INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
RICE UNIVERSITY

Robust Empirical Likelihood

by

Nancy Glenn

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

APPROVED. THESIS COMMITTEE:

Dr. David Scott, Chairman
Professor of Statistics

Dr. Yin Zhang, Co-Chairman
Associate Professor of Computational and
Applied Mathematics

Dr. Katherine Ensrud
Professor of Statistics

Dr. Nathaniel Dean
Associate Professor of Computational and
Applied Mathematics

Houston, Texas
April, 2002
Abstract

Robust Empirical Likelihood

by

Nancy Glenn

This research introduces a new nonparametric technique: robust empirical likelihood. Robust empirical likelihood employs the empirical likelihood method to compute robust parameter estimates and confidence intervals. The technique uses constrained optimization to solve a robust version of the empirical likelihood function, thus allowing data analysts to estimate parameters accurately despite any potential contamination.

Empirical likelihood combines the utility of a parametric likelihood with the flexibility of a nonparametric method. Parametric likelihoods are valuable because they have a wide variety of uses; in particular, they are used to construct confidence intervals. Nonparametric methods are flexible because they produce accurate results without requiring knowledge about the data's distribution. Robust empirical likelihood's applications include regression models, hypothesis testing, and all areas that use likelihood methods.
Contents

1 Introduction .................................................. 1
   1.1 Statement of Problem .................................. 2
   1.2 Robust Empirical Likelihood ......................... 2

2 Link Between Empirical Likelihood and Robust Empirical Likelihood 4
   2.1 Empirical Likelihood History and Development ..... 5
   2.2 One-Sample Empirical Likelihood .................... 7
   2.3 Tilting .................................................. 12

3 Robust Empirical Likelihood Point Estimators 14
   3.1 Robust Empirical Likelihood Definition ............. 14
   3.2 Point Estimator Derivation .......................... 15

4 Theoretical Analysis ...................................... 17
   4.1 KKT Conditions ....................................... 17
   4.2 KKT and Univariate Point Estimators ................ 18
   4.3 KKT and Multivariate Point Estimators .............. 22

5 Robust Empirical Likelihood Curve 24
   5.1 Scale Estimation ...................................... 24
   5.2 The Curve ............................................. 29
5.3 Sensitivity Analysis ........................................... 32

6 Multivariate Extensions ..................................... 35
   6.1 Bivariate Nonlinear Programming Problem .......... 35
   6.2 Bivariate Point Estimators ............................. 36

7 Confidence Intervals ....................................... 40
   7.1 Distributional Properties ............................... 40
   7.2 Bootstrap Confidence Intervals ....................... 42
   7.3 Sequential Linearization ............................... 44

8 Conclusions and Future Work ............................. 46
List of Figures

2.1 Plot of $\gamma$ vs. the product of the elements of $p_\gamma$ ................. 9
2.2 Empirical likelihood ratio curve for the mean ...................... 10
2.3 Empirical likelihood ratio curve for the mean and corresponding confidence intervals .................................................. 12

5.1 Plot of $p_i X_i$ vs. Deviations ................................. 28
5.2 Plot of Data I vs. Objective Function .......................... 31

7.1 $Q$–$Q$ plot of bootstrap resampled values of $-2 \log R(\mu_0)$ plotted against $\chi^2(1)$ quantiles ..................................................... 41
7.2 95% confidence interval for the mean of Data Set I ............... 43
Chapter 1

Introduction

This dissertation introduces a new nonparametric procedure, robust empirical likelihood, that improves upon the empirical likelihood (EL) method of confidence interval construction; the goal is to construct robust EL confidence intervals and to provide multidimensional robust point estimators. This allows data analysts to accurately estimate parameters despite any potential contamination.

There are two main robustification approaches. One approach modifies the likelihood function, the other uses a robust instead of nonrobust statistic [11]. Robust EL uses the first approach in the following way: robust EL modifies the EL function by introducing a weight vector, \( \mathbf{w} \), that differently weighs the contribution from \( X_i \) to the log-EL function.

This modification incorporates a nonlinear programming problem that is solved using MATLAB's optimization toolbox. The solution is a length-\( n \) vector, \( \mathbf{p} \), whose elements represent the probability mass placed on \( X_i \). After obtaining \( \mathbf{p} \), I choose the best weight vector, \( \mathbf{w} \). This produces a point estimate, which is a single value. I next obtain a confidence interval.
1.1 Statement of Problem

Because empirical likelihood confidence intervals for the mean extend in the direction that the sample is skewed, using contaminated data skews the interval in the direction of outliers. The major implication of this problem is that coverage probabilities of interval estimators may be incorrect.

Constructing confidence regions in the midst of contaminated data is especially a problem with multivariate data. Measuring distance in multidimensions is not a trivial task. Hence, outlier identification is not trivial.

This research downweights outliers so that they have less influence on the log-EL function than data values that are representative of the true distribution. The following section describes a solution to this problem.

1.2 Robust Empirical Likelihood

As noted in Section 1.1, the EL method of confidence interval construction is not robust when the data are contaminated. The solution, robust empirical likelihood, extends Choi, Hall and Presnell's [1] robust parametric method, described in Chapter 2, to the nonparametric setting. This involves "tilting" the log-empirical likelihood function so that outliers do not have as much influence on the log-likelihood function.

In Choi, Hall and Presnell's robust parametric procedure, the authors weigh by $w_i$ the contribution of $X_i$ to the log-likelihood. Since robust EL is a nonparametric procedure, I weigh the contribution from $X_i$ to the log-empirical likelihood. Weighing the contribution of $X_i$ amounts to starting with a discrete uniform distribution on $n$ points and tilting it to achieve a desired empirical log-likelihood. The resulting tilted distribution is a multinomial distribution on $n$ points. In simulations, $w_i$ is smaller when $X_i$ is an outlier.

This dissertation centers around obtaining robustness through tilting the log-EL function. Tilting and EL are discussed in Chapter 2. Robust EL point estimators
are presented in Chapter 3. Chapters 4 describes the underlying theoretical analysis. Chapters 5 constructs the robust empirical likelihood curve. Chapter 6 considers multivariate extensions. Chapter 7 gives distributional properties then constructs confidence intervals. In the concluding chapter, I discuss future directions for robust EL.
Chapter 2

Link Between Empirical Likelihood and Robust Empirical Likelihood

The empirical likelihood ratio is a nonparametric likelihood ratio. The distinction between EL and the parametric likelihood ratio is that EL does not require knowledge about the data’s distribution. Hence, EL has a wider range of applications than its parametric counterpart.

Robust empirical likelihood’s distinguishing attribute is that it is not as sensitive to contamination as empirical likelihood. The insensitivity of robust EL to outliers holds true even when the statistic is sensitive to outliers. Therefore, robust EL can solve a wider variety of problems than empirical likelihood.

Because EL is the foundation of robust EL, this chapter details the EL method of confidence interval construction. The chapter starts with an EL history and development section. Section 2.2 analyzes the intuitive approach to understanding the empirical likelihood method. Tilting, the topic of Section 2.3, is the link between EL and robust EL.
2.1 Empirical Likelihood History and Development


In 1938 Wilks [14] published a landmark paper in which he proved that $-2 \log(LR)$ is asymptotically $\chi^2(r)$, where $LR$ is the likelihood ratio and $r$ is the degrees of freedom. Wilks's result is a commonly used statistical fact; in particular, it shows that $-2 \log(LR)$ is a pivotal quantity, which is used to construct parametric confidence intervals and regions.

In 1975 Thomas and Grunkemeier [13] introduced the EL ratio method of confidence interval construction. They build upon Wilks's results and give heuristic arguments for nonparametric confidence intervals. Their extension of Wilks's results to nonparametric data is quite an advancement because nonparametric procedures require fewer distributional assumptions than parametric procedures do. Thomas and Grunkemeier focus on survival probabilities for censored data; however, their 1975 article's most important result is that it gives a heuristic argument for the limiting distribution of the $-2 \log(ELR)$ statistic for the case $r = 1$. Therefore, $-2 \log(ELR)$ is also a pivotal quantity.

