

Challenges for EL

This chapter is devoted to problems where empirical likelihood has difficulties, and to ways of mitigating those difficulties. As a case in point, empirical likelihood inferences on the number of distinct possible values from a distribution is completely degenerate. The confidence interval only includes the number of distinct values in the sample, and that value is completely wrong for continuous distributions.

As a less outlandish example, the natural way to define empirical likelihood tests for symmetry or independence are degenerate. The root cause is that these conditions are equivalent to an infinite number of estimating equations. It is, however, possible to use a known point of symmetry, or known independence, as a side condition in inferences. In some settings this is degenerate, in some it reduces to ordinary empirical likelihood, and in others it gives something new. It is also possible to test for approximate symmetry or approximate independence, defined through a finite subset of the infinite set of constraints.

For some parametric likelihoods, the usual asymptotic theory does not hold. This can happen when the range of data values depends on the parameter, when the true value of the parameter is on the boundary of the set of possible values, or when the value of one parameter does not affect the predictions, if a second parameter is zero. Empirical likelihood based on the estimating equations from these likelihoods cannot be expected to have a χ^2 calibration.

10.1 Symmetry

The distribution F of $X \in \mathbb{R}$, is symmetric about a center c , if every interval (a, b) has the same probability as the interval $(c - b, c - a)$. In the familiar case where F has a density or mass function f , symmetry means that $f(c + x) = f(c - x)$, which we can rewrite as $f(x) = f(2c - x)$.

A natural approach to nonparametric inference under symmetry is to construct a family \mathcal{F}_S of symmetric distributions that put positive probability on every observation. Such a family can be represented with $n + 1$ parameters: the center c of symmetry, and weights w_i attached to X_i , which by symmetry also attach to $\tilde{X}_i = 2c - X_i$. We suppose that $\sum_{i=1}^n w_i = 1/2$. Under this setting the probability that F puts on x is $\sum_{i=1}^n w_i(1_{X_i=x} + 1_{\tilde{X}_i=x})$, so that x is double counted if it is both a data point and the reflection of a data point.

First we consider whether nonparametric likelihood can help identify the center

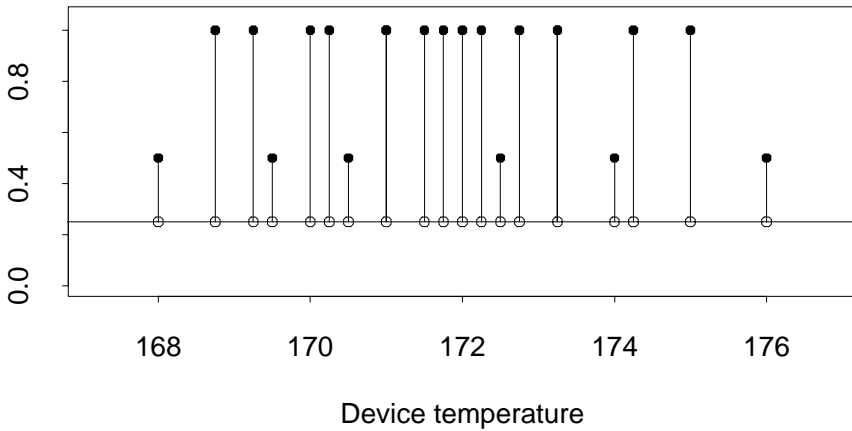


Figure 10.1 The raw data are temperatures of six solid-state electronic devices, in degrees Celsius. The function shown is the empirical likelihood for the center of symmetry, assuming a symmetric distribution. The likelihood ratio takes the value $1/2$ at each of the six data values. Source: Hahn & Meeker (1991, Chapter 13).

c of symmetry. Define the points $c_{ij} = (X_i + X_j)/2$ for $1 \leq i \leq j \leq n$. For continuously distributed data, there are no ties among the c_{ij} . If c is not one of the c_{ij} , then the nonparametric likelihood is maximized at $w_i = 1/(2n)$ and takes the value $L_0 = (2n)^{-n}$. When $c = c_{ii} = X_i$, the likelihood $(2w_i) \prod_{j \neq i} w_j$ takes on a maximum value of $L_1 = 2(2n)^{-n}$. Finally, if $c = c_{ij}$ for $j \neq i$, then the likelihood $(w_i + w_j)^2 \prod_{k \notin \{i,j\}} w_k$ takes on maximum value $L_2 = 4(2n)^{-n}$.

Thus nonparametric likelihood arguments lead to a degenerate inference on c , under sampling from continuous F . The $n(n-1)/2$ points c_{ij} with $i < j$ maximize the likelihood, next come the n points $c_{ii} = X_i$ with a likelihood ratio of $1/2$, and finally every other point in the real line, with a likelihood ratio of $1/4$. Figure 10.1 shows this function for the operating temperatures of six solid-state electronic devices. Even the midpoint between the two largest X_i is a mode of the likelihood ratio function.

Although testing for symmetry and estimating a point of symmetry are degenerate, imposing a known symmetry as side information is not necessarily degenerate. In some cases this side information does not change empirical likelihood inferences, and in some it gives a new and nondegenerate method.

Suppose we know that F is symmetric about c . Then $F(\{x\}) = F(\{2c - x\})$ for all x . This symmetry can be imposed as a side constraint on F , to sharpen inferences on θ defined by $E(m(X, \theta)) = 0$, for some function m . We place weights w_i on X_i and weights w_i on $\tilde{X}_i = 2c - X_i$, where $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1/2$.

