
<td>0.888</td>
<td>0.892</td>
<td>0.892</td>
<td>0.887</td>
</tr>
<tr>
<td>0.05</td>
<td>0.683</td>
<td>0.712</td>
<td>0.735</td>
<td>0.725</td>
<td>0.724</td>
<td>0.722</td>
</tr>
<tr>
<td>0.02</td>
<td>0.525</td>
<td>0.529</td>
<td>0.503</td>
<td>0.497</td>
<td>0.497</td>
<td>0.493</td>
</tr>
<tr>
<td>0.01</td>
<td>0.368</td>
<td>0.369</td>
<td>0.338</td>
<td>0.336</td>
<td>0.335</td>
<td>0.333</td>
</tr>
</tbody>
</table>

\[ \mu = (1.65, 0, 0, 0) \]

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.898</td>
<td>0.902</td>
<td>0.914</td>
<td>0.919</td>
<td>0.922</td>
<td>0.921</td>
</tr>
<tr>
<td>0.05</td>
<td>0.770</td>
<td>0.794</td>
<td>0.804</td>
<td>0.798</td>
<td>0.796</td>
<td>0.798</td>
</tr>
<tr>
<td>0.02</td>
<td>0.604</td>
<td>0.612</td>
<td>0.586</td>
<td>0.581</td>
<td>0.580</td>
<td>0.578</td>
</tr>
<tr>
<td>0.01</td>
<td>0.449</td>
<td>0.444</td>
<td>0.418</td>
<td>0.419</td>
<td>0.417</td>
<td>0.418</td>
</tr>
</tbody>
</table>
\( \mu = (1.8, 0, 0, 0) \)

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.934</td>
<td>0.946</td>
<td>0.960</td>
<td>0.959</td>
<td>0.959</td>
<td>0.957</td>
</tr>
<tr>
<td>0.05</td>
<td>0.814</td>
<td>0.836</td>
<td>0.861</td>
<td>0.861</td>
<td>0.860</td>
<td>0.858</td>
</tr>
<tr>
<td>0.02</td>
<td>0.678</td>
<td>0.691</td>
<td>0.671</td>
<td>0.664</td>
<td>0.664</td>
<td>0.661</td>
</tr>
<tr>
<td>0.01</td>
<td>0.525</td>
<td>0.526</td>
<td>0.514</td>
<td>0.509</td>
<td>0.508</td>
<td>0.508</td>
</tr>
</tbody>
</table>

\[ \chi^2_{(1)} \text{ Data, } d = 8, \ n = 40 \]

The following tables compare the power at the alternative \( \mu \) listed above each table for the ordinary empirical likelihood method (EL); the adjusted empirical likelihood method of Chen et al. (2008) (AEL); the balanced empirical likelihood method for three different values of \( s \) (BEL(1.0), BEL(1.9), and BEL(2.5)); and Hotelling’s T-square test (Hotelling). The data for these examples come from a multivariate distribution with independent \( \chi^2_{(1)} \) margins.

\( \mu = 0.35(1, 1, 1, 1, 1, 1, 1, 1) \)

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.814</td>
<td>0.744</td>
<td>0.727</td>
<td>0.559</td>
<td>0.519</td>
<td>0.442</td>
</tr>
<tr>
<td>0.05</td>
<td>0.511</td>
<td>0.395</td>
<td>0.373</td>
<td>0.261</td>
<td>0.246</td>
<td>0.190</td>
</tr>
<tr>
<td>0.02</td>
<td>0.255</td>
<td>0.143</td>
<td>0.124</td>
<td>0.052</td>
<td>0.039</td>
<td>0.027</td>
</tr>
<tr>
<td>0.01</td>
<td>0.109</td>
<td>0.021</td>
<td>0.014</td>
<td>0.012</td>
<td>0.010</td>
<td>0.009</td>
</tr>
</tbody>
</table>

\( \mu = -0.35(1, 1, 1, 1, 1, 1, 1, 1) \)