In 1988 Owen [9] rigorously proved Thomas and Grunkemeier's argument for any value of $r$. Owen's contribution allows one to nonparametrically compute threshold values for constructing a $100(1 - \alpha)\%$ confidence interval based on limiting values of the $\chi^2(r)$ distribution. Hence, Owen shows that his nonparametric result for EL is comparable to Wilks's parametric result for the likelihood ratio.

DiCiccio, Hall and Romano [4] gave a detailed comparison of Wilks's parametric LR and Owen's EL ratio functions in 1989. They assert that EL approximates the parametric likelihood ratio when a parametric likelihood is based on a least favorable family, but show this approximation solely for the mean. Their work shows that parametric and nonparametric likelihood ratios are approximately the same for the

In 1990, Owen [10] gave an argument for multivariate generalizations to his 1988 findings. Multivariate extensions, contours and regions are important because they have many real-life applications. Owen does not theoretically prove that EL contours approximate parametric likelihood contours; however, in the same issue of *The Annals of Statistics*, Hall [6] rigorously proves that EL regions approximate pseudo-likelihood regions. This proof helps to fill the gap left by Owen [10] since “pseudo” indicates that the likelihood is based on some function of the data instead of the entire data set. One of the objectives of this dissertation is to show that robust empirical likelihood also has multivariate extensions.

A future objective of this research is to show that robust empirical likelihood is Bartlett correctable. Bartlett correction is a rescaling procedure which had previously been reserved for parametric likelihood ratios. In 1990, Hall and La Scala [7] showed that EL is Bartlett correctable. Bartlett correction reduces the empirical likelihood coverage error from $O(n^{-1})$ to $O(n^{-2})$. This is a notable reduction, but the correction factor is difficult to compute. DiCiccio, Hall and Romano [3] present a general formula for computing the Bartlett correction and demonstrate its use with various statistics. If this general formula is applicable to robust EL, I will apply it to improve the robust empirical likelihood coverage error.

Robust EL improves EL by assigning unequal weights to the empirical log-likelihood function. This tilts the log-likelihood. The purpose of tilting is to reduce the effects of contamination by assigning smaller weights to outliers. Tilting is an improvement because weighing outliers and inliers equally can degrade the performance of statistical estimators.

This dissertation develops a new procedure that tilts away from outliers in order to render accurate confidence intervals and regions. It is a robust empirical likelihood.
2.2 One–Sample Empirical Likelihood

Since the robust empirical likelihood method is built upon the one–sample empirical likelihood method, understanding the one–sample empirical likelihood method is instrumental to understanding the robust empirical likelihood method. I have constructed an example that facilitates understanding by giving an intuitive approach to the one–sample empirical likelihood method for constructing confidence intervals. The example uses an approach that facilitates understanding, but is not computationally practical for more than three points. Following the example, I give a less intuitive approach. The second approach is computationally practical. It is the one used in practice. Both examples use the same data set to construct the one–sample empirical likelihood confidence interval for the mean. I now define empirical likelihood:

Definition 2.1 Suppose \( \{X_1, X_2, \ldots, X_n\} \) is a sample from an unknown distribution. Suppose \( p_i \) is the probability mass placed on \( X_i \). Let \( \theta(p) = \sum_{i=1}^{n} p_i X_i \) denote the distribution’s mean. The **Empirical Likelihood** for the parameter \( \theta \) evaluated at \( \theta_0 \) is defined as

\[
L(\theta_0) = \max_{p: \theta(p) = \theta_0} \prod_{i=1}^{n} p_i. \tag{2.1}
\]

The corresponding empirical log–likelihood ratio is

\[
l(\theta_0) = \log L(\theta_0) / L(\hat{\theta}) = \max_{p: \theta(p) = \theta_0} \sum_{i=1}^{n} \log(n p_i), \tag{2.2}
\]

where \( L(\hat{\theta}) = n^{-n} \).

Computing the empirical likelihood function is equivalent to computing the profile likelihood of a general multinomial distribution supported on the data. The following simple example gives an intuitive approach to understanding the above definition.

Example 2.1 Suppose that \( X = \{1, 3, 7\} \). Plot the empirical likelihood ratio for the mean, then plot the 90% and 95% EL confidence intervals for the sample mean, \( \bar{x} \).
Solution: With 3 data points, an equilateral triangle can be used to represent the probability vector, \( \mathbf{p} = (p_1, p_2, p_3) \), that corresponds to specific mean values. Since \( n = 3 \), \( \mathbf{p} = (p_1, p_2, p_3) \) can be represented in Baricentric coordinates. Each data point is associated with a vertex of an equilateral triangle. Without loss of generality, suppose that \( X_1 = 1 \) and \( X_2 = 3 \) are associated with the bottom left and right vertices, respectively, and that \( X_3 = 7 \) is associated with the top vertex. The \( p_i's, i = 1, 2, 3 \) represent probabilities and \( \mathbf{p} = (p_1, p_2, p_3) \) gives a multinomial distribution on the \( n = 3 \) points. Hence, the probability vectors that correspond to 1 and 3 are \( (1, 0, 0) \) and \( (0, 1, 0) \), respectively. For mean values other than \( \{1, 3, 7\} \) a convex combination determines probability vectors that correspond to specific mean values:

\[
\mathbf{p}_\gamma = \gamma \mathbf{p}_i + (1 - \gamma) \mathbf{p}_j, \tag{2.3}
\]

where \( \gamma \in [0, 1] \) and \( i, j \in \{1, 2, 3\}, i \neq j \).

Since EL is supported on the data, confidence regions do not extend beyond the convex hull of the data. Therefore, mean values that lie outside the convex hull of the data are not considered.

All probability vectors corresponding to specific mean values lie on lines that intersect distinct sides of the equilateral triangle. Figure 2.1 is a plot of the product of the elements of \( \mathbf{p}_\gamma \) for five mean values.

The plot displays integer values of the mean, but real numbers could have been used. The maximum value of each curve is the value of the empirical likelihood for that particular value of the mean. Hence, the empirical likelihood is a profile likelihood. The empirical likelihood ratio is the empirical likelihood divided by the largest possible value of the empirical likelihood, \( \left( \frac{1}{n} \right)^n \).

Example 2.1 describes the intuitive approach to finding the empirical likelihood value for the mean. The steps are summarized below:

1. Choose an interval of hypothesized mean values.
Figure 2.1: Plots of $\gamma \in [0,1]$ vs. the product of the elements of $p_\gamma$. The indicated maximums (filled points) are the values of the EL function for corresponding mean values, $\mu$.

2. For each mean value find several corresponding probability vectors, $p_\gamma$.

3. For each mean value, graph $\gamma \in [0,1]$ vs. the product of the elements of $p_\gamma$.

Graph this product for several values of $p_\gamma$; see Figure 2.1.

4. Find the maximum value of the curve for each mean value; see Figure 2.1.

5. Each optimal value found in the previous step is the EL for each corresponding mean value. The EL ratio function is the EL divided by the nonparametric maximum likelihood estimator (MLE) of the cumulative distribution function; see Figure 2.2.

This view is intuitive, but it is difficult to visualize when there are four or more data points because the picture extends from a triangle to a tetrahedron. The four data points correspond to the vertices of the tetrahedron. Instead of lines intersecting opposite sides, there are surfaces intersecting the faces of a tetrahedron. The profile,
Figure 2.2: Empirical likelihood ratio curve for the mean.

The empirical likelihood curve is the curve that intersects the maximum point of each surface. This approach is even more difficult to visualize with five or more data points.

Owen [9] gives a computationally feasible, but less intuitive approach for finding the EL value for the mean, in which he derives an expression $G$, that gives rise to the form of the optimal probability vectors:

$$G = \sum_{i=1}^{n} p_i X_i + \lambda_1 (1 - \sum_{i=1}^{n} p_i) + \lambda_2 \{\log c - \sum_{i=1}^{n} \log (n p_i)\}$$  \hspace{1cm} (2.4)

where

- $p_i$ is the probability mass placed on $X_i$,
- $\lambda_1$ and $\lambda_2$ are Lagrange multipliers,
- $c \in (0, 1)$, and
- $p_i$'s are constrained by: $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$, and $\prod_{i=1}^{n} n p_i \geq c$.

