Self-concordance for empirical likelihood

Art B. Owen
Stanford University

July 2012

Abstract

The usual approach to computing empirical likelihood for the mean uses Newton's method after eliminating a Lagrange multiplier and replacing the function $-\log(x)$ by a quadratic Taylor approximation to the left of $1/n$. This paper replaces the quadratic approximation by a quartic. The result is a self-concordant function for which Newton's method with backtracking has theoretical convergence guarantees.

1 Introduction

This paper presents an improved computational strategy for the empirical likelihood. The new algorithm is a damped Newton iteration applied to a convex and self-concordant function. Self-concordance controls the rate at which the second derivative of a function changes. It is a checkable sufficient condition to ensure that damped Newton iterations converge to the global solution and it also provides a computable criterion (the Newton decrement) by which to judge convergence.

The standard approach to empirical likelihood calculation replaces a constrained optimization over n parameters by an unconstrained dual optimization over d parameters. Here n is the number of observations and the number d of parameters is usually much smaller than n.

Recently Yang and Small (2012) presented five algorithms for computing the empirical log likelihood. They were motivated by some failures of a code written by the present author. That code computes the empirical likelihood when it exists and was designed to fail gracefully when the empirical log likelihood is $-\infty$, yielding a negative quantity with absolute value in the hundreds. Yang and Small (2012) found some anomalous behavior in the solutions at points where the empirical likelihood is almost undefined. They traced the cause to a Levenberg-Marquardt style stepping from the Newton direction towards the gradient direction. Their improvement replaced that search by a damped Newton style line search. Their code includes Davidon-Fletcher-Powell and BFGS optimizations. In addition to being less ad hoc than Levenberg-Marquardt line search methods are guaranteed to converge to the global optimum under certain conditions. Typically they require either a suitable starting point or a uniform bound on the condition of the Hessian matrix of the function to be optimized See Rheinboldt (1998) or Polyak (1987).

In this paper, a damped Newton method is retained but the dual function is changed in order to make it self-concordant. Self-concordance alone is a strong enough condition to ensure that the damped Newton algorithm converges to the global optimum (Boyd and Vandeberghe, 2004). It removes the need to check whether the Hessian matrix has bounded condition or the starting point is close enough to the solution.

Chen et al. (2012) show that a step reducing Newton method converges to a point where the gradient of the log likelihood has a suitably small norm. They use results from Polyak (1987) after showing that their log empirical likelihood has first and second derivatives satisfying certain bounds. Self concordance gives similar gaurantees and allows one to use the Newton decrement to bound the sub-optimality of an estimate.

An outline of this paper is as follows. Section 2 describes empirical likelihood. Section 3 describes the dual problem underlying most algorithms for computing empirical likelihood. Section 4 reviews basic properties of self-concordant functions and proves the main result: a quartic extension of the negative empirical log likelihood is self-concordant. It also shows one of the challenging data sets from Yang and Small (2012). Section 5 shows that we can even replace the entire negative log likelihood by a Taylor approximating quartic function and retain self-concordance. The quartic log empirical likelihood function is known to be Bartlett correctable (Corcoran, 1998).

2 Empirical likelihood

The empirical likelihood is a nonparametric likelihood ratio technique suitable for generating confidence regions and tests. It is based on an analogue of Wilks' theorem but does not require the data to be sampled from any known parametric family of distributions.

Given IID data $X_1, \dots, X_n \in \mathbb{R}^d$, the profile empirical likelihood function for the mean is

$$\mathcal{R}(\mu) = \max\left\{ \prod_{i=1}^n nw_i \mid 0 \le w_i, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i X_i = \mu \right\}. \tag{1}$$

Asymptotically $-2\log\mathcal{R}(\mu_0) \xrightarrow{\mathrm{d}} \chi^2_{(r)}$ holds where $\mu_0 = \mathbb{E}(X_i)$ and r is the rank of $\mathrm{var}(X_i)$ (usually d).