The function m has even and odd parts defined by

$$m_E(X, \theta) = \frac{1}{2} \left[m(X, \theta) + m(2c - X, \theta) \right], \quad \text{and}$$

$$m_O(X, \theta) = \frac{1}{2} \left[m(X, \theta) - m(2c - X, \theta) \right].$$

Now $m = m_E + m_O$, and the odd part has weighted mean zero by symmetry. Thus the definition of θ may be replaced by $E(m_E(X, \theta)) = 0$. Operationally, we replace m by m_E and apply empirical likelihood as usual. If $m = m_E$ then empirical likelihood is unchanged by this operation, and so imposing symmetry this way does not make any difference. If $m = m_O$ then the equations are degenerate, reducing to $E(0) = 0$ for all θ . An odd function is known to have mean zero under a symmetric distribution, and data are not required to draw that conclusion. If m has nonzero odd and even parts, then the result can be nondegenerate and different from the usual empirical likelihood.

The following three examples illustrate the possible cases. In each of them X is a real random variable known to be symmetric about a value c .

Example 10.1 (Mean under symmetry) For $x, \mu \in \mathbb{R}$, the estimating function $m(x, \mu) = x - \mu$, defines a univariate mean μ . Then $m_E(x, \mu) = c - \mu = 0$, telling us what we already know ($\mu = c$), and without making any use of the data.

Example 10.2 (Variance under symmetry) Because $E(X) = c$, the variance σ^2 of X is defined by the estimating equation $E((X - c)^2 - \sigma^2) = 0$, so that $m = (x - c)^2 - \sigma^2$. This m is even, and so empirical likelihood inferences are unchanged by replacing m by m_E .

Example 10.3 (Tail probability under symmetry) Suppose that we are wish to estimate a tail probability $\theta = \Pr(X > x)$. To avoid trivialities, assume that $x \neq c$, and $\min_i X_i < x < \max_i X_i$. The estimating function for θ is $m(X, \theta) = 1_{X > x} - \theta$. This m has nondegenerate even and odd parts

$$m_E(X, \theta) = \frac{1}{2} \left(1_{X > x} + 1_{X < 2c - x} \right) - \theta,$$

$$m_O(X, \theta) = \frac{1}{2} \left(1_{X > x} - 1_{X < 2c - x} \right) - \theta.$$

Thus, for this case, empirical likelihood inferences imposing symmetry are nondegenerate and different from empirical likelihood inferences that do not impose symmetry. Knowing that X is symmetric allows us to use data from one tail of the distribution to estimate a probability in the other tail. In practice, of course, we would have to be fairly sure of the symmetry to trust an estimate of one tail based on data from the other. For $x > c$, we could get the same result by replacing every X_i by $Z_i = \max(X_i, 2c - X_i)$, drawing inferences on $\Pr(Z > x) = 2\theta$, and dividing out the factor of 2 from this probability.

One consequence of symmetry of F is that odd (antisymmetric) functions of $X - c$ have mean zero. Suppose that $\phi(x) = -\phi(-x)$, so that ϕ is odd. If $\int |\phi(x - c)|dF(x) < \infty$, then $\int \phi(x - c)dF(x) = 0$. Conversely, this property can be used as a definition of symmetry about c , through specially chosen functions of the form

$$\phi_{a,b}(x) = 1_{a < x < b} - 1_{-b < x < -a}. \quad (10.1)$$

We might replace symmetry by a weaker concept using only some finite number r of conditions $\int \phi_j(x - c)dF(x) = 0$, for $j = 1, \dots, r$. Instead of testing for symmetry about c , we test $E(\phi_j(X - c)) = 0$ for $j = 1, \dots, r$. If we have a specific value of c such as $c = 0$ in mind, then we may simply test whether these r functions all have mean 0. If we wish to test for symmetry about an unknown center, then all values of c that are not rejected form a conservative confidence region for the center of symmetry. If this region is empty, then we infer that F is not a symmetric distribution.

Here are some signed moment functions that might be used to approximate symmetry

$$M_j(z) = \text{sign}(z) |z|^{j-1}, \quad j = 1, 2, \dots \quad (10.2)$$

A vanishing signed moment $E(M_j(X - c))$ corresponds to X having median c , when $j = 1$ and mean c , when $j = 2$. An alternative is to use a set of functions ϕ_{a_j, b_j} of the form (10.1) for a set of intervals (a_j, b_j) .

For 672 National Football League (NFL) games of American professional football, the observed pointspread was compared to the pointspread established by professional book-makers. We will look at the values of $F - U - S$ where F is the actual number of points scored by the team favored to win, U is the number scored by the underdog team, and S is the published pointspread — the number of points by which the favorite was expected to win. The spreads S take values that are integer multiples of 1/2 point. Figure 10.2 shows a histogram of these observed minus expected pointspreads. The data appear to be nearly normally distributed, with a center near zero and a surprisingly large standard deviation of 13.86 points. Thus about 95% of observed pointspreads came within plus or minus 28 points of the prediction.

It is plausible that data like these should be centered around 0. Even-money bets on whether the favorite beats the pointspread make most sense if the median discrepancy is close to zero. Otherwise bettors would catch on and exploit the difference. Because the data are nearly normally distributed, the mean may be a better estimate of the center of symmetry than the median. Also, if there are bets made that pay proportionally to the number of points by which the pointspread is beaten (or missed), then those bettors might pay attention to the mean difference between observed and predicted point spreads.