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.854</td>
<td>0.859</td>
<td>0.860</td>
<td>0.864</td>
<td>0.871</td>
<td>0.870</td>
</tr>
<tr>
<td>0.05</td>
<td>0.749</td>
<td>0.785</td>
<td>0.775</td>
<td>0.786</td>
<td>0.788</td>
<td>0.780</td>
</tr>
<tr>
<td>0.02</td>
<td>0.625</td>
<td>0.650</td>
<td>0.640</td>
<td>0.619</td>
<td>0.605</td>
<td>0.595</td>
</tr>
<tr>
<td>0.01</td>
<td>0.476</td>
<td>0.467</td>
<td>0.473</td>
<td>0.475</td>
<td>0.479</td>
<td>0.484</td>
</tr>
</tbody>
</table>
### APPENDIX C. BEL POWER COMPARISONS

#### $\mu = (1, 0, 0, 0, 0, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.000</td>
<td>0.998</td>
<td>0.992</td>
<td>0.935</td>
<td>0.898</td>
<td>0.832</td>
</tr>
<tr>
<td>0.05</td>
<td>1.000</td>
<td>0.962</td>
<td>0.906</td>
<td>0.709</td>
<td>0.651</td>
<td>0.550</td>
</tr>
<tr>
<td>0.02</td>
<td>1.000</td>
<td>0.714</td>
<td>0.568</td>
<td>0.315</td>
<td>0.267</td>
<td>0.210</td>
</tr>
<tr>
<td>0.01</td>
<td>1.000</td>
<td>0.241</td>
<td>0.199</td>
<td>0.129</td>
<td>0.118</td>
<td>0.102</td>
</tr>
</tbody>
</table>

#### $\mu = (-1, 0, 0, 0, 0, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.641</td>
<td>0.657</td>
<td>0.651</td>
<td>0.685</td>
<td>0.698</td>
<td>0.704</td>
</tr>
<tr>
<td>0.05</td>
<td>0.484</td>
<td>0.532</td>
<td>0.513</td>
<td>0.546</td>
<td>0.561</td>
<td>0.580</td>
</tr>
<tr>
<td>0.02</td>
<td>0.366</td>
<td>0.385</td>
<td>0.375</td>
<td>0.380</td>
<td>0.383</td>
<td>0.389</td>
</tr>
<tr>
<td>0.01</td>
<td>0.239</td>
<td>0.256</td>
<td>0.257</td>
<td>0.281</td>
<td>0.292</td>
<td>0.306</td>
</tr>
</tbody>
</table>

#### $\mu = 0.71(1, -1, 0, 0, 0, 0, 0, 0)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.973</td>
<td>0.889</td>
<td>0.892</td>
<td>0.798</td>
<td>0.789</td>
<td>0.746</td>
</tr>
<tr>
<td>0.05</td>
<td>0.890</td>
<td>0.712</td>
<td>0.713</td>
<td>0.603</td>
<td>0.586</td>
<td>0.557</td>
</tr>
<tr>
<td>0.02</td>
<td>0.724</td>
<td>0.468</td>
<td>0.464</td>
<td>0.377</td>
<td>0.351</td>
<td>0.334</td>
</tr>
<tr>
<td>0.01</td>
<td>0.512</td>
<td>0.265</td>
<td>0.265</td>
<td>0.244</td>
<td>0.242</td>
<td>0.247</td>
</tr>
</tbody>
</table>

#### $\mu = 0.35(1, 1, 1, 1, -1, -1, -1, -1)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.863</td>
<td>0.841</td>
<td>0.844</td>
<td>0.790</td>
<td>0.783</td>
<td>0.762</td>
</tr>
<tr>
<td>0.05</td>
<td>0.685</td>
<td>0.645</td>
<td>0.631</td>
<td>0.597</td>
<td>0.597</td>
<td>0.583</td>
</tr>
<tr>
<td>0.02</td>
<td>0.504</td>
<td>0.434</td>
<td>0.425</td>
<td>0.372</td>
<td>0.357</td>
<td>0.345</td>
</tr>
<tr>
<td>0.01</td>
<td>0.326</td>
<td>0.244</td>
<td>0.243</td>
<td>0.237</td>
<td>0.240</td>
<td>0.243</td>
</tr>
</tbody>
</table>