Extensions to parameters other than the mean are based on estimating equations. For a parameter θ defined by estimating equations $\mathbb{E}(m(X,\theta)) = 0$ the usual plug-in estimator is $\hat{\theta}$ defined by $(1/n)\sum_{i=1}^n m(X_i, \hat{\theta}) = 0$. The empirical likelihood for θ is

$$\mathcal{R}(\theta) = \max\biggl\{\prod_{i=1}^n nw_i \mid 0 \le w_i, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(X_i,\theta) = 0\biggr\}. \tag{2}$$

Empirical likelihood tests have very competitive power (Kitamura, 2003), even when there is a parametric likelihood that one could have used (Lazar and Mykland, 1998). For $d > 1$, a confidence region is defined by a region (not just two endpoints) and empirical likelihood automatically determines the shape of this region in a way that Hall (1990) shows is correct to high order. Finally, when there is side information of the form $\mathbb{E}(h(X,\theta)) = 0$ for some function $h(\cdot,\cdot)$ one can exploit this fact to obtain sharper confidence regions and tests (Owen, 1991; Qin and Lawless, 1994). For these and other facts about empirical likelihood see Owen (2001).

3 Empirical likelihood optimization

Here we describe the optimization problem required to compute the empirical likelihood for a mean. This account summarizes details from Owen (2001). Suppose that μ is an interior point of the convex hull of $X_1,\dots,X_n \in \mathbb{R}^d$. In case $X_1,\dots,X_n$ lie in an affine subspace of dimension less than d, we take this to mean that μ is in the relative interior of the convex hull. Then we call μ an interior point.

To optimize the empirical likelihood (1) we maximize $\log(\mathcal{R}(\mu))$. For this we construct the Lagrangian

$$G = \sum_{i=1}^n \log(nw_i) - n\lambda^\mathsf{T} \sum_{i=1}^n w_i(X_i - \mu) + \delta\Bigl(\sum_{i=1}^n w_i - 1\Bigr)$$

where $-n\lambda \in \mathbb{R}^d$ and $\delta \in \mathbb{R}$ are Lagrangian multipliers. Setting $\partial G/\partial w_i$ to zero yields

$$0 = \frac{1}{w_i} - n\lambda^\mathsf{T}(X_i - \mu) + \delta. \tag{3}$$

Multiplying (3) by w_i and summing over i, we find that $\delta = -n$ when μ is an interior point. We may then eliminate that multiplier, and obtain

$$w_i = \frac{1}{n}\frac{1}{1 + \lambda^\mathsf{T}(X_i - \mu)},$$

where λ satisfies

$$\sum_{i=1}^n \frac{X_i - \mu}{1 + \lambda^\mathsf{T}(X_i - \mu)} = 0. \tag{4}$$

The left side of equation (4) is the gradient with respect to λ of $-f$, given by

$$f(\lambda) = -\sum_{i=1}^{n} \log(1 + \lambda^{\mathsf{T}} Z_i)$$

for $Z_i = X_i - \mu$. The function f is convex on $\{\lambda \in \mathbb{R}^d \mid 1 + \lambda^{\mathsf{T}} Z_i > 0,\ i = 1, \dots, n\}$. This set is convex and non-empty (it contains the origin). The minimizer of f over λ recovers the maximizing weights w_i of the empirical log likelihood.

When μ is an interior point, then there is a solution to $\sum_{i=1}^n w_i X_i = \mu$ with $w_i > 0$ and $\sum_{i=1}^n w_i = 1$. The optimal weight vector must also have $w_i > 0$ or else it would have $\mathcal{R} = 0$. Because the solution has all $w_i > 0$ and $\sum_i w_i = 1$, we conclude that $\max_i w_i < 1$ at the solution. Accordingly $1 + \lambda^{\mathsf{T}} Z_i > 1/n$ holds at the solution.

The optimization strategy in Owen (2001) uses the function

$$\log_\star(x) = \begin{cases} \log(x), & x > 1/n \\ \log(1/n) - 3/2 + 2nx - (nx)^2/2, & x \le 1/n. \end{cases} \tag{5}$$

This $\log_\star$ function is a second degree Taylor approximation to the logarithm at the point $1/n$. Using this construction

$$f_\star(\lambda) \equiv -\sum_{i=1}^{n} \log_\star(1 + \lambda^{\mathsf{T}} Z_i)$$

is convex on all of $\lambda \in \mathbb{R}^d$ but has the same minimum as f when μ is an interior point. Thus we may minimize $f_\star$ without first checking whether μ is an interior point.