Figure 10.3 shows empirical likelihood curves for the mean and the median of the pointspread data. The sample mean is very close to zero, and the true mean seems to be within roughly one point of zero. The sample median is between -1

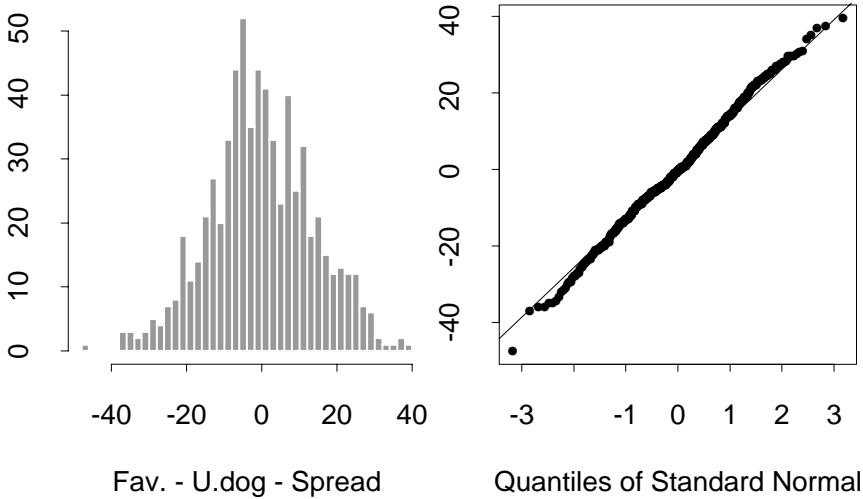


Figure 10.2 For each of 672 NFL games, the betting-line pointspread has been subtracted from the observed one as described in the text. The data come from Stern (1991). The data are plotted as a histogram on the left, and in a normal QQ plot on the right.

point and $-1/2$ point, and the true median appears to be between -2 points and $1/2$ point.

Now suppose that we consider a center of symmetry θ for which

$$0 = E(1_{X>\theta} - 1_{X<\theta}), \quad \text{and}$$

$$0 = E(X - \theta),$$

corresponding to signed moment functions M_1 and M_2 from equation (10.2) above. The empirical likelihood ratio function for θ is plotted in the lower panel of Figure 10.3. The unusual shape is easily explained. Consider a joint likelihood surface for the mean and median of the data. For a fixed value of the mean, that surface has step discontinuities as the median crosses observed data values. For a fixed value of the median, the surface is smooth as the mean varies. The function shown is a transect along the 45° line of the joint mean-median likelihood surface. This function takes jumps at the same points where the likelihood function for the median does. Between those jump points, it varies smoothly. The sawtooth shape of the likelihood curve gives a 95% confidence region that is a union of four disjoint intervals. The likelihood is relatively high at, and just to the left of, multiples of $1/2$ point. The smallest interval containing all four parts of the confidence interval for θ is narrower than either of the intervals for the mean or the median.

If we were not confident that the mean and median were at the same point, we could test this by examining the maximum value over θ of the empirical likelihood

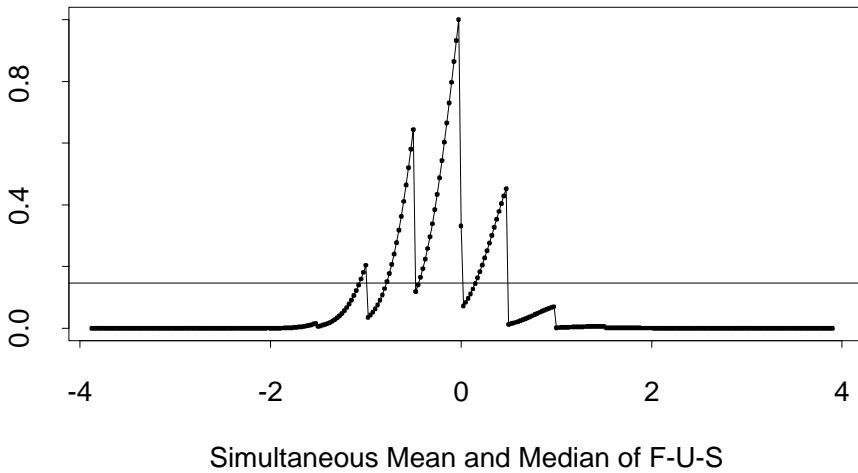
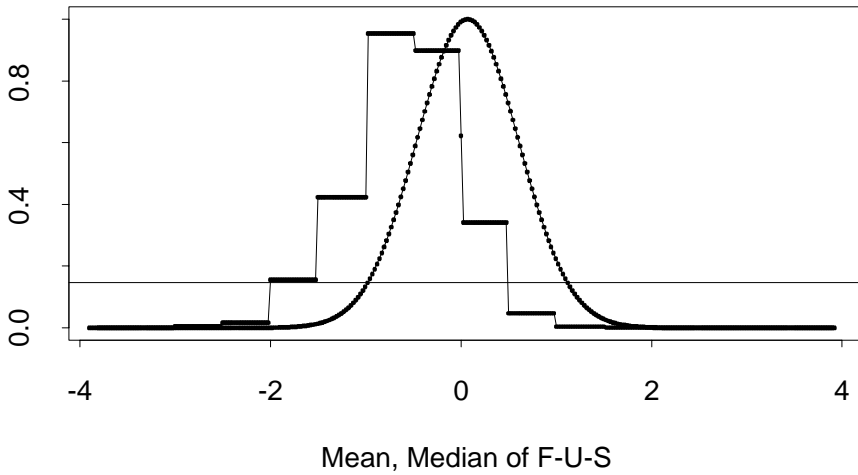


Figure 10.3 The top plot shows empirical likelihood curves for the mean and for the median of the pointspread data in Figure 10.2. The bottom plot shows the empirical likelihood for a center θ that is simultaneously the mean and the median of the distribution.

for the hypothesis that θ is both the mean and median of X . At the MELE $\hat{\theta}$, the log empirical likelihood is -0.498 . Using a $\chi^2_{(1)}$ calibration, we obtain a p -value of $\Pr(\chi^2_{(1)} \geq -2 \times (-0.498)) = 0.318$, so there is no reason to reject the presumed equality of the mean and median.