Taking the partial derivative of $G$ with respect to $p_i$ and setting this equal to 0 yields
the form of the elements of optimal probability vectors \[9\],

\[ p_i = \frac{\lambda_2}{X_i - \lambda_1}. \tag{2.5} \]

Figure 2.2 depicts the empirical likelihood ratio curve for \( X = \{1, 3, 7\} \).

My \textit{S}plus program that generates Figures 2.2 and 2.3 starts with a grid of Lagrange multiplier values. I compute the Lagrange multipliers from an ad hoc method devised by Owen \[9\]. Each pair of Lagrange multipliers yields corresponding probability vectors, which are constructed from (2.5). Each probability vector yields a corresponding mean value. Because of this construction, several Lagrange multiplier pairs may correspond to the same mean value. In particular, several pairs correspond to the mean value that maximizes the empirical likelihood ratio function. As can be shown by a Lagrange multiplier argument, the value that maximizes the empirical likelihood ratio function is always the sample mean, that is, \( p_i = 1/n, \forall i \). Since many Lagrange multiplier pairs correspond to the sample mean, there is a cluster of points around the sample mean in Figures 2.2 and 2.3.

The beauty of empirical likelihood is that it is a nonparametric method from which one can compute threshold values for constructing a \(100(1 - \alpha)\%\) confidence interval based on limiting values of a parametric distribution, the \( \chi^2 \) distribution. Since \(-2 \log(ELR)\) is asymptotically \( \chi^2(r) \), where \( r \) is the degrees of freedom, \(-2 \log(ELR)\) is the pivotal quantity used to construct EL confidence intervals. Figure 2.3 shows the ELR and corresponding confidence intervals for \( X = \{1, 3, 7\} \).

In a 90\% confidence interval, the .10 quantile is \( \chi^2_{.10}(1) = 2.706 \). This means that \(-2 \log(ELR) \approx 2.706\), and that \( ELR \approx .258 \) is the threshold value for constructing a 90\% confidence interval. Similarly, \( ELR \approx .147 \) is the threshold value for establishing a 95\% confidence interval.

Understanding the one–sample empirical likelihood method is the first step to understanding robust EL. The second step is tilting. Tilting is the ingredient that makes the robust EL method \textit{robust}. Section (2.3) details the tilting procedure.
Figure 2.3: Empirical likelihood ratio curve for the mean and corresponding confidence intervals. Equation 2.5 generates the points from a grid of Lagrange multiplier values.

### 2.3 Tilting

The term "tilting" refers to a nonuniform probability assignment on $n$ data points. It has a dual purpose: tilt away from outliers and tilt towards inliers. Tilting changes the value of the maximum likelihood estimator (MLE) of the contaminated distribution to a value that better represents the true distribution's parameter. When comparing parameter values, $\theta_1$ having a greater likelihood function than $\theta_2$ indicates that the observed sample is more likely to have occurred if $\theta_1$ were the true parameter. When working with robust EL parameter analysis is done by starting with a discrete uniform distribution and assigning a greater probability to parameters that have a greater likelihood. This probability assignment tilts the distribution.

In this dissertation, probability assignments are brought about by using a discrete uniform distribution as the initial weight vector in a nonlinear programming problem, then appropriately tilting the distribution. This process maximizes the robust EL
objective function subject to both the given weight vector and the constraints on the probability vector.

Choi, Hall and Presnell [1] empirically tilt the distribution in their robust parametric procedure. Choi, Hall and Presnell’s data are drawn from a distribution, $f(\cdot|\theta)$, that may be contaminated. They assume that $\mathbf{w} = (w_1, \ldots, w_n)$ is a multinomial distribution on $n$ points and define the log-likelihood as a function of $\theta$:

$$L = \sum_{i=1}^{n} w_i \log(f(x_i|\theta)).$$  \hspace{1cm} (2.6)

This formulation weighs by $w_i$ the contribution of $x_i$ to the log-likelihood. Therefore, $w_i$ is smaller when $x_i$ is an outlier.

This distribution tilting procedure relates to constructing confidence intervals in the following way. In general, a confidence interval for a parameter, $\theta$, contains the values of $\hat{\theta}$ that cannot be rejected by a hypothesis test. Similar to the nonparametric procedure weighted bootstrap [5], deciding which $\mathbf{w}$ distribution is most representative of the observed data answers the question of inclusion/exclusion of a $\hat{\theta}$ value. Minimizing the divergence distance from $\mathbf{w}$ to the uniform distribution on $n$ points gives the $\mathbf{w}$ distribution that is most representative of the observed data. A significance value is assigned to $\hat{\theta}$ based on this estimated distribution.
Chapter 3

Robust Empirical Likelihood Point Estimators

Constructing robust EL confidence intervals and regions begins with point estimators. The function from which one derives the point estimator is the robust EL function. This chapter defines the term robust empirical likelihood and then derives the point estimator.

3.1 Robust Empirical Likelihood Definition

Definition 3.1 Suppose \( \{X_1, X_2, \ldots, X_n\} \) is a sample from an unknown distribution that may be contaminated. Let \( \theta(p) = \sum_{i=1}^{n} p_i X_i \) denote the distribution’s mean. Given the weight vector, \( w \), the Robust Empirical Likelihood for the parameter \( \theta \) evaluated at \( \theta_0 \) is defined as

\[
L(\theta_0) = \max_{\theta(p) = \theta_0} \prod_{i=1}^{n} (p_i)^{n w_i}.
\]  

(3.1)
The corresponding robust empirical log–likelihood ratio is

\[ l(\theta_0) = \log L(\theta_0)/L(\hat{\theta}) = \max_{p: \hat{\theta}(p)=\theta_0} \sum_{i=1}^{n} n w_i \log(p_i/w_i). \tag{3.2} \]

where \( L(\hat{\theta}) = (w_i)^{-nw_i}. \)

By Lagrange multipliers, the form of the elements of the optimal probability vector is

\[ p_i^* = \frac{nw_i}{n - \lambda_2(\theta_0 - X_i)}. \tag{3.3} \]

See Chapter 4 for the derivation of Equation (3.3). Rewriting (3.2) gives

\[ l(\theta_0) = \sum_{i=1}^{n} n w_i \log(p_i^*/w_i). \tag{3.4} \]

Equation (3.4) is the robust empirical log–likelihood ratio function for the parameter \( \theta \) evaluated at \( \theta_0 \). The following section explains how to derive robust empirical likelihood point estimators.

### 3.2 Point Estimator Derivation

The procedure for deriving robust empirical likelihood point estimators for the mean, \( \mu \), is analogous to that in Section 2.2 for deriving non–robust empirical likelihood point estimators for \( \mu \). The robust EL procedure starts with solving the following nonlinear programming problem:

\[ \max_{p} \sum_{i=1}^{n} n w_i \log(p_i/w_i) \tag{3.5} \]
subject to

\[ \sum_{i=1}^{n} p_i X_i = \mu \quad (3.6) \]
\[ \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \quad (3.7) \]
given \[ \sum_{i=1}^{n} w_i = 1, \quad w_i > 0. \quad (3.8) \]

The above nonlinear programming problem yields optimal probability vectors, \( \mathbf{p} \), for robust empirical likelihood. Optimal probability vectors for robust EL generate robust EL curves in the same way that non-robust probability vectors generate EL curves. The optimal probability vectors derived from Equation (3.3) are analogous to the non-robust optimal probability vectors derived from Equation (2.5).

To formulate the objective function (3.5), first assume that the weight vector, \( \mathbf{w} \), is given. As with EL, maximize the product of the elements of the probability vector since the maximum indicates the most plausible parameter value. Because this product is small, results are numerically unstable. Therefore, I maximize the sum of the logarithms instead. Dividing this product by the MLE, \( (w_i)^{nw_i} \), yields the robust empirical log-likelihood ratio. At this stage, the objective function is

\[ \max_{\mathbf{p}} \sum_{i=1}^{n} w_i \log(p_i / w_i). \quad (3.9) \]

The final step for deriving the robust empirical likelihood objective function is to scale by the sample size, \( n \), so that robust EL and non-robust empirical likelihood curves are identical when \( w_i = 1/n \) and there are no outliers:

\[ \max_{\mathbf{p}} \sum_{i=1}^{n} nw_i \log(p_i / w_i). \quad (3.10) \]

The steps outlined in this paragraph yield the objective function in Equation (3.5).
Chapter 4

Theoretical Analysis

The robust EL method for confidence interval construction begins with a point estimator. I now present the underlying theoretical analysis for obtaining point estimators, which centers around solving a nonlinear programming problem. Because the robust EL objective function is a strictly concave function on a convex set of probability vectors, the Karush Kuhn Tucker (KKT) Theorem states that a unique global maximum exists. Section 4.1 describes the KKT conditions. Sections 4.2 and 4.3 use the KKT Theorem to derive univariate and multivariate robust EL point estimators.