In the event that μ is not an interior point, the function $f_\star$ is still convex on all of $\mathbb{R}^d$, but is unbounded below. Then algorithms based on Newton's method diverge. Empirically, $\|\lambda\| \to \infty$, along with $\|\nabla f_\star(\lambda)\| \to 0$. One can then stop the algorithm when an upper limit on the number of steps is reached or when a lower limit on the norm of the gradient vector is reached, or whichever comes first, and declare $\mathcal{R}(\mu) = -\infty$. Inspecting the solutions in such cases we find $\sum_{i=1}^n w_i < 1$. Usually the sum of the weights is near zero. If μ is on a face of the convex hull of $X_1, \dots, X_n$, then the number of non-negligible w_i is (in the author's experience) typically the dimension of that face (e.g., 1 for a vertex, 2 for an edge, and so on).

4 Self concordance

A convex function $g(x)$ on $x \in \mathbb{R}$ is self-concordant if it has three derivatives and $|g'''(x)| \le 2g''(x)^{3/2}$. A function $g(\boldsymbol{x})$ on $\boldsymbol{x} \in \mathbb{R}^d$ is self-concordant if $g(\boldsymbol{x}_0 + t\boldsymbol{x}_1)$ is a self-concordant function of t for all $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{R}^d$. Furthermore there is a computable quantity (the Newton decrement) that when small enough, yields a guaranteed lower bound for the minimum of the objective function. These results are due to Nesterov and Nemirovskii (1994). The description here is based on Boyd and Vandeberghe (2004, Chapter 9).

For optimization problems it is enough to have

$$|g'''(x)| \le C g''(x)^{3/2} \tag{6}$$

hold for some $C < \infty$. Then $(C^2/4)g(x)$ is self-concordant using the original constant $C = 2$ and of course $(C^2/4)g$ has the same minimizer as g.

Here we switch to the notation commonly used in optimization problems, replacing λ by $\boldsymbol{x}$. The negative empirical log likelihood takes the form

$$f(\boldsymbol{x}) = -\sum_{i=1}^{n} \log(1 + Z_i^{\mathsf{T}} \boldsymbol{x})$$

over $\boldsymbol{x} \in \mathbb{R}^d$, where $Z_i \in \mathbb{R}^d$ are given by the estimating equations. Usually $Z_i = Z_i(\theta_0)$ for a null value of a parameter θ. The function $-\log(x)$ is self-concordant on $(0, \infty)$. Therefore $-\log(1 + Z_i^{\mathsf{T}} \boldsymbol{x})$ is self-concordant on $\boldsymbol{x} \in \mathbb{R}^d$ such that $1 + Z_i^{\mathsf{T}} \boldsymbol{x} > 0$. Self-concordance is preserved under summation, and so $f(\boldsymbol{x})$ is self-concordant on its domain $\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \min_i 1 + Z_i^{\mathsf{T}} \boldsymbol{x} > 0\} \subset \mathbb{R}^d$. In this section we extend $f(\boldsymbol{x})$ to a function that is self-concordant on all of $\mathbb{R}^d$. When 0 is an interior point of $Z_1, \dots, Z_n$ the extension has the same minimizer as f.

Note that the function $-\log_\star$ from Section 3 is not self-concordant. It is self-concordant on $(-\infty, 1/n)$ and it is also self-concordant on $(1/n, \infty)$. It fails to be self-concordant on $\mathbb{R}$ because $-\log_\star'''(x)$ does not exist at $x = 1/n$.

4.1 Self-concordant approximate negative logarithm

In Section 3 we considered the function $\log_\star$ which was a Taylor approximation to the logarithm at the point $1/n$ keeping terms up to the quadratic. Here we work with Taylor approximations to $-\log$ at the point ϵ keeping polynomial terms up to degree k.

For $\epsilon > 0$ and an integer $k \ge 0$, let

$$L_k(x) = L_k(x; \epsilon) = \begin{cases} -\log x, & x \ge \epsilon \\ h_k(x - \epsilon), & x < \epsilon, \end{cases} \tag{7}$$

where

$$h_k(y) = h_k(y; \epsilon) = -\sum_{t=0}^{k} \log^{(t)}(\epsilon) \frac{y^t}{t!}. \tag{8}$$

The function L_k has k continuous derivatives.

The function L_2 is quadratic and convex, but does not have a third derivative at $x = \epsilon$, so it cannot be self-concordant. The function L_3 is not convex, so it is not self-concordant either. We show here that L_4 is convex and self-concordant with $C = 2$. The Taylor approximation h_4 is also self-concordant on $\mathbb{R}$, but with $C = 98/25 = 3.92$ (as in equation (6)). That is $(C^2/4)h_4$ is self-concordant in the usual sense as is $4h_4$.