Further analysis shows that using the first three signed moments does not make

much difference to the empirical likelihood ratio function, but that using the fourth and higher signed moments calls into question the symmetry of the data. Because these higher order signed moments have very heavy tails, dominated by the small number of extreme games, it does not seem wise to use them. If one were interested in imposing higher order symmetry, it would be preferable to use a set of bounded functions ϕ_j , perhaps of the form given in (10.1).

10.2 Independence

Suppose that $(X, Y) \in \mathbb{R}^2$ and that we want to test for independence of X and Y . As with the problem of symmetry, a direct formulation of empirical likelihood leads to a degenerate answer.

Suppose that the data are (X_i, Y_i) pairs, $i = 1, \dots, n$, from a continuous joint distribution. Consider a distribution for X that has weight u_i on X_i and a distribution for Y with weight v_j on Y_j . Under independence of X and Y , the weight on the pair (X_i, Y_j) is $u_i v_j$. Without an assumption of independence, a distribution can put weight w_{ij} where w_{ij} is not necessarily of product form. To maximize the likelihood without assuming independence, put $w_{ij} = 0$ if $i \neq j$ and $w_{ii} = 1/n$. To maximize the likelihood ratio assuming independence put $u_i = v_j = 1/n$. As a result, the likelihood ratio for independence is $n^{-2n}/n^{-n} = n^{-n}$, without regard to the data. Thus the empirical likelihood test for independence is degenerate on continuous data.

If the (X_i, Y_i) pairs are from a discrete distribution, with finitely many possible values, the situation changes. Then the empirical likelihood test becomes a standard multinomial likelihood test of independence. Significance levels for that likelihood ratio test can be set using the $\chi^2_{(r-1)(s-1)}$ distribution where X takes r distinct values and Y takes s distinct values.

For the problem of symmetry, tests were degenerate, but some nondegenerate methods arose when symmetry was imposed as a side constraint. A similar effect occurs when independence is imposed as a side constraint.

Suppose that we know that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are independent. Then there is no difference between n paired observations (X, Y) and two unpaired samples of X and Y . Because the data are effectively unpaired, there is no real need for the sample sizes to be equal. Suppose then that $X_i \sim F_0$ for $i = 1, \dots, n$, and $Y_j \sim G_0$ for $j = 1, \dots, m$ are independent. We are interested in θ defined by

$$E(h(X, Y, \theta)) = 0.$$

Here are example functions h that are of interest in applications,

$$\begin{aligned} h_1(X, Y, \theta) &= X - Y - \theta, \quad \text{and} \\ h_2(X, Y, \theta) &= 1_{X>Y} - \theta. \end{aligned}$$

When $p = q$, the function h_1 can be used for inferences on $E(X) - E(Y)$. When $p = q = 1$, the function h_2 can be used for inferences on $\Pr(X > Y)$.

Let $F = \sum_{i=1}^n u_i \delta_{X_i}$ and $G = \sum_{j=1}^m v_j \delta_{Y_j}$ where $u_i \geq 0$, $v_j \geq 0$, and

$\sum_{i=1}^n u_i = \sum_{j=1}^m v_j = 1$. The likelihood of the pair (F, G) is $\prod_{i=1}^n u_i \prod_{j=1}^m v_j$. This is maximized by taking $u_i = 1/n$ and $v_j = 1/m$, so the likelihood ratio function is $\prod_{i=1}^n n u_i \prod_{j=1}^m m v_j$. The estimating equation is

$$\sum_{i=1}^n \sum_{j=1}^m u_i v_j h(X_i, Y_j, \theta) = 0.$$

The profile empirical likelihood ratio function for θ is

$$\mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^n n u_i \prod_{j=1}^m m v_j \mid \sum_{i=1}^n \sum_{j=1}^m u_i v_j h(X_i, Y_j, \theta) = 0, \right. \\ \left. u_i \geq 0, \sum_{i=1}^n u_i = 1, v_j \geq 0, \sum_{j=1}^m v_j = 1 \right\}.$$

Under mild conditions $-2 \log \mathcal{R}(\theta_0) \rightarrow \chi_{(d)}^2$ where d is the dimension of θ . A nonrigorous argument is given in Chapter 11.4 for the case of two samples with a scalar function $h(X, Y, \theta)$ and a scalar θ .

Independence corresponds to an infinite set of constraints. These take the form $E(\phi(X)\eta(Y)) = E(\phi(X))E(\eta(Y))$ for every ϕ, η pair with $E(|\phi(X)\eta(Y)|) < \infty$. If independence of X and Y is in doubt, and we want to test it, we can retain the original (X_i, Y_i) pairings and then for a list $j = 1, \dots, r$ of functions ϕ_j and η_j test whether $E(\phi_j(X)\eta_j(Y)) = E(\phi_j(X))E(\eta_j(Y))$.