4.1 KKT Conditions

The KKT conditions are first-order necessary conditions because they concern properties of the gradient of the objective function and constraints. Consider the following nonlinear minimization problem:

$$\max_x f(x)$$

subject to

$$c_i(x) = 0, \, \forall \, i \in \mathcal{E}$$

$$c_i(x) \geq 0, \, \forall \, i \in \mathcal{I}$$
where
\[ f(x), c_i(x) : \mathbb{R}^n \rightarrow \mathbb{R} \] (4.2)
are continuously differentiable. Suppose \( x^* \) is a local solution of (4.1). Let \( L(x, \lambda) \) be the Lagrangian,
\[ L(x, \lambda) = f(x) + \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x). \] (4.3)
Call \( c_i, i \in \mathcal{E} \) the equality constraints and \( c_i, i \in \mathcal{I} \) the inequality constraints. Then there is a Lagrange multiplier vector \( \lambda^* \) such that the KKT conditions are satisfied at \( (x^*, \lambda^*) \). Provided the linear independence constraint qualification holds at \( x^* \), the KKT conditions are [8]:

\[ \nabla_x L(x^*, \lambda^*) = 0 \] (4.4)

\[ c_i(x^*) = 0, \ \forall \ i \in \mathcal{E} \] (4.5)

\[ c_i(x^*) \geq 0, \ \forall \ i \in \mathcal{I} \] (4.6)

\[ \lambda_i^* \geq 0, \ \forall \ i \in \mathcal{I} \] (4.7)

\[ \lambda_i^* c_i(x^*) = 0, \ \forall \ i \in \mathcal{E} \cup \mathcal{I} \] (4.8)

Section 4.2 uses the KKT theorem to derive univariate point estimators.

### 4.2 KKT and Univariate Point Estimators

Obtaining univariate robust EL point estimators starts with solving the nonlinear programming problem given in (3.5). The first step in this application is to find the Lagrangian for (3.5), which is

\[ L(p, \lambda_1, \lambda_2 | w, x, \mu) = \sum_{i=1}^{n} nw_i \log(p_i/w_i) + \lambda_1 (1 - \sum_{i=1}^{n} p_i) + \lambda_2 (\mu - \sum_{i=1}^{n} X_i p_i). \] (4.9)
Equation (4.9) gives rise to the form of the optimal univariate robust EL probability vectors, as Equation (2.4) is an analogous equation for empirical likelihood. In the following paragraphs, I use Equation (4.9) to derive the form of the optimal robust EL probability vectors.

The optimality conditions for the constrained optimization problem (3.5) are the KKT conditions. Therefore, a first-order necessary condition is:

$$\nabla L(p^*, \lambda_1^*, \lambda_2^* | w, x, \mu) = 0. \quad (4.10)$$

In order to derive the form of the optimal probability vectors for robust EL, first derive an expression for the Lagrange multiplier, $\lambda_1$:

Evaluating (4.10) yields

$$\frac{\partial L}{\partial p_k} = \frac{n w_k}{p_k} - \lambda_1 - \lambda_2 X_k = 0, \quad k = 1, 2, \cdots, n. \quad (4.11)$$

From equation (4.11),

$$n w_k - \lambda_1 p_k - \lambda_2 X_k p_k = 0. \quad (4.12)$$

Since both $w$ and $p$ satisfy the axioms of probability vectors, summing each term in (4.12) gives $n \sum_{i=1}^{n} w_i - \lambda_1 - \lambda_2 \mu = 0$. Therefore,

$$\lambda_1 = n \sum_{i=1}^{n} w_i - \lambda_2 \mu \quad (4.13)$$

is an expression for $\lambda_1$.

From this $\lambda_1$ expression one can derive the form of the elements of the optimal probability vectors for robust EL. Solving (4.12) for $w_k$ yields $n w_k = \lambda_1 p_k + \lambda_2 X_k p_k$. Substituting (4.13) for $\lambda_1$ implies that $n w_k = (n - \lambda_2 \mu) p_k + \lambda_2 X_k p_k$. Therefore,

$$p_k = \frac{n w_k}{n \sum_{i=1}^{n} w_i - \lambda_2 (\mu - X_k)} \quad (4.14)$$
is the form of the elements of the optimal probability vector for robust EL.

To evaluate (4.14), provide values for \( w_k, \lambda_2, \) and \( \mu \). Let \( \mathbf{w} = (\frac{1}{n}, \ldots, \frac{1}{n}) \) be the initial weight vector. Iteratively reweight the EL function to find an appropriate \( \mathbf{w} \) that downweights outliers. Reweighing tilts the EL likelihood function, as described in Chapter 2, so that parameters with a greater probability have a greater likelihood. Values for the parameter \( \mu \) lie within the convex hull of the data. That is, robust EL is supported on the data. Therefore, \( \mu \) must be in the interval \((X_{(1)}, X_{(n)})\) in order to satisfy the constraint \( \sum_{i=1}^{n} p_i X_i = \mu \), where \( p_i \geq 0 \). To find \( \mu^* \) given \( \mathbf{w} \) note that

\[
REL(\mu|\mathbf{w}, \mathbf{x}) = \sum_{i=1}^{n} nw_i \log \left( \frac{p^*_i(\mu)}{w_i} \right). \tag{4.15}
\]

where \( p^*_i(\mu) \) is \( i \)th element of \( \mathbf{p} \); it satisfies the constraints \( \sum_{i=1}^{n} p_i = 1 \) and \( \sum_{i=1}^{n} p_i X_i = \mu \) for a fixed \( \mu \). So,

\[
\mu^* = \arg \max_{\mu} REL(\mu|\mathbf{w}, \mathbf{x}). \tag{4.16}
\]

The Lagrange multiplier, \( \lambda_2 \), is chosen such that it satisfies the constraints

\[
\sum_{i=1}^{n} p_i = 1 \tag{4.17}
\]

and

\[
\sum_{i=1}^{n} p_i X_i = \mu. \tag{4.18}
\]

Suppose \( p_k \) is defined as in (4.14). Let \( \lambda_2^* \) be a Lagrange multiplier value that satisfies the constraints (4.17) and (4.18). The first of three steps for choosing \( \lambda_2^* \) is to recall (4.14), the form of the elements of robust EL's optimal probability vector. Define the function

\[
f(\lambda_2) = \sum_{k=1}^{n} \frac{n w_k}{n - \lambda_2 (\mu - X_k)} - 1. \tag{4.19}
\]

The Lagrange multiplier \( \lambda_2^* \), is the value that satisfies the constraints (4.17) and
(4.18). Hence, \( f(\lambda_2^*) = 0 \). In order to find the root \( \lambda_2^* \) of \( f(\lambda_2) \) examine the function

\[
f'(\lambda_2) = \sum_{k=1}^{n} \frac{n w_k (\mu - X_k)}{[n - \lambda_2 (\mu - X_k)]^2}
\]

(4.20)
evaluated at \( \lambda_2 = 0 \). This separates into three cases. All three cases use the fact that \( \lambda^* = 0 \) is always a solution. Since cases II and III have two roots, these cases also have a root that is either negative or positive.

**case I:** \( f'(0) = 0 \)

In this case, (4.19) has one root. Therefore, \( \lambda_2^* = 0 \).

**case II:** \( f'(0) > 0 \)

This implies that \( \sum_{k=1}^{n} w_k (\mu - X_k) > 0 \). Therefore, \( \lambda_2^* < 0 \).

