

Ec241a

Econometrics

Spring 2004

UC Berkeley Department of Economics

## MAXIMUM LIKELIHOOD ESTIMATION I: DURATION MODELS:

## LIKELIHOOD FUNCTIONS AND COMPUTATIONAL ISSUES (LANCASTER, LUENBERGER)

Lancaster (1979) is interested in determining “the causes of variation between unemployed persons in the length of time they are out of work .... bearing as it does upon the design and effect of welfare policy.” He has data on unemployment durations of 479 unskilled workers, as well as some of their individual characteristics such as age, the local unemployment rate and the replacement ratio, measured as “how much they had coming in from all these sources (unemployment benefit, supplementary benefit, and family income supplement) during the main period of their unemployment”, divided by the answer to the question “how much did you earn, after deductions, in your last job.” Especially the coefficient on the last variable is viewed as relevant for social policy.

The economic theory underlying Lancaster’s analysis is job search theory. An unemployed individual is assumed to receive job offers, arriving according to some rate  $\lambda(t)$ , such that the expected number of job offers arriving in a short interval of length  $dt$  is  $\lambda(t)dt$ . Each offer consists of some wage rate  $w$ , drawn independently of previous wages, from some distribution with distribution function  $F(w)$ . The offer is compared to some reservation wage  $\bar{w}(t)$ , and if the offer is better than the reservation wage, that is with probability  $1 - F(\bar{w}(t))$ , the offer is accepted. The reservation wage is set to maximize utility. Suppose that the arrival rate is constant over time. In that case the optimal reservation wage is also constant over time, and the probability of receiving an acceptable offer in a short period of time  $dt$  is  $\theta dt$ , with  $\theta = \lambda \cdot (1 - F(\bar{w}))$  (e.g., Karlin, 1962). The constant acceptance rate  $\theta$  implies that the distribution for the unemployment duration is exponential with mean  $1/\theta$ , and probability density function

$$f(y) = \theta \exp(-y\theta).$$

This distribution is widely used for durations of various types, e.g., life times, unemployment durations, failure times, etcetera.

The mean and variance for this distribution are  $1/\theta$  and  $1/\theta^2$  respectively. The expected value of the remaining duration conditional on survival up to  $c$ ,  $E[Y - c | Y > c] = 1/\theta$ , and does not depend on the elapsed duration. This is known as the lack of memory property. One minus the distribution function,  $S(y) = 1 - F(y) = \exp(-y\theta)$ , is known as the survivor function. The ratio of the density and the survivor function is the hazard or failure rate (in our example the rate at which a job is offered and accepted):

$$h(y) = \lim_{dy \downarrow 0} \frac{\Pr(y < Y < y + dy)}{\Pr(y < Y)} = \frac{f(y)}{S(y)} = \theta.$$

For the exponential distribution this rate is constant. Intuitively it means that the probability of finding a job on the 51st day, conditional on not having found a job in the first 50 days, is the same as the probability of finding a job on the first day. Whether this is reasonable in practice is an important question, and in the case of unemployment durations of great interest for policy purposes. We shall return to this question later.

Given a parametric distribution for the unemployment duration,  $f(y|\theta)$ , with implied survivor function  $S(y|\theta)$  and hazard function  $h(y|\theta)$ , let us look at some possible sampling schemes and the implied likelihood functions.

### 1. (OBSERVATION OF EXACT FAILURE TIMES)

The simplest case is if we observe for a random sample of people entering unemployment the exact unemployment durations. In that case the likelihood function is the product of the density functions:

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(y_i|\theta) = \prod_{i=1}^N h(y_i|\theta) \cdot S(y_i|\theta),$$

where the second representation uses the equality of the density and the product of the hazard and survivor functions.

## 2. (OBSERVATION OF INDICATOR OF SURVIVAL)

Suppose we observe a number of people becoming unemployed, but we only observe whether they exited unemployment before a fixed point in time, say  $c$ . In that case the likelihood function is, using  $d_i = 1$  to denote that individual  $i$  left employment before time  $c$  and  $d_i = 0$  to denote this individual was still unemployed at time  $c$ :

$$\mathcal{L}(\theta) = \prod_{i=1}^N F(c|\theta)^{d_i} \cdot (1 - F(c|\theta))^{1-d_i} = \prod_{i=1}^N (1 - S(c|\theta))^{d_i} \cdot S(c|\theta)^{1-d_i}.$$