10.3 Comparison to permutation tests

A disadvantage of testing approximate symmetry or approximate independence is that symmetry or independence can be violated by a data distribution satisfying all of the estimating equations used in the approximate test. This motivates taking a large value of r , to get a test sensitive to more kinds of departures. For larger r , a larger $\chi_{(r)}^2$ threshold must be surpassed in order to reject independence. So large r can result in less power than small r . Judgment is required in order to select r conditions that cover either the likely departures, or if possible, the consequential departures, from symmetry or independence. A sieve method (Chapter 9.10) letting $r = r(n) \rightarrow \infty$ as $n \rightarrow \infty$ might provide an effective approach.

Permutation tests are commonly used for testing independence, and similar procedures are available for testing symmetry. For a permutation test of independence one starts with a statistic $T((X_i, Y_i), i = 1, \dots, n)$ for which larger values represent greater departures from independence of X and Y . Then the Y_i are permuted with respect to the X_i , replacing Y_i by $Y_{\pi(i)}$ where $(\pi(1), \dots, \pi(n))$ is a uniform random permutation of $(1, \dots, n)$. A uniform random permutation takes each of the $n!$ possible permutations with probability $1/n!$. Then a histogram is made of the values $T((X_i, Y_{\pi(i)}), i = 1, \dots, n)$ taken by T under a large number of independent random permutations. If the original T value is in the largest 5% of this histogram, then independence is rejected at the 5% level. When $n!$ is not

too large then it is possible to consider all the permutations instead of sampling them.

As with empirical likelihood tests of approximate independence, these permutation tests can fail to detect those departures from independence which do not have a strong effect on T . The statistic T can be constructed as a composite of some number r of test statistics, each of which is sensitive to one aspect of dependence between X and Y . Once again there is a trade-off in the value of r .

Compared to permutation tests, empirical likelihood offers the advantage of using the data to combine the r test statistics. Consider testing $E(m_j(X_i, Y_i)) = 0$, simultaneously for all $j = 1, \dots, r$, for real-valued estimating functions m_j . A permutation test could take

$$T = \sum_{j=1}^r \left(\sum_{i=1}^n m_j(X_i, Y_i) \right)^2$$

as its test criterion. But it might be better to weight the criteria, either a priori or based on the data. An empirical likelihood test based on $\mathcal{R}_0 = \max \prod_{i=1}^n n w_i$ subject to constraints $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$, and $\sum_{i=1}^n w_i m_j(X_i, Y_i) = 0$ for $j = 1, \dots, r$ uses the data to combine the individual tests.

The significance level for the test can be based on an asymptotic χ^2 distribution for \mathcal{R}_0 under the null hypothesis that X_i and Y_i are independent. We can also use the permutation distribution of \mathcal{R}_0 directly. For N random permutations $\pi(i)$ of $1, \dots, n$, compute \mathcal{R}_0 on the data pairs $(X_i, Y_{\pi(i)})$, $i = 1, \dots, n$. If the original \mathcal{R}_0 is smaller than αN of the permuted \mathcal{R}_0 values, then reject the null hypothesis at level α .

10.4 Convex hull condition

Empirical likelihood confidence regions for the mean are nested within the convex hull of the data. Their coverage level is necessarily smaller than that of the convex hull itself. This constraint is limiting when n is small, p is large, or the confidence level $1 - \alpha$ is high.

The Euclidean likelihood and some other members of the Cressie-Read family do not restrict the weights w_i to be nonnegative, thus allowing the reweighted mean $\sum_i w_i X_i$ to escape the convex hull of the data.

The empirical likelihood-t method also produces confidence regions for the mean that can escape from the convex hull. For testing whether $E(X) = \mu$, the constraint $\sum_{i=1}^n w_i X_i = \mu$ is replaced by

$$\begin{aligned} & \left(\sum_{i=1}^n w_i (X_i - \mu_w)(X_i - \mu_w)' \right)^{-1/2} (\mu_w - \bar{X}) \\ &= \left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \right)^{-1/2} (\bar{X} - \mu), \end{aligned}$$

where $\mu_w = \sum_{i=1}^n w_i X_i$.

For $d = 1$, the new constraint equates two quantities. The first is the number of weighted standard deviations by which the reweighted mean differs from \bar{X} . The second is the number of unweighted standard deviations by which \bar{X} differs from the candidate mean μ . Notice that the reweighted mean is not μ . If it were, the result could not escape the convex hull. For $d > 1$ a similar interpretation can be made in terms of analogous Mahalanobis distances $(u - \mu)' S^{-1} (u - \mu)$.

The empirical likelihood-t method is analogous to the bootstrap-t method. It replaces a quantity by a studentized version of that quantity. The empirical likelihood-t constraints can be applied to nonparametric likelihoods other than empirical likelihood. For empirical likelihood-t, a Bartlett correction is available. See Chapter 10.8.

10.5 Inequality and qualitative constraints

Many interesting hypotheses are expressed in terms of inequalities, not equalities. For example, if μ_1, \dots, μ_k are the means of k different populations, the constraints

$$\mu_{j+1} - \mu_j \geq 0, \quad j = 1, \dots, k - 1, \quad (10.3)$$

impose a qualitative constraint that μ_j is a nondecreasing function of the group labels. Estimating and testing of parameters subject to isotone constraints like (10.3) is known as order restricted inference.

Another qualitative constraint that might be useful is stochastic monotonicity. Suppose that $(X, Y) \in \mathbb{R}^2$ are jointly observed and we are sure that Y can only increase if X increases. Stochastic monotonicity corresponds to infinitely many constraints of the form $\Pr(Y \geq y | X = x) \geq \Pr(Y \geq y | X = x')$ for all y and all $x \geq x'$.

A standard technique for imposing unimodality constraints is to impose monotonicity constraints on either side of a mode $c \in \mathbb{R}$. When c is not known, then an outer loop searching over candidate modes is required.