**case III:** \( f'(0) < 0 \)

This implies that \( \sum_{k=1}^{n} w_k (\mu - X_k) < 0 \). Therefore, \( \lambda_2^* > 0 \).

From the two previous steps, one can infer whether \( \lambda_2^* \) is positive, negative, or zero. The third and final step for choosing \( \lambda_2^* \) is to find its exact value. I applied the bisection method to:

\[
\sum_{k=1}^{n} w_k \left( \frac{n}{n - \lambda_2 (\mu - X_k)} \right) - 1 = 0,
\]

(4.21)

which uses the lower (\( \lambda_- \)) and upper (\( \lambda_+ \)) bounds of \( \lambda_2^* \). To derive the bounds, notice that

- \( \mu > X_k \) implies \( \lambda_2 < \frac{1}{\mu - X_k} \)
- \( \mu < X_k \) implies \( \lambda_2 > \frac{1}{\mu - X_k} \)
- \( \mu = X_k \) implies \( \lambda_2 \) unrestricted.

The bounds for \( \lambda \) are \( \lambda_- = \frac{1}{\mu - \lambda(X)} \) and \( \lambda_+ = \frac{1}{\mu - \lambda(X)} \). Therefore, \( \lambda_2^* \) is the value of \( \lambda_2 \) that lies in the interval (\( \lambda_-^*, \lambda_+^* \)) and allows the constraints (4.17) and (4.18) to be
satisfied. The above argument focuses on (4.17). However, when (4.17) is satisfied, (4.18) is implicitly satisfied.

The above method for finding Lagrange multipliers is intuitive, however, it is difficult to find the form of Lagrange multipliers in bivariate and d-variate cases. In these instances, I employ Matlab’s optimization toolbox.

### 4.3 KKT and Multivariate Point Estimators

Chapter 6 presents the bivariate robust EL problem. The robust EL trivariate non-linear programming problem for the mean vector, $\mu$, is:

$$\max \sum_{i=1}^{n} n w_i \log(p_i/w_i)$$  \hspace{1cm} (4.22)

subject to

$$\sum_{i=1}^{n} p_i X_{i1} = \mu_1$$
$$\sum_{i=1}^{n} p_i X_{i2} = \mu_2$$
$$\sum_{i=1}^{n} p_i X_{i3} = \mu_3$$
$$\sum_{i=1}^{n} p_i = 1, p_i \geq 0$$
$$\text{given} \sum_{i=1}^{n} w_i = 1, w_i > 0.$$  

The $p$ that maximizes the objective function subject to the given constraints is the optimal trivariate robust EL probability vector. The corresponding Lagrangian is:
\[ L(p, \lambda_1, \lambda_2 \mid w, x, \mu) = \sum_{k=1}^{n} nw_k \log(p_k/w_k) + \lambda_1(1 - \sum_{k=1}^{n} p_k) + \lambda_2(\mu_1 - \sum_{k=1}^{n} X_{k1}p_k) + \lambda_3(\mu_2 - \sum_{k=1}^{n} X_{k2}p_k) + \lambda_4(\mu_3 - \sum_{k=1}^{n} X_{k3}p_k). \] (4.23)

An induction proof combined with an argument analogous to the Lagrange multiplier derivation in Section (4.2), proves that the elements of the \(d\)-variate optimal probability vectors have the form:

\[ p_k = \frac{nw_k}{n - \lambda_2(\mu_1 - X_{k1}) - \lambda_3(\mu_2 - X_{k2}) - \ldots - \lambda_{d+1}(\mu_d - X_{kd})}. \] (4.24)

The optimization problems in this chapter assume that the weight vector, \(w\), is given. The next logical question focuses on the choice of \(w\). It is the weight vector that differentiates robust EL from EL. Each element, \(w_i\), of the weight vector measures the contribution from \(X_i\) to the empirical log–likelihood function. I assume that \(w_i\)'s associated with inliers are equal. Then use the absolute deviation to determine an appropriate \(w\).
Chapter 5

Robust Empirical Likelihood Curve

Constructing robust EL curves is a two-part problem: find $w$ then construct the curve. Section 5.1 uses scale estimation to find $w$. Section 5.2 uses the optimization procedure discussed in Chapter 4 to build the robust EL curve and to find the optimal mean values. Section 5.3 discusses ways use sensitivity analysis to find $w$.

5.1 Scale Estimation

A scale estimator is a statistic that is equivariant under scale transformations. That is, if the parameter is transformed then the answer is also transformed. The scale estimate I use to determine an appropriate $w$ is the absolute deviation,

$$AD_n = |p_i X_i - M_n|, \quad (5.1)$$

where $M_n = \text{median}\{p_i X_i\}$. I use $AD_n$ because it identifies the outlier, and it signals when the outlier has been downweighted to a point where it does not greatly influence the optimal mean. The procedure is as follows:

- Let $w_i = 1/n$, $i = 1, \ldots, n$.
- Find $p_i^*$. 

24
• Compute $p_i^*X_i$.

• Find $AD_n$.

• The largest $AD_n$ corresponds to the outlier.

• Downweight the $w_i$ corresponding to the outlier until the outlier’s deviation lies within the range of the inliers’ deviations.

That is, downweight the outlier so that the distance between the outlier’s absolute deviation and the center of the absolute deviations is no greater than the distance between any inlier’s absolute deviation and the center. The center of the deviations is zero. To downweight the outlier, define a new weight vector:

$$w_{new} = w_b + cw_k,$$  \hspace{1cm} (5.2)

where

• $w_{new}$ and $w_b$ are weight vectors

• $w_b = (\frac{1}{n}, \ldots, \frac{1}{n})$

• $c$ is a scalar that moves the outlier’s deviation within the range of inliers’ deviations.

• $d_k$ is the vector of unit length that downweights $k$ outliers. Proposition 5.1 explains how to compute $d_k$.

Proposition 5.1 If $X = \{X_1, X_2, \ldots, X_n\}$ is an independent sample of size $n$, then $d_k$ is the vector of unit length defined as

$$d_k = \left(\frac{1}{\sqrt{\frac{k}{n-k}}}\right)y,$$ \hspace{1cm} (5.3)

where $y$ is the vector defined as
\[
\begin{pmatrix}
\frac{k}{n-k} \\
\vdots \\
\frac{k}{n-k} \\
-1 \\
\vdots \\
-1
\end{pmatrix}
\]

The rank ordered vector \( \mathbf{y} \) contains \( n-k \) elements equal to \( \frac{k}{n-k} \) and \( k \) elements equal to \(-1\).

**Proof**

By definition of unit vector, \( \mathbf{d}_k = \left( \frac{1}{\|\mathbf{y}\|} \right) \mathbf{y} \). Since \( \|\mathbf{y}\| = \sqrt{\frac{kn}{n-k}} \),

\[
\mathbf{d}_k = \frac{1}{\sqrt{\frac{kn}{n-k}}} \begin{pmatrix}
\frac{k}{n-k} \\
\vdots \\
\frac{k}{n-k} \\
-1 \\
\vdots \\
-1
\end{pmatrix}
\]

This completes the proof.

The scalar \( c \) is defined such that the outlier receives a weight that is less than or equal to the weight the inliers receive. When there is one outlier, \( 0 \leq c < \frac{1}{n} \sqrt{\frac{n}{n-1}} \). Proposition 5.2 confirms this.

**Proposition 5.2** Suppose \( \mathbf{X} = \{X_1, X_2, ..., X_n\} \) is an independent sample of size \( n \). Without loss of generality, assume \( \mathbf{X} \) contains one outlier. If the probability vector \( \mathbf{w}_{\text{new}} \) is defined as in Equation (5.2) then

\[
0 \leq c < \frac{1}{n} \sqrt{\frac{n}{n-1}}. \quad (5.4)
\]
Proof:

Without loss of generality, assume that \( X_n \) is an outlier. From Equation (5.2),

\[
\mathbf{w}_{new} = \left( \begin{array}{c} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{array} \right) + \frac{c}{\sqrt{\frac{n}{n-1}}} \left( \begin{array}{c} \frac{1}{n-1} \\ \vdots \\ \frac{1}{n-1} \\ -1 \end{array} \right). 
\]

Therefore,

\[
\mathbf{w}_{new}[n] = \frac{1}{n} + \frac{-c}{\sqrt{\frac{n}{n-1}}}
\]

is the \( \mathbf{w}_{new} \) element that corresponds to \( X_n \). In order for the outlier to receive a weight less than or equal to an inlier’s weight, the following must hold:

\[
0 < \mathbf{w}_{new}[n] \leq \frac{1}{n}.
\]

Therefore,

\[
0 < \frac{1}{n} + \frac{-c}{\sqrt{\frac{n}{n-1}}} \leq \frac{1}{n}.
\]

Solving (5.8) for \( c \) yields,

\[
0 \leq c < \frac{1}{n} \sqrt{\frac{n}{n-1}}.
\]

This completes the proof.