We will need the derivatives of h_k. For $r \in \{0, 1, \dots, k\}$,

$$h_k^{(r)}(y) = -\sum_{t=0}^{k-r} \log^{(t+r)}(\epsilon)\frac{y^t}{t!}.$$

For $t > 0$, $-\log^{(t)}(\epsilon) = (-1)^t\epsilon^{-t}(t-1)!$ and $-\log^{(0)}(\epsilon) = -\log(\epsilon)$. For $r > 0$,

$$\begin{aligned}
h_k^{(r)}(y) &= \sum_{t=0}^{k-r}(-1)^{t+r}\epsilon^{-t-r}y^t\frac{(t+r-1)!}{t!} \\
&= (-\epsilon)^{-r}\sum_{t=0}^{k-r}\frac{(t+r-1)!}{t!}\Bigl(\frac{-y}{\epsilon}\Bigr)^t.
\end{aligned}$$

For $k = 4$,

$$\begin{aligned}
h_4''(y) &= \epsilon^{-2}\Bigl(1 - 2\frac{y}{\epsilon} + 3\bigl(\frac{y}{\epsilon}\bigr)^2\Bigr) = \epsilon^{-2}\Bigl(\bigl(1 - \frac{y}{\epsilon}\bigr)^2 + \bigl(\frac{y}{\epsilon}\bigr)^2\Bigr), \quad \text{and} \\
h_4'''(y) &= \epsilon^{-3}\Bigl(-2 + 6\frac{y}{\epsilon}\Bigr).
\end{aligned}$$

Theorem 1. *For any $\epsilon > 0$, the function $L_4(x)$ given by equation (7) with $k = 4$ is self-concordant on $\mathbb{R}$.*

Proof. For $x \geq \epsilon$ we have $|L_k'''(x)| \leq L_k''(x)^{3/2}$ for $k \geq 3$ because the logarithm is self-concordant. For self-concordance of L_4, we also need $|h_4'''(y)| \leq 2h_4''(y)^{3/2}$ to hold for all $y \leq 0$. Self-concordance of $h_4(\cdot)$ is equivalent to self-concordance of $h_4(\epsilon \times \cdot)$. For $z \leq 0$, define

$$\rho(z) = \frac{|h_4'''(z\epsilon)|}{h_4''(z\epsilon)^{3/2}} = \frac{2 - 6z}{D(z)^{3/2}}$$

where $D(z) = (z-1)^2 + z^2$. The derivative

$$\begin{aligned}
\rho'(z) &= \frac{-6D(z)^{3/2} - (2-6z)(3/2)D(z)^{1/2}(6z-2)}{D(z)^3} \\
&= \frac{-6(z-1)^2 - 6z^2 + 6(3z-1)^2}{D(z)^{5/2}} \\
&= \frac{-6z(4-7z)}{D(z)^{5/2}} \qquad (9)
\end{aligned}$$

is non-negative for $z \leq 0$. Therefore $\rho(z) \leq \rho(0) = 2$ on $(-\infty, 0]$ and so L_4 is self-concordant. □

Theorem 1 shows that L_4 is self-concordant on $(-\infty, \epsilon]$ because h_4 is self-concordant on $(-\infty, 0]$. This allows us to replace the negative logarithm whose domain is $(0, \infty)$ by the piece-wise defined self-concordant function L_4 whose domain is $\mathbb{R}$. Interestingly, we can do more. The fourth degree Taylor approximation to $-\log$ at the point ϵ can be scaled to self-concordance on all of $\mathbb{R}$. The constant C necessary to the right of ϵ is somewhat larger than the $C = 2$ that is needed to the left of ϵ:

Theorem 2. *For any $\epsilon > 0$ the function $h_4(y)$ given by* (8) *with $k = 4$ satisfies*

$$|h_4'''(y)| \le \frac{98}{25} h_4''(y)^{3/2} = 3.92 h_4''(y)^{3/2}.$$

Proof. For $y > 0$, we proceed as in the proof of Theorem 1, with $\rho(z) = (2 - 6z)D(z)^{-3/2}$ for $z \le 1/3$ and $\rho(z) = (6z - 2)D(z)^{-3/2}$ for $z \ge 1/3$. On the interval $[0, 1/3]$, $\rho'(z)$ is given by equation (9) and it is not positive there. Therefore the maximum of ρ on $[0, 1/3]$ is $\rho(0) = 2$.