A widely studied qualitative constraint for $X \in [0, \infty)$ is that f has a non-increasing density. The NPMLE in this setting is known to be a mixture of $U[0, x)$ distributions, characterized as the smallest concave function lying above the empirical CDF.

Some results have been obtained for empirical likelihood ratios in inequality constrained problems. Suppose that a parameter $\theta \in \mathbb{R}^p$ is defined by

$$E(m(X, \theta)) = 0,$$

where $m(X, \theta) \in \mathbb{R}^p$, and consider the hypotheses

$$\begin{aligned} \mathcal{H}_0 : g_j(\theta) &= 0, \quad j = 1, \dots, k, \quad \text{and} \\ \mathcal{H}_1 : g_j(\theta) &\geq 0, \quad j = 1, \dots, k, \end{aligned}$$

where $k \leq p$. The functions g_j may represent all the constraints of the qualitative feature we are interested in or a judiciously chosen subset of them.

Two statistical problems of interest are the testing of \mathcal{H}_1 and the testing of \mathcal{H}_0 assuming that \mathcal{H}_1 holds. Following parametric likelihood theory, we define

$$\mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i m(X_i, \theta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}$$

and then

$$\begin{aligned} \mathcal{R}_0 &= \max \{ \mathcal{R}(\theta) \mid g_j(\theta) = 0, j = 1, \dots, k \}, \quad \text{and} \\ \mathcal{R}_1 &= \max \{ \mathcal{R}(\theta) \mid g_j(\theta) \geq 0, j = 1, \dots, k \}. \end{aligned}$$

The hypothesis \mathcal{H}_1 is rejected if \mathcal{R}_1 is too small, and assuming \mathcal{H}_1 , the further constraint \mathcal{H}_0 is rejected if $\mathcal{R}_0/\mathcal{R}_1$ is too small.

This poses two new problems for empirical likelihood. The first is how to maximize an empirical likelihood subject to inequality constraints. The second is how to calibrate the likelihood ratios. These issues are described in the references in Chapter 10.8. Here we summarize the calibration results.

The asymptotic distribution of $-2 \log \mathcal{R}_1$ is not necessarily chisquared. The null hypothesis is quite heterogenous. Perhaps the true value θ_0 has $g_j(\theta_0) = 0$ for some j , and $g_j(\theta_0) > 0$, for other j . There may be as many as 2^k different cases to consider. It is known that under mild conditions, for $Q > 0$,

$$\lim_{n \rightarrow \infty} \Pr(-2 \log \mathcal{R}_1 \leq Q) = \sum_{i=0}^k \gamma_i \Pr(\chi_{(i)}^2 \leq Q),$$

where $\gamma_i = \gamma_i(\theta_0) \geq 0$ are weights that sum to 1, and $\chi_{(0)}^2 = 0$. This limiting distribution of $-2 \log \mathcal{R}_1$ is known as a chi-bar squared, often written $\bar{\chi}^2$. A chi-bar squared distribution commonly applies to parametric likelihoods in the presence of inequality constraints. Similarly,

$$\lim_{n \rightarrow \infty} \Pr(-2 \log(\mathcal{R}_0/\mathcal{R}_1) \leq Q) = \sum_{i=0}^k \gamma_i \Pr(\chi_{(k-i)}^2 \leq Q)$$

with the same $\gamma_i(\theta_0)$.

Each γ_i is a sum of products of probabilities that certain multivariate normal random variables have no negative components. The normal vectors have mean zero and it is possible to construct sample estimates of their covariance matrices.

10.6 Nonsmooth estimating equations

Theorem 3.4 allows us to construct tests and confidence regions for parameters defined through very general estimating equations. **Theorem 3.6** justifies profiling out nuisance parameters, but only from smooth estimating equations.

Let θ and a nuisance parameter ν be defined through $E(m(X, \theta, \nu)) = 0$ for

a nonsmooth estimating function $m(X, \theta, \nu)$. Then profiling out ν raises both theoretical and numerical challenges.

In this text we have encountered several problems of this type: the interquartile range defined through (3.21), a problem in regression tolerance intervals defined by (4.16) through (4.19), the sensitivity and specificity of logistic regression (Chapter 4.7), and the difference between medians of censored data.

Some results are known for problems in which the estimating equations are smoothed, with decreasing smoothness as $n \rightarrow \infty$. The theorem below requires no smoothing:

Theorem 10.1 For $i = 1, \dots, k$, let $X_{ij} \in [0, \infty)$ for $j = 1, \dots, n_i$ be failure times from distribution F_i . Suppose that all F_i have a common median θ_0 , and that F_i is a continuous distribution with density f_i where $f_i(\theta_0) > 0$. Let $Y_{ij} \in [0, \infty)$ be corresponding right censoring times from distribution G_i , where $G_i([0, \infty)) > 0$. Assume that all X_{ij} and Y_{ij} are independent. Let $Z_{ij} = \min(X_{ij}, Y_{ij})$ and $\delta_{ij} = 1_{X_{ij} \leq Y_{ij}}$, and define

$$L(F_1, \dots, F_k) = \prod_{i=1}^k \prod_{j=1}^{n_i} F(\{Z_{ij}\})^{\delta_{ij}} F((Z_{ij}, \infty))^{1-\delta_{ij}}$$

$$\mathcal{R}(\theta_1, \dots, \theta_k) = \frac{\max \{L(F_1, \dots, F_k) \mid F_i([\theta_i, \infty)) = 1/2, i = 1, \dots, k\}}{\max_{F_1, \dots, F_k} \{L(F_1, \dots, F_k)\}},$$

and $\mathcal{R}(\theta) = \mathcal{R}(\theta, \theta, \dots, \theta)$. Then $-2 \max_{\theta} \log \mathcal{R}(\theta) \rightarrow \chi_{(k-1)}^2$ as $\min n_i \rightarrow \infty$.