Example 5.1 Data Set I contains 19 \( N(0,1) \) independent values and 1 \( N(6,1) \) outlier. The symbol \( N(\mu, \sigma^2) \) stands for normally distributed, with mean \( \mu \) and variance \( \sigma^2 \). Use Equation (5.1) to find an appropriate weight vector. Find a robust EL point estimate for the mean.
Answer: Begin by finding an appropriate weight vector, \( w \). This is best illustrated graphically. The two frames in Figure 5.2 are \( p_i X_i \) vs. Deviations. The first

\[
\begin{array}{cccccc}
-0.0198 & -0.1567 & -1.6041 & 0.2573 \\
-1.0565 & 1.4151 & -0.8051 & 0.5287 \\
0.2193 & -0.9219 & -2.1707 & -0.0592 \\
-1.0106 & 0.6145 & 0.5077 & 1.6924 \\
0.5913 & -0.6436 & 0.3803 & -4.9909 \\
\end{array}
\]

Figure 5.1: Plots of \( p_i X_i \) vs. Deviations. Frame 1 indicates that the outlier's deviation does not lie within the range of the inlier's deviation when \( c = 0 \). Frame 2 indicates that the outlier's deviation lies within the range of the inlier's deviation when \( c = 0.285 \). The scalar \( c \) downweights the outlier.

frame represents \( w \) as defined by \( c = 0 \). The outlier is in the upper right quadrant. Increasing \( c \) decreases the outlier's deviation from \( p_i X_i \). In the last frame, \( c = 0.285 \).
At this point, the outlier's deviation lies within the range of the inlier's deviation. Therefore, \( c = .0285 \) defines an appropriate weight vector for calculating the robust EL point estimate for the mean. Note that

- \( 0 \leq c < .0513 \)
- \( \mu = 0 \) is true mean
- \( \bar{x} = .1375 \)
- \( \bar{x}_{\text{inliers}} = -.1180 \)
- \( \mu^* = -0.0044 \).

Figure 5.2 indicates that an appropriate choice for \( w \) corresponds to \( c = .0285 \). This implies that \( w_i = .0515 \) for inliers, and \( w_i = .0222 \) for the outlier. Therefore, a robust EL point estimate for the mean is \( \mu^* = -0.0044 \), which lies between the mean of the inliers and the mean of the entire data set.

### 5.2 The Curve

Constructing robust EL curves and finding \( \mu^* \) can be viewed as an optimization problem within an optimization problem. Since the curve is supported on the data, \( X_{(1)} < \mu < X_{(n)} \). There is a corresponding optimum probability vector for each \( \mu \) in the support. Therefore, the inner optimization problem is

\[
\max_{p} \sum_{i=1}^{n} n_w \log\left(\frac{p_i}{w_i}\right)
\]  

(5.10)
subject to
\[ \sum_{i=1}^{n} p_i X_i = \mu \quad (5.11) \]
\[ \sum_{i=1}^{n} p_i = 1, \, p_i \geq 0, \quad (5.12) \]
\[ \text{given } \sum_{i=1}^{n} w_i = 1, \, w_i > 0. \quad (5.13) \]

The outer optimization problem is to find \( \mu^* \), the value of \( \mu \) that maximizes the robust EL curve:
\[ \max_{p} \sum_{i=1}^{n} n w_i \log(p_i/w_i) \quad (5.14) \]

subject to
\[ X_{(1)} < \sum_{i=1}^{n} p_i X_i < X_{(n)} \quad (5.15) \]
\[ \sum_{i=1}^{n} p_i = 1, \, p_i \geq 0, \quad (5.16) \]
\[ \text{given } \sum_{i=1}^{n} w_i = 1, \, w_i > 0. \quad (5.17) \]

The outer optimization problem does not require nonlinear programming techniques to find \( p^* \). Proposition 5.3 shows that when \( \mu = \mu^* \), \( p^* \) is simply a function of the weight vector.

**Proposition 5.3** Assume \( \mu^* \) is defined as in Equation (4.16). If \( \mu = \mu^* \) then
\[ p_i^* = \frac{w_i}{\sum_{k=1}^{n} w_k}. \]
Figure 5.2: Plot of Data I vs. Objective Function. Plot indicates that $\mu^*$ for robust EL is closer to the mean of the inliers.

**Proof:**

Since $p_i^*$ and $\lambda_2$ are functions of $\mu$,

$$p_i^*(\mu) = \frac{n w_i}{n \sum_{k=1}^{n} w_k - \lambda_2(\mu)(\mu - X_i)}.$$  \hspace{1cm} (5.18)

From Equation (3.4),

$$l(\mu) = \sum_{i=1}^{n} n w_i \log\left(\frac{p_i^*(\mu)}{w_i}\right).$$ \hspace{1cm} (5.19)
Taking the derivative of $l(\mu)$ with respect to $\mu$ yields:

\[
l'(\mu) = \sum_{i=1}^{n} n w_i \frac{(p_i^*(\mu))'}{p_i^*(\mu)}
\]

\[
= \sum_{i=1}^{n} n w_i \frac{\lambda_2'(\mu) + \lambda_2''(\mu)(\mu - X_i)}{n \sum_{k=1}^{n} w_k - \lambda_2(\mu - X_i)} 
\]

\[
= \sum_{i=1}^{n} p_i^*(\mu)(\lambda_2'(\mu) + \lambda_2''(\mu)(\mu - X_i))
\]

\[
= \lambda_2'(\mu) + \left( \sum_{i=1}^{n} p_i^*(\mu) (\mu - X_i) \right) \lambda_2''(\mu)
\]

\[
= \lambda_2(\mu).
\]  \hspace{1cm} (5.23)

\[
= \lambda_2(\mu).
\]  \hspace{1cm} (5.24)

Because $\sum_{i=1}^{n} p_i^*(\mu)(\mu - X_i) = 0$, equation (5.23) simplifies to Equation (5.24). Since $\mu^*$ is a stationary point, $l'(\mu^*) = 0$. This implies that $\lambda_2(\mu^*) = 0$. Therefore,

\[
p_i^*(\mu) = \frac{w_i}{\sum_{k=1}^{n} w_k}.
\]  \hspace{1cm} (5.25)

This completes the proof.

### 5.3 Sensitivity Analysis

Section 5.1 uses scale estimation to find an appropriate weight vector, $w$. An alternative to scale estimation is sensitivity analysis. This section shows how to apply Proposition 5.3 and use post–optimal sensitivity analysis to find an appropriate weight vector.

Post–optimal sensitivity analysis measures the sensitivity of the optimum to small perturbations in parameters. My application is to measure the sensitivity of the optimal mean value, $\mu^*$, to infinitesimal changes in weight vector values. After making infinitesimal perturbations to the weight vector, I observe how the optimum value of the mean changes as each $X_i$ is reweighted. I start with $w_i = \frac{1}{n}$, $i \in \{1, \ldots, n\}$ then
downweight each outlier one at a time while maintaining the constraints \( \sum_{i=1}^{n} w_i = 1 \) and \( w_i > 0 \). The optimal, \( \mu^* \), is most sensitive when evaluated at outliers. Downweighting outliers reduces this sensitivity. Downweighting moves the sample mean of the entire data set closer to the sample mean of the inliers: this causes the empirical likelihood estimator to be robust. A derivation for the sensitivities follows. From Equation (5.11), \( \mu = \sum_{k=1}^{n} p_k X_k \). Recall that

\[
p_k^* = \frac{w_k}{\sum_{i=1}^{n} w_i}.
\]

(5.26)

When \( \sum_{i=1}^{n} w_i = 1 \), \( p_k^* = w_k \). From Proposition (5.3),

\[
\mu^* = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}.
\]

(5.27)

Therefore, the sensitivity is the gradient:

\[
\Delta \mu^* = \frac{X}{\sum_{i=1}^{n} w_i} - \frac{\sum_{i=1}^{n} w_i X_i}{(\sum_{i=1}^{n} w_i)^2} e
\]

(5.28)

\[
= \frac{X - \mu^* e}{\sum_{i=1}^{n} w_i},
\]

(5.29)

where \( e = (1, 1, \ldots, 1)' \).