#### $\mu = 0.53(1, 1, 1, 1, 1, 1, 1, 1)$

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.05</td>
<td>1.000</td>
<td>0.998</td>
<td>0.997</td>
<td>0.979</td>
<td>0.973</td>
<td>0.951</td>
</tr>
<tr>
<td>0.02</td>
<td>0.987</td>
<td>0.948</td>
<td>0.930</td>
<td>0.803</td>
<td>0.751</td>
<td>0.681</td>
</tr>
<tr>
<td>0.01</td>
<td>0.857</td>
<td>0.656</td>
<td>0.604</td>
<td>0.497</td>
<td>0.480</td>
<td>0.442</td>
</tr>
</tbody>
</table>
### APPENDIX C. BEL POWER COMPARISONS

\[ \mu = -0.53(1, 1, 1, 1, 1, 1, 1) \]

<table>
<thead>
<tr>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.990</td>
<td>0.992</td>
<td>0.991</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>0.05</td>
<td>0.964</td>
<td>0.975</td>
<td>0.974</td>
<td>0.980</td>
<td>0.981</td>
</tr>
<tr>
<td>0.02</td>
<td>0.913</td>
<td>0.926</td>
<td>0.923</td>
<td>0.914</td>
<td>0.907</td>
</tr>
<tr>
<td>0.01</td>
<td>0.826</td>
<td>0.835</td>
<td>0.831</td>
<td>0.836</td>
<td>0.839</td>
</tr>
</tbody>
</table>

\[ \mu = 1.5(1, 0, 0, 0, 0, 0, 0) \]

<table>
<thead>
<tr>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.05</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.992</td>
<td>0.988</td>
</tr>
<tr>
<td>0.02</td>
<td>1.000</td>
<td>0.998</td>
<td>0.987</td>
<td>0.925</td>
<td>0.902</td>
</tr>
<tr>
<td>0.01</td>
<td>1.000</td>
<td>0.955</td>
<td>0.882</td>
<td>0.780</td>
<td>0.763</td>
</tr>
</tbody>
</table>

\[ \mu = -1.5(1, 0, 0, 0, 0, 0, 0) \]

<table>
<thead>
<tr>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.854</td>
<td>0.872</td>
<td>0.869</td>
<td>0.894</td>
<td>0.902</td>
</tr>
<tr>
<td>0.05</td>
<td>0.752</td>
<td>0.793</td>
<td>0.775</td>
<td>0.824</td>
<td>0.834</td>
</tr>
<tr>
<td>0.02</td>
<td>0.650</td>
<td>0.707</td>
<td>0.680</td>
<td>0.706</td>
<td>0.710</td>
</tr>
<tr>
<td>0.01</td>
<td>0.514</td>
<td>0.569</td>
<td>0.566</td>
<td>0.617</td>
<td>0.629</td>
</tr>
</tbody>
</table>

\[ \mu = 1.06(1, -1, 0, 0, 0, 0, 0) \]

<table>
<thead>
<tr>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.996</td>
</tr>
<tr>
<td>0.05</td>
<td>1.000</td>
<td>0.999</td>
<td>0.995</td>
<td>0.976</td>
<td>0.971</td>
</tr>
<tr>
<td>0.02</td>
<td>1.000</td>
<td>0.969</td>
<td>0.948</td>
<td>0.867</td>
<td>0.846</td>
</tr>
<tr>
<td>0.01</td>
<td>1.000</td>
<td>0.817</td>
<td>0.798</td>
<td>0.728</td>
<td>0.719</td>
</tr>
</tbody>
</table>

\[ \mu = 0.53(1, 1, 1, -1, -1, -1, -1) \]

<table>
<thead>
<tr>
<th>EL</th>
<th>AEL</th>
<th>BEL(1.0)</th>
<th>BEL(1.9)</th>
<th>BEL(2.5)</th>
<th>Hotelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.995</td>
<td>0.996</td>
</tr>
<tr>
<td>0.05</td>
<td>0.993</td>
<td>0.985</td>
<td>0.985</td>
<td>0.971</td>
<td>0.966</td>
</tr>
<tr>
<td>0.02</td>
<td>0.971</td>
<td>0.929</td>
<td>0.921</td>
<td>0.860</td>
<td>0.837</td>
</tr>
<tr>
<td>0.01</td>
<td>0.884</td>
<td>0.762</td>
<td>0.745</td>
<td>0.713</td>
<td>0.711</td>
</tr>
</tbody>
</table>
Bibliography


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


Welch, B. L. (1938), ‘The significance of the difference between two means when the population variances are unequal’, *Biometrika* **29**(3-4), 350–362.