Proof. Naik-Nimbalkar & Rajarshi (1997). \square

Now suppose that $X_1, \dots, X_n \sim F$ are independent real random variables with CDF $F(x) = F((-\infty, x])$. As usual, the order statistics are denoted by $X_{(1)} \leq \dots \leq X_{(n)}$. Statistics known as L -estimators can be written as $T_n = \sum_{i=1}^n c_{ni} X_{(i)}$ where $c_{ni} = \int_{(i-1)/n}^{i/n} dG(u)$. Here $G(u)$ can be a distribution function, or more generally, a weighted difference of distribution functions, $G(u) = c_+ G_+(u) - c_- G_-(u)$, where G_{\pm} are distribution functions and $c_{\pm} \geq 0$. The population value of T is $T(F) = \int_0^1 F^{-1}(u) dG(u)$, where $F^{-1} = \inf\{x \mid F(x) \geq u\}$.

For example, when $0 < \alpha < 1/2$, the α -trimmed mean has $dG(u) = (1 - 2\alpha)^{-1}$ if $\alpha/2 < u < 1 - \alpha/2$ and $dG(u) = 0$ otherwise. As a second example, the interquartile range has $c_{\pm} = 1$, $G_+(u) = 1_{u \leq 3/4}$ and $G_-(u) = 1_{u \leq 1/4}$.

The empirical likelihood function is

$$\mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^n n w_i \mid T \left(\sum_{i=1}^n w_i \delta_{X_i} \right) = \theta, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\},$$

where we suppose that T uses F determined from F in the obvious way.

We suppose here that $dG(u) = g(u)$ where $\int_0^1 |g(u)| < \infty$. This rules out the interquartile range, though it can be closely approximated by an L -estimate satisfying this condition, using smoothed versions of $G_{\pm} = 1_{u \leq (2\pm 1)/4}$.

Theorem 10.2 Let X_1, \dots, X_n be IID from F_0 . Suppose that for some $p \in (0, \infty)$, some $M \in (0, \infty)$, and some $d \in (1/6, 1/2)$ both

$$|g(u)| \leq M[u(1-u)]^{\frac{1}{p}-\frac{1}{2}+d}, \quad \text{and}$$

$$|F^{-1}(u)| \leq M[u(1-u)]^{-\frac{1}{p}}$$

hold for all $0 < u < 1$. Then $-2 \log \mathcal{R}(T(F_0)) \rightarrow \chi_{(1)}^2$ as $n \rightarrow \infty$.

Proof. La Rocca (1995a) \square

10.7 Adverse estimating equations and black boxes

In parametric likelihood settings there can be difficulty in passing from maximizing the log likelihood $\sum_{i=1}^n \log f(X_i; \theta)$, to solving $\sum_{i=1}^n m(X_i, \theta) = 0$. The set of solutions θ may include multiple global maxima, local maxima that are not global maxima, local minima, and saddlepoints. Such equations pose a challenge for empirical likelihood too, for solutions to $\sum_{i=1}^n w_i m(X_i, \theta) = 0$ might not correspond to unique global maxima of $\sum_{i=1}^n w_i \log f(X_i; \theta)$ as we would have hoped.

Parametric likelihood methods have developed the farthest for models in which solving the likelihood equation does indeed give the maximum likelihood estimate. The most widely used parametric models are those with log concave likelihoods, such as the normal distribution, the binomial, and the Poisson. When $\log f(X; \theta)$ is concave in θ then so is $\sum_{i=1}^n w_i \log f(X_i; \theta)$ and so empirical likelihood is on firmer ground in these cases, too.

Another big challenge comes from methods such as classification and regression trees (CART) and projection pursuit regression. These may be thought of as black boxes that take data in and produce answers. It would be a daunting computational challenge to find the weighting of the data closest to equality for which a classification tree predicted $y_0 \in \mathbb{R}$ given a feature vector X_0 . One reason is that the method is nonsmooth. A small change in the weighting of the data can change the shape of the tree and the variables used to split the data. Bootstrap resampling is ideally suited to computations for black box estimators of this type. In addition to the computational challenge, the theoretical behavior of the resulting likelihood ratio would have to be at least as complicated as that leading to the chi-bar squared distribution in Chapter 10.5.

10.8 Bibliographic notes

Brown & Chen (1998) study empirical likelihood for an assumed common value of the mean and median. They prefer the Euclidean log likelihood for this problem because it allows negative weights and so produces a bounded log likelihood. Brown & Chen (1998) prefer a smoothed version of the log likelihood ratio. They discuss robustness and asymptotic normality of the combined mean/median estimator.

The signed moments are due to Qu (1995). He shows that a location estimator defined as an MELE using a small number of signed moments is nearly fully efficient compared to some parametric maximum likelihood estimators from symmetric distributions. He extends the signed moment arguments to multiple regression estimators with symmetrically distributed residuals.

The approach in Chapter 10.1 symmetrizes the estimating equations. By contrast, Jing (1995a) symmetrizes a data set around its sample mean, then constructs an empirical likelihood test of whether the resulting $2n$ observations have mean μ . He obtains a χ^2 limit but finds that Bartlett correctability does not hold.