Now that the sensitivity is formulated, I can choose \( w \). Possible ways of choosing \( w \) involves solving the two optimization problems:

\[
\min_{w} \sum_{i=1}^{n} |X_i - \sum_{j=1}^{n} w_j X_j|
\]

(5.30)

subject to

\[
\sum_{i=1}^{n} w_i = 1
\]

\[
0 \leq w_i \leq 0.
\]
In Equation (5.30), the objective function is the 1–norm of the sensitivity vector found in (5.29): this is true for $\sum_{k=1}^{n} w_k = 1$. This implies that $\sum_{j=1}^{n} w_j X_j = \text{median}(X)$, if $n$ is odd. When $n$ is even, the value is close to the median. It can be shown that $w^*$ is not unique. One way to find a unique solution is to solve following problem:

$$\min_{w} \sum_{i=1}^{n} |w_i - 1/n|$$  \hspace{1cm} (5.31)

subject to

$$\sum_{i=1}^{n} w_i X_i = \text{median}(X)$$

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

$$w_i \geq 0.$$  

Both problems above can be solved as linear programming problems. I considered the univariate case, but these formulations can be extended to the multivariate case. The formulations presented in this section will be explored in future research. Chapter 6 presents multivariate extensions for robust EL.
Chapter 6

Multivariate Extensions

The examples in Chapter 5 are univariate. They involve random variables. A random variable is a function whose domain is a sample space and whose range is the real line. I now explore multivariate extensions to random vectors. A random vector is a function whose domain is a sample space and whose range is the \(d\)-dimensional Euclidean space. Random vector analysis is important because random vectors are common in real life applications. Regression analysis is one example. Section 6.1 describes the bivariate problem. Section 6.2 gives an example of finding a bivariate point estimator.

6.1 Bivariate Nonlinear Programming Problem

The nonlinear programming problem for finding bivariate robust EL point estimators for the mean parallels its univariate counterpart of Chapter 3. Bivariate data can be viewed as making two measurements on each subject. For example, each row of Equation (6.1) represents a different subject and each column represents a different
measurement:

\[
\begin{pmatrix}
X_{11} & X_{12} \\
X_{21} & X_{22} \\
\vdots & \vdots \\
X_{n1} & X_{n2}
\end{pmatrix}
\]

(6.1)

The robust EL bivariate nonlinear programming problem for the mean vector, \( \mu \), is:

\[
\max_p \sum_{i=1}^n n w_i \log(p_i/w_i)
\]

(6.2)

subject to

\[
\sum_{i=1}^n p_i X_{i1} = \mu_1
\]

(6.3)

\[
\sum_{i=1}^n p_i X_{i2} = \mu_2
\]

(6.4)

\[
\sum_{i=1}^n p_i = 1, \ p_i \geq 0.
\]

(6.5)

given \( \sum_{i=1}^n w_i = 1, \ w_i > 0. \)

(6.6)

The \( p \) that maximizes the objective function subject to the given constraints is the optimal bivariate robust EL probability vector. An application of (6.2) is in the next section.

### 6.2 Bivariate Point Estimators

Finding bivariate robust EL point estimators is similar to finding univariate point estimators. Instead of using the absolute deviation, it uses the Mahalanobis distance [12]. The bivariate procedure also extends to \( d \)-variates. The steps are:

- Compute vectors \( p_i X_{i1}, p_i X_{i2}, 1 \leq i \leq n \)
- Compute Mahalanobis Distance:

\[
\delta \left( \left( \begin{array}{c} p_1 X_{11} \\ p_1 X_{12} \end{array} \right) , \left( \begin{array}{c} M_1 \\ M_2 \end{array} \right) \right) = \\
\sqrt{\left( p_1 X_{11} - M_1 \right)^2 + \left( p_1 X_{21} - M_2 \right)^2} \\
\vdots \\
\sqrt{\left( x_{n1} p_n - M_1 \right)^2 + \left( x_{n2} p_n - M_2 \right)^2}
\]

The following example uses Data Set II to illustrate the procedure.

<table>
<thead>
<tr>
<th>Data Set II (n=20)</th>
<th>X_1</th>
<th>X_2</th>
<th>X_1</th>
<th>X_2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.4326</td>
<td>-0.8323</td>
<td>1.1909</td>
<td>-0.6918</td>
</tr>
<tr>
<td></td>
<td>-1.6656</td>
<td>0.2944</td>
<td>1.1892</td>
<td>0.8580</td>
</tr>
<tr>
<td></td>
<td>0.1253</td>
<td>-1.3362</td>
<td>-0.0376</td>
<td>1.2540</td>
</tr>
<tr>
<td></td>
<td>0.2877</td>
<td>0.7143</td>
<td>0.3273</td>
<td>-1.5937</td>
</tr>
<tr>
<td></td>
<td>-1.1465</td>
<td>1.6236</td>
<td>0.1746</td>
<td>-1.4410</td>
</tr>
<tr>
<td></td>
<td>-0.1867</td>
<td>0.5711</td>
<td>0.7258</td>
<td>-0.3999</td>
</tr>
<tr>
<td></td>
<td>-0.5883</td>
<td>0.6900</td>
<td>2.1832</td>
<td>0.8156</td>
</tr>
<tr>
<td></td>
<td>-0.1364</td>
<td>0.7119</td>
<td>0.1139</td>
<td>1.2902</td>
</tr>
<tr>
<td></td>
<td>1.0668</td>
<td>0.6686</td>
<td>0.0593</td>
<td>1.1908</td>
</tr>
<tr>
<td></td>
<td>-0.0956</td>
<td>-1.2025</td>
<td>5.9802</td>
<td>5.8433</td>
</tr>
</tbody>
</table>

**Example 6.1** Data Set II contains 19–bivariate \(N(0, 0, 1, 1, 0)\) independent values and 1–bivariate \(N(6, 6, 1, 1, 0)\) outlier. The symbol \(N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)\) stands for bivariate normal distribution with mean vector elements \(\mu_X, \mu_Y\); variance–covariance matrix values \(\sigma_X^2, \sigma_Y^2, \rho\), and correlation coefficient \(\rho\). Compute a robust EL point estimate for the mean vector.

**Answer:** The bivariate procedure yields that an appropriate choice for \(w\) corresponds to \(c = 0.0342\). This implies that \(w_i = .0518\) for inliers and \(w_i = 0.0167\) for the outlier.
Therefore, a robust EL point estimate for the mean vector is $\mu^* = (0.2634, 0.2626)$, which lies between the mean of the inliers and the mean of the entire data set.

Note that

- $(0, 0)$ is the true mean vector
- $\bar{x} = (0.4567, 0.4514)$
- $\bar{x}_{inliers} = (0.1660, 0.1676)$.

The following figures illustrate the first, middle and last frames of downweighting the outlier.
Chapter 7

Confidence Intervals

The essence of this dissertation is to use the EL method to form robust EL confidence intervals. Prior chapters concern the first step, obtaining a point estimator. I now turn to interval estimators. Creating intervals and regions begins with an investigation into robust EL's distributional properties. Section 7.1 provides simulations that illuminate robust EL's distributional properties. Section 7.2 considers these properties when constructing confidence intervals for the mean. A method for finding confidence intervals for functions other than the mean is presented in Section 7.3.