On the interval $z \ge 1/3$,

$$\rho'(z) = \frac{6z(4 - 7z)}{D(z)^{5/2}} \tag{10}$$

which vanishes at $z = 4/7$ and is negative thereafter. It follows that the largest value of $\rho(z)$ for $z \in \mathbb{R}$ is

$$\rho\Big(\frac{4}{7}\Big) = \frac{6 \times 4/7 - 2}{D(4/7)^{3/2}} = \frac{98}{25} = 3.92. \quad \square$$

4.2 Backtracking and the Newton decrement

An unguarded Newton's method for minimizing $f(\boldsymbol{x})$ proceeds via updates $\boldsymbol{x} \leftarrow \boldsymbol{x} + \Delta\boldsymbol{x}$ for $\Delta\boldsymbol{x} = -(\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})$ where ∇ and ∇^2 denote the gradient and Hessian respectively. This method converges under mild conditions if started near enough the solution, but in practice it is hard to know whether a given starting point is near enough.

A backtracking line search replaces the update $\Delta\boldsymbol{x}$ by a shorter vector if f does not decrease sufficiently. Given $\alpha \in (0, 0.5)$ and $\beta \in (0, 1)$ the algorithm starts with $t = 1$ and replaces t by βt until $f(\boldsymbol{x} + t\Delta\boldsymbol{x}) \le f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^\mathsf{T} \Delta\boldsymbol{x}$. See Boyd and Vandeberghe (2004, Algorithm 9.2, p. 464) for backtracking and Boyd and Vandeberghe (2004, Algorithm 9.5, p. 487) for Newton's method incorporating backtracking.

Newton iterations with back-tracking are provably effective on self-concordant functions. The self concordance property yields a bound on the number of Newton steps required to minimize a function. That bound depends on the gap between the initial and minimal value of the objective function, so it is not very useful. But self concordance does supply a usable stopping criterion, based on the Newton decrement. The Newton decrement is

$$\nu = \nu(\boldsymbol{x}) = \big(\nabla f(\boldsymbol{x})^\mathsf{T} (\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})\big)^{1/2}.$$

If f is a strictly convex self-concordant function, and $\nu(\tilde{\boldsymbol{x}}) \le 0.68$, then $\inf_{\boldsymbol{x}} f(\boldsymbol{x}) \ge f(\tilde{\boldsymbol{x}}) - \nu(\tilde{\boldsymbol{x}})^2$ (Boyd and Vandeberghe, 2004, Equation (9.50)). Thus stopping when $\nu(\boldsymbol{x}) < \epsilon < 0.68$ ensures that the objective function is within ϵ^2 of the minimum when f is self-concordant. This bound is necesarily conservative, but not by much. The quantity $\epsilon^2/2$ is often used as an estimate of $f(\boldsymbol{x}) - \inf_{\boldsymbol{x}} f(\boldsymbol{x})$.

If $(C^2/4)f$ is self concordant for $C > 2$, then a similar guarantee holds. The Newton decrement for $(C^2/4)f$ is the same as that for f. So $(C^2/4)f$ is within ϵ^2 of its minimum when $\nu(\boldsymbol{x}) < \epsilon < 0.68$ and hence f is within $4\epsilon^2/C^2$ of its minimum.

4.3 An example from Yang and Small (2012)

Yang and Small (2012) encountered numerical difficulties with empirical likelihood in an instrumental variables model. They needed to compute the empirical likelihood for a mean of zero in four dimensions where the variables were $Z_1(Y - \beta_1 W - \alpha_1)$, $Y - \beta_1 W - \alpha_1$, $Z_2(Y - (\beta_1 + \delta)W - \alpha_2)$, and $Y - (\beta_1 + \delta)W - \alpha_2$. The variables Z_1 and Z_2 are binary instrumental variables. The other factors are residuals. Their model specified values for the parameters α_1, β_2, α_2 and δ. For the meaning of these variables, see Yang and Small (2012). They encountered difficulties when profiling β_1 over a series of values with fixed levels of the other parameters on a bootstrap resample of the underlying data.

Figure 1 shows pairwise scatterplots of the four variables at one value of β_1. That value is $\beta_1 = 1.84$ (others are similar). The hypothesized mean of 0 is just barely inside the convex hull of the data. In particular the fifth plot, shows that the two kinds of residuals lie very nearly on a straight line. On closer inspection they fall onto two very close parallel lines (W is binary) with the origin in between. The data matrix depicted in Figure 1 is given as an R file at `stat.stanford.edu/~owen/reports/e85samp`. To use it within R, save the file in a directory, run R in that directory, and use the R command `load`("e85samp").