Romano (1988) considers bootstrap versions of permutation tests for symmetry and independence.

Breiman, Friedman, Olshen & Stone (1984) describe classification and regression trees. Friedman & Stuetzle (1981) introduce projection pursuit regression.

Zhou (2000) considers a location family in which $X_{ij} - \mu_i \sim F$ for $i = 1, \dots, k$ and $j = 1, \dots, n_i$. As with symmetry and independence, this location family constraint corresponds to an infinite number of moment constraints.

Empirical likelihood-t was proposed by Baggerly (1999), who also considers an entropy version. Baggerly (1999) notes that one price to be paid for escaping the convex hull is a loss of transformation invariance. Baggerly (1999) reports some simulations in which the studentized empirical entropy method gives very good coverage accuracy in small samples. He also applies a Bartlett correction to those intervals. Bartlett correction of empirical entropy appears in simulations to make a worthwhile improvement in the coverage error, although it does not improve the error rate from $O(1/n)$. Baggerly speculates that this is because the empirical entropy distance gives relative error of $O(1/n)$ instead of an absolute error of that order, pointing to work by Jing, Feuerverger & Robinson (1994).

The discussion of empirical likelihood ratios under inequality constraints in Chapter 10.5 is based on El Barmi (1996). El Barmi gives expressions for the weights in the $\bar{\chi}^2$ distributions. El Barmi & Dykstra (1994) describe maximization of multinomial likelihoods subject to constraints expressed as the intersection of a finite number of convex sets. Hoff (2000) considers maximizing empirical likelihoods subject to stochastic monotonicity constraints linking a finite number of distributions.

Grenander (1956) constructs an NPMLE for a distribution known to have a monotone density on $[0, \infty)$. Groeneboom & Wellner (1992) characterize the result as the least concave majorant of the empirical CDF. Lindsay (1995) describes mixture-based methods for this and other statistical problems. Banerjee & Wellner (2000) present describe the asymptotics of likelihood ratios under monotonicity constraints. The results are characterized by $n^{1/3}$ rate asymptotics involving Brownian motion with quadratic drift. Azzalini & Hall (2000) show how qualitative information can be used to reduce variability. The result can be like having $O(n^{2/3})$ extra observations. Although this effect becomes negligible as $n \rightarrow \infty$, the benefit may be substantial at practical sample sizes.

Qin & Zhao (1997) prove a chisquared limit for the difference of two smoothed

quantiles from two different populations, by extending the arguments of Chen & Hall (1993) for smoothing of a single quantile. The amount of smoothing applied to the two quantiles tends to zero as the sample sizes tend to infinity. Naik-Nimbalkar & Rajarshi (1997) prove [Theorem 10.1](#) without requiring that the underlying statistics be smoothed.

La Rocca (1995a) proves [Theorem 10.2](#), and also considers more general statistics defined as L estimators applied to values $h(X_{(i)})$ where h is a function of bounded variation. La Rocca (1995a) shows empirical likelihood ratio curves for trimmed means of Newcomb's passage time of light data, and presents some simulations. La Rocca (1996) obtains a χ^2 calibration for the difference between the trimmed means of two populations.

10.9 Exercises

Exercise 10.1 Suppose that X_i are independent random variables from a distribution with maximum value $\theta < \infty$. That is $\Pr(X \leq \theta) = 1$, but $\Pr(X \leq \theta - \epsilon) < 1$ for any $\epsilon > 0$. Let $\hat{\theta} = \max_{1 \leq i \leq n} X_i$. Describe how the empirical likelihood confidence region for θ fails. Invent and describe a nondegenerate approach to empirical likelihood for the sample maximum.

Exercise 10.2 Derive the values of L_1 and L_2 given in Chapter 10.1. For L_1 , when $c = X_i$ what is the value that w_i takes in order to maximize the likelihood? For L_2 , when $c = (X_i + X_j)/2$, what is the maximizing value of w_i ?

Exercise 10.3 Let $(X, Y) \in \mathbb{R}^2$ be from a distribution F that is thought to be symmetric about the line $x \cos \theta + y \sin \theta = r$.

- a) For $(x, y) \in \mathbb{R}^2$, find its reflection (\tilde{x}, \tilde{y}) under this symmetry.
- b) For $i = 1, \dots, n$, suppose that (X_i, Y_i) are sampled from a continuous distribution. Consider the $n + 2$ dimensional family of distributions given by parameters θ, r , and weights w_i summing to $1/2$ that apply equally to (X_i, Y_i) and $(\tilde{X}_i, \tilde{Y}_i)$. For almost all lines, the maximum over w_i of $\prod_i w_i$ is $(2n)^{-n}$. Describe geometrically any lines that give a larger likelihood. How much larger is that likelihood? Do not bother with any kind of line that is not sure to arise in the sample for large enough n .

Exercise 10.4 Suppose that F is known to be symmetric about the line $x + y = 0$ (for which $\theta = \pi/4$ and $r = 0$ in the notation of the previous question). Find an example estimating equation $E(m(X, Y, \theta)) = 0$, for each of the following cases:

- a) Imposing the symmetry constraint is degenerate.
- b) Imposing the symmetry constraint leaves empirical likelihood unchanged.
- c) Imposing the symmetry constraint with empirical likelihood gives something new.

Exercise 10.5 Construct a family of moment-like estimating equations to use for testing or imposing approximate symmetry of F about the line $x \cos \theta + y \sin \theta = r$.

Exercise 10.6 Now suppose that F is thought to be symmetric about the point (x_c, y_c) . Construct a set of moment-like estimating equations for imposing approximate symmetry of this kind.