7.1 Distributional Properties

Empirical likelihood converges in distribution to a \( \chi^2(r) \); \( r \) is the degrees of freedom [11]. This result is based on the assumption that the data are independent and identically distributed. Robust EL assumes that the data are independent, but not identically distributed. Because EL's asymptotic distribution is a central limit theorem result, robust EL's nonidentical distribution may preclude it from being asymptotically \( \chi^2(r) \). This section uses a quantile–quantile plot (Q–Q plot) to compare the \(-2\log R(\mu_0)\) statistic to the \( \chi^2(1) \) distribution. The symbol \( R \) represents robust EL; \( \mu_0 \) is the hypothesized mean value. Figure 7.1 shows the Q–Q plot of
Figure 7.1: Plot of $\chi^2(1)$ Quantile vs. Objective Function Values using data from a normal distribution. Each data set's objective function value is $-2 \log R(\mu_0)$. Plot indicates that Robust EL is right skewed compared to the $\chi^2(1)$ distribution. A 45° line is included in the plot.

values of $-2 \log R(\mu_0)$ based on 500 simulations, each of sample size 20. Each of the 500 simulations consists of 19 $N(0, 1)$ values and 1 $N(0, 6)$ outlier.

The figure indicates that a $\chi^2(1)$ calibration is not reasonable for this small data set. If $-2 \log R(\mu_0)$ were asymptotically $\chi^2(1)$, the points would have approximated the line $y = x$. Since $-2 \log R(\mu_0)$ is not obviously $\chi^2(1)$, I construct bootstrap confidence intervals. Section 7.2 discusses bootstrap confidence intervals.
7.2 Bootstrap Confidence Intervals

The bootstrap is a powerful nonparametric procedure that is often used in estimation [2]. Its general idea is to form bootstrap replicates by sampling with replacement from the original sample. I use the percentile method to bootstrap robust EL confidence intervals. The percentile method uses quantiles of bootstrap replicates to determine the confidence interval’s upper \( U \) and lower \( L \) bounds. To apply the bootstrap to robust EL, I do the following:

**Step 1** From the original sample, generate \( B \) bootstrap samples of size \( n \). Sample with replacement from the original sample.

**Step 2** Determine the number of outliers in each sample.

**Step 3** Use the distance measures discussed in Chapters 5 and 6 to determine the weights to assign to the outliers.

**Step 4** The inliers are upweighted based on the weights of the outliers. Inliers receive equal weights.

**Step 5** Compute the statistic \(-2 \log R^*(\mu^*)\) for each sample. The symbol \( R^* \) represents the robust EL ratio computed from a bootstrap sample. The value \( \mu^* \) maximizes the robust EL function for the original sample.

**Step 6** Order the \( B \) bootstrap replicates. Call the \( i^{th} \) replicate \( R^*_{(i)} \).

**Step 7** For a 95\% confidence interval, choose the 95\% percentile of \(-2 \log R(\mu^*)\).

The horizontal line determined by the 95\% percentile of \(-2 \log R(\mu^*)\) intersects the robust EL curve at two points. The corresponding horizontal axis values of the two points represent the 95\% confidence interval.

The following example demonstrates these steps.
Example 7.1 Use the bootstrap percentile method to form a robust EL confidence interval for Data Set I.

Answer:
I employ the MATLAB function bootstrp, which draws \( B \) bootstrap samples and analyzes them with a user supplied function. I supplied an m-file that finds \(-2 \log R(\mu^*)\) for the weight vector that moves the deviation of the outlier within the range of the inliers' deviations. The bootstrap percentile result is 1.6648. I use this value to form a 95% confidence interval for Data I. See Figure 7.2.

![Graph](image)

Figure 7.2: Plot of \( \mu \) vs. \(-2 \log R(\mu^*)\) for Data Set I. Plot indicates that a 95% confidence interval for the mean of Data Set I lies approximately between -0.18 and 0.39.

This section demonstrates how to construct confidence intervals for the mean. Functions other than the mean are more complicated because they often contain
non-linear constraints instead of the linear constraint, $\sum_{i=1}^{n} p_i X_i = \mu$. Sequential linearization simplifies solving robust EL for functions other than the mean. Section 7.3 describes the sequential linearization process.

7.3 Sequential Linearization

Constructing robust EL confidence intervals for the mean involves linear constraints, which is the simplest case. However, constructing confidence intervals for parameters other than the mean may require nonlinear constraints. I will use sequential linearization to extend robust EL to functions other than the mean, specifically smooth functionals of the mean [15]. The main idea of sequential linearization is to linearize the nonlinear constraints.

Linearizing the nonlinear constraints involves replacing the exact constraint with its Taylor series approximation. As a result, solving robust EL optimization problems with linearized constraints reduces to calculating robust EL for a mean with pseudo observations.

Wood, Do and Broom’s sequential linearization procedure follows [15]. Suppose $\theta$ is the function of interest, $p$ is the multinomial distribution on $n$ points, and $w$ is given. The log-robust EL at $\theta = \theta_0$ is

$$L(\theta_0) = \max_p \sum_{i=1}^{n} nw_i \log(p_i/w_i)$$

subject to:

$$\theta(p) = \theta_0$$
$$\sum_{i=1}^{n} p_i = 1, p_i \geq 0,$$
$$\text{given } \sum_{i=1}^{n} w_i = 1, w_i > 0,$$
Notice that (7.1) generalizes (3.5). The Taylor series expansion of \( \theta(p) \) about \( p = \hat{p}^{(0)} = (\frac{1}{n}, ..., \frac{1}{n}) \) is

\[
\theta(p) \simeq \theta(\hat{p}^{(0)}) + \sum_{i=1}^{n} (p_i - \hat{p}_i^{(0)}) \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i}.
\]  

(7.2)

Replacing the exact constraint, \( \theta(p) = \theta_0 \), by the above approximation yields

\[
\theta(\hat{p}^{(0)}) + \sum_{i=1}^{n} (p_i - \hat{p}_i^{(0)}) \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i} = \theta_0.
\]  

(7.3)

Note that \( \hat{p}^{(0)} \) is the initial or untilted vector of Section 2.3. Equation (7.3) can be rewritten as

\[
\sum_{i=1}^{n} p_i \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i} = \theta_0 - \theta(\hat{p}^{(0)}) + \sum_{i=1}^{n} \hat{p}_i^{(0)} \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i}.
\]  

(7.4)

Setting \( \theta_0 = \mu \) equates (7.4) with the mean constraint.

\[
\sum_{i=1}^{n} p_i X_i = \theta_0
\]  

(7.5)

of (3.5); instead of \( X_i \), there is the pseudo observation

\[
Y_i^{(0)} = \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i}.
\]  

(7.6)

Additionally, \( \theta_0 \) is modified:

\[
\theta_0^{(0)} = \theta_0 - \theta(\hat{p}^{(0)}) + \sum_{i=1}^{n} \hat{p}_i^{(0)} \frac{\partial \theta(\hat{p}^{(0)})}{\partial p_i}.
\]  

(7.7)

One iterates the sequential linearization procedure \( m \) times. The optimal probability vector in the linearized problem is \( \hat{p}^{(1)} \). This vector becomes the initial probability vector in the second iteration, and so on. Wood, et al. found that \( m = 3 \) works well for empirical likelihood because it minimizes coverage error and computational difficulty. In future research I will investigate a good \( m \) value for robust EL.
Chapter 8

Conclusions and Future Work

The purpose of this work is to use the empirical likelihood method of confidence interval construction to construct robust empirical likelihood confidence intervals. This dissertation has shown the following:

- Robust EL allows data analysts to construct robust confidence intervals for the mean.

- The weight vector tilts the likelihood function away from outliers towards inliers. This process robustifies the EL function.

- This method generalizes to multidimensions.

Future work will focus on obtaining confidence regions, and on following:

- an optimal $c$ for choosing an appropriate weight vector, $w$

- multiple outliers receiving different weights

- functions other than the mean
Acknowledgments

Thank you to my committee members: Dr. David W. Scott (Committee Chairperson) and Dr. Katherine Ensor of the Statistics Department; Dr. Yin Zhang (Committee Co-Chairperson) and Dr. Nathaniel Dean of the Computational and Applied Math Department.

Special thank you to Dr. Art Owen of Stanford University, Dr. Keith Baggerly of MD Anderson Cancer Center, and Dr. Juan Meza of Sandia National Laboratory. I also thank Paul David Glenn II, my son.
Bibliography