By using a backtracking Newton method, Yang and Small (2012) were able to compute the empirical likelihood for this case. They obtain a log empirical likelihood of -339.6937 for this data in 9 Newton steps with a Newton decrement of 6.74277×10^{-16}. Using backtracking along with the self-concordant version of the empirical log likelihood yields identical values to this level of precision. Both algorithms used the same backtracking parameters $\alpha = 0.3$ and $\beta = 0.8$ and the same start of 0 for the Lagrange multiplier.

5 Quartic log likelihood

Corcoran (1998) considered replacing $-\log$ by its fourth order Taylor approximation around the point $\epsilon = 1$. The resulting alternative nonparametric likelihood function preserves many of the properties of empirical likelihood, including Bartlett correctability. Theorem 2 shows that $3.92^2/4$ times this alternative log likelihood is self-concordant.

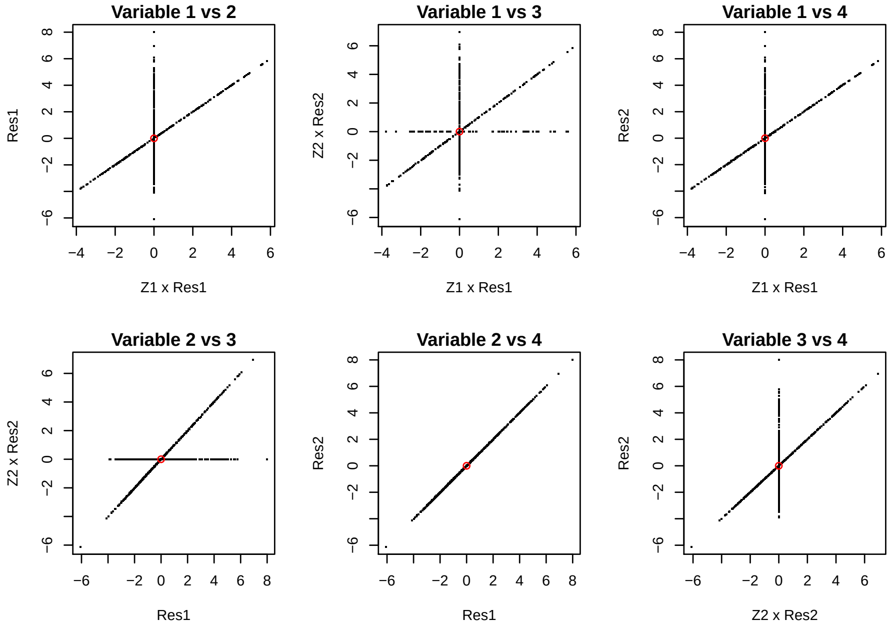

Figure 1: Pairwise scatterplots of four variables in the instrumental variables model of Yang and Small (2012). They hypothesized mean of 0 is shown as a circle. The 1000 data points are plotted as dots.

We can thus construct the quartic empirical log likelihood problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} h_4(nw_i - 1; 1), \\ \text{subject to} \quad & \sum_{i=1}^{n} w_i = 1, \\ & \sum_{i=1}^{n} w_i X_i = \mu \end{aligned} \tag{11}$$

and define $\log(\mathcal{R}_Q(\mu))$ to be the minimizing value. This problem is a linearly constrained convex minimization. There are now n variables given by $\boldsymbol{w} \in \mathbb{R}^n$. The vector $\boldsymbol{w}$ is feasible as long as it satisfies the linear constraints, because the objective function is finite for any $\boldsymbol{w}$. There is always a feasible vector $\boldsymbol{w}$, so long as μ is in the affine hull of $X_1, \dots, X_n$.

When, as usual, the X_i span all of $\mathbb{R}^d$, then there is a feasible vector $\boldsymbol{w}$. Confidence regions based on the quartic empirical log likelihood can extend beyond the convex hull of the data. For any point μ outside the convex hull of X_i, the corresponding vector $\boldsymbol{w}$ must contain at least one negative element.

5.1 Convexity of the quartic empirical likelihood regions

The quartic empirical likelihood ratio confidence regions are of the form

$$\{\mu \in \mathbb{R}^d \mid f(\mu) \le F_*\} \tag{12}$$

for a critical value $F_* < \infty$. The quartic empirical likelihood ratio regions need not be nested within the convex hull of the data. They therefore share this critical property that motivated the adjusted empirical likelihood of Chen et al. (2008). Convexity of adjusted empirical likelihood confidence regions was recently established by Chen and Huang (2012). Quartic empirical likelihood confidence regions are easily seen to be convex. The proof holds for functions more general than h_4.

Theorem 3. *Let $h(\cdot)$ be a convex function on $\mathbb{R}$ and let $X_1, \dots, X_n \in \mathbb{R}^d$ for integers $n \ge 1$ and $d \ge 1$. Define*

$$\mathcal{H}(\mu) = \max\Bigl\{\sum_{i=1}^n h(nw_i - 1) \mid \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i X_i = \mu\Bigr\},$$

and

$$\mathcal{C}(\tau) = \{\mu \in \mathbb{R}^d \mid \mathcal{H}(\mu) \le \tau\}.$$

Then $\mathcal{C}(\tau)$ is a convex set.

Proof. If $\mathcal{C}(\tau)$ has fewer than two points, then it is convex. Otherwise, choose any $\mu, \mu' \in \mathcal{C}(\tau)$. Let w_i satisfy $\sum_{i=1}^n w_i = 1$, $\sum_{i=1}^n w_i X_i = \mu$, and $\sum_{i=1}^n h(nw_i - 1) \le \tau$. Similarly let w_i' satisfy $\sum_{i=1}^n w_i' = 1$, $\sum_{i=1}^n w_i X_i = \mu'$, and $\sum_{i=1}^n h(nw_i' - 1) \le \tau$. For $0 < \theta < 1$ let $\widetilde{\mu} = \theta\mu + (1-\theta)\mu'$ and $\widetilde{w}_i = \theta w_i + (1-\theta)w_i'$. Then $\sum_{i=1}^n \widetilde{w}_i = 1$ and $\sum_{i=1}^n \widetilde{w}_i = \theta\mu + (1-\theta)\mu'$ and

$$\begin{aligned}\sum_{i=1}^n h(n\widetilde{w}_i - 1) &= \sum_{i=1}^n h(\theta(nw_i - 1) + (1-\theta)(nw_i' - 1))\\ &\le \theta \sum_{i=1}^n h(nw_i - 1) + (1-\theta)\sum_{i=1}^n h(nw_i' - 1)\\ &\le \tau.\end{aligned}$$

It follows that $\widetilde{\mu} \in \mathcal{C}(\tau)$ as well and so $\mathcal{C}(\tau)$ is convex. □

5.2 Computation for quartic empirical likelihood

The quartic empirical log likelihood problem (11) is a convex optimization in n variables with $d+1$ linear constraints. For empirical likelihood, we could eliminate the Lagrange multiplier corresponding to $\sum_{i=1}^n w_i = 1$ by summing w_i times the gradient of the Lagrangian with respect to w_i. That multiplier cannot be eliminated for the present problem. In the example of the next section an equality constrained convex optimization using a primal-dual algorithm (Boyd and Vandeberghe, 2004, Algorithm 10.2) was used.

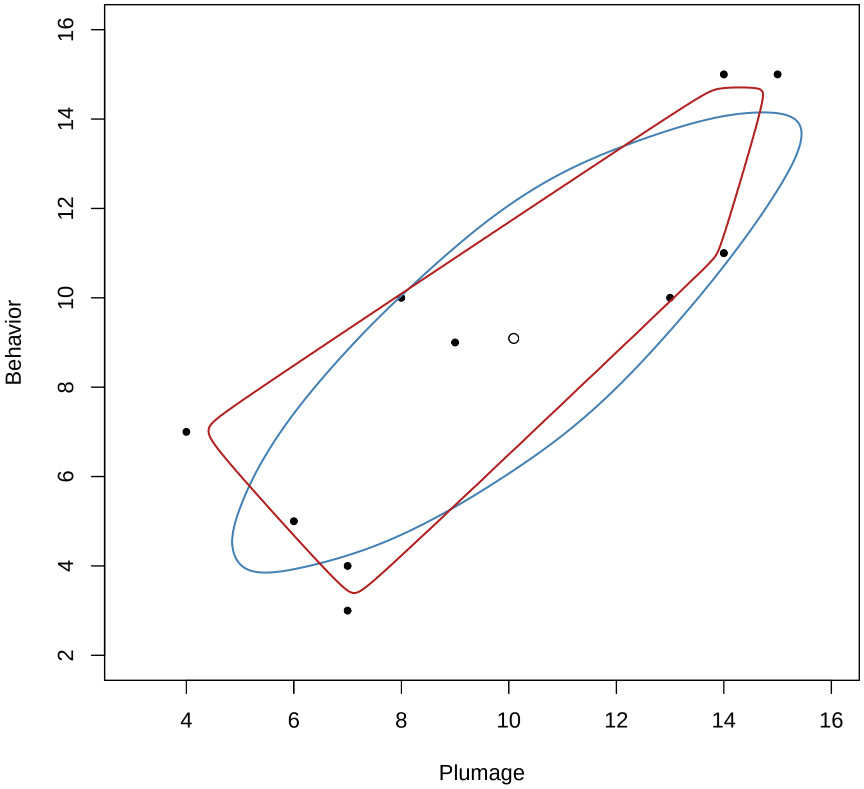

Figure 2: Empirical likelihood (nearly piecewise linear) and quartic empirical likelihood (smooth) contours for the mean of the given data points. An open circle marks the sample mean.

5.3 Example of quartic empirical likelihood

Figure 2 shows the duck data of Larsen and Marx (1986). Subjective plumage and behavior scores were obtained for 11 ducks. The figure shows extreme empirical likelihood and quartic empirical likelihod contours for the mean. They are both at a nominal coverage level of $1 - 10^{-10}$ from a $\chi^2_{(2)}$ calibration for $-2\log(\mathcal{R})$. The contours for empirical likelihood are close to the convex hull of the data. The quartic empirical likelihood contours are almost ellipsoidal and extend out of the convex hull.

While the quartic empirical likelihood confidence region extends outside of the convex hull, it does not extend very far out of it. Both of these regions have roughly the same area and neither is plausible as a $1-10^{-10}$ confidence region. The Hotelling's T^2 region for 0.99 confidence has roughly the same size as these regions. It is thus not reasonable to expect the usual χ^2 calibration to be effective at small sample sizes for the quartic empirical likelihood. Instead, some technique used for improved calibration of empirical likelihood, such as bootstrap calibration or adding extra points will be needed.

6 Conclusions

This paper has shown that a quartic extension to the logarithm yields a convex self-concordant objective function which is equivalent to minus the empirical log likelihood. Self-concordance implies gauranteed convergence for back-tracking Newton methods. Without self-concordance, convergence depends on hard to verify properties of the objective function. This does not necessarily mean that backtracking with the usual quadratic extension will fail often. Indeed it may require special data circumstances for difficulties to arise even with Levenberg-Marquardt style step reductions. Self-concordance does however allow one to translate a desired accuracy in the log likelihood into a convergence criterion. Finally, the quartic log empirical likelihood studied by Corcoran (1998) is self-concordant and yields convex confidence regions for the mean.

Acknowledgments

I thank Dylan Small and Wang Dan for helpful comments and Jiahua Chen for sharing an early version of Chen and Huang (2012). This work was supported by the U.S. NSF under grant DMS-0906056.

References

Boyd, S. and Vandeberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Chen, J. and Huang, Y. (2012). Finite-sample properties of adjusted empirical likelihood. Technical report, University of British Columbia.

Chen, J., Sitter, R. R., and Wu, C. (2012). Using empirical likelihood methods to obtain range restricted weightes in regression estimators for surveys. *Biometrika*, 89(1):230–237.

Chen, J., Variyath, A. M., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 2:426–443.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85:967–972.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics*, 18:121–140.

Kitamura, Y. (2003). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 69(6):1661–1672.

Larsen, R. J. and Marx, M. L. (1986). *An introduction to mathematical statistics and its applications.* Prentice-Hall, Englewood Cliffs, NJ.

Lazar, N. and Mykland, P. A. (1998). An evaluation of the power and conditionality properties of empirical likelihood. *Biometrika*, 85:523–534.

Nesterov, Y. and Nemirovskii, A. (1994). *Interior-point polynomial methods in convex programming.* Society for Industrial and Applied Mathematics, Philadelphia.

Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19:1725–1747.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRCpress, Boca Raton, FL.

Polyak, B. T. (1987). *Introduction to Optimization.* Optimization Software Inc., New York.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325.

Rheinboldt, W. C. (1998). *Methods for solving systems of nonlinear equations.* Society for Industrial and Applied Mathematics, Philadelphia, second edition.

Yang, D. and Small, D. S. (2012). An R package and a study of methods for computing empirical likelihood. *Journal of Statistical Computing and Simulation.* (to appear).