## 3. (OBSERVATION OVER FIXED PERIOD OF TIME)

A variation of this observation scheme occurs if we observe the exact exit or failure time if it occurs before  $c$ , but only an indicator if exit occurs after  $c$ :

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(y_i|\theta)^{d_i} \cdot S(c|\theta)^{1-d_i} = \prod_{i=1}^N h(y_i|\theta)^{d_i} \cdot S(y_i|\theta)^{d_i} \cdot S(c|\theta)^{1-d_i}.$$

Let  $t_i = \min(y_i, c) = d_i \cdot y_i + (1 - d_i) \cdot c$  be the minimum of the actual duration and the censoring time, so the likelihood can be written as

$$\mathcal{L}(\theta) = \prod_{i=1}^N h(t_i|\theta)^{d_i} \cdot S(t_i|\theta).$$

Two naive solutions to this problem of right censoring are not adequate. First, one can pretend the observations censored at  $c$  in fact exited at  $c$ . Second, one can discard the observations who did not exit before  $c$ . To see the biases from these “solutions”, let us look at the exponential case in a little more detail. In that case  $f(y|\theta) = \theta \exp(-y\theta)$ , so the likelihood function simplifies to:

$$\mathcal{L}(\theta) = \theta^{\sum_{i=1}^N d_i} \cdot \exp\left(-\sum_{i=1}^N t_i \theta\right).$$

The maximum likelihood estimator is

$$\hat{\theta}_{mle} = \sum_{i=1}^N d_i / \sum_{i=1}^N t_i = 1 / (\bar{t} / \bar{d}) = \bar{d} / \bar{t}.$$

The first naive estimator ignores the fact that observations with  $d_i = 0$  were censored, and estimates  $\theta$  as  $\tilde{\theta} = 1/\bar{t}$ . This leads to an over-estimate of  $\theta$ , as the numerator

increases from  $\bar{d}$  to 1. The second naive estimator, which discards the censored observations, estimates  $\theta$  as  $\bar{\theta} = \sum d_i / \sum d_i t_i$ . This also leads to an over-estimate as the denominator decreases from  $\sum_{i=1}^N t_i$  to  $\sum_{i=1}^N d_i \cdot t_i$ .

#### 4. (GENERAL RIGHTHAND CENSORING)

More generally the censoring time when we cease to observe unit  $i$  can differ from individual to individual. Assuming it is independent of the failure or exit time, and again letting  $t_i$  denote the minimum of the exit time  $y_i$  and the censoring time  $c_i$ , the likelihood function is:

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(t_i|\theta)^{d_i} \cdot S(t_i|\theta)^{1-d_i} = \prod_{i=1}^N h(t_i|\theta)^{d_i} \cdot S(t_i|\theta).$$

#### 5. (STOCK SAMPLING)

In all four sampling schemes so far we assumed we start observing the individuals at the time of entry into unemployment. This is also known as flow sampling. An alternative arises if we sample from the stock of unemployed. For example, we draw someone from the stock of unemployed who has been unemployed for  $s_i = 4$  weeks, and who eventually finds a job after 6 more weeks, leading to a total duration of  $y_i = 10$  weeks. In general, let  $s_i$  be the incomplete duration of the unemployment spell for individual  $i$  at the time the individual is first observed. If we continue observing each individual till he or she exits unemployment with a total duration of  $y_i$ , the likelihood conditional on the incomplete duration is:

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(y_i|\theta)/S(s_i|\theta) = \prod_{i=1}^N h(y_i|\theta) \cdot \frac{S(y_i|\theta)}{S(s_i|\theta)}.$$

The other likelihood functions (with right censoring) can similarly be accommodated to allow for stock sampling. More complicated issues arise if we do not observe how long an individual has already been unemployed at the point where we start observing him or her. To deal with this problem of left censoring we need to make assumptions regarding the flows into, and out, of unemployment.

Let  $\{Z\}_{i=1}^{\infty}$  be a sequence of independent and identically distributed random variables with common cumulative distribution function  $F(z|\theta_0)$ . Either the  $Z_i$  are discrete random variables and  $f(z|\theta_0)$  is the probability  $Pr(Z_i = z)$ , or the  $Z_i$  are continuous random variables and  $f(z|\theta_0)$  is the probability density function.

We know the function  $F(z|\theta_0)$ , but we do not know the value of the parameter  $\theta_0$ , other than that it is in some set  $\Theta$ , a subset of  $R^K$ . The problem we face is that of estimating  $\theta_0$  given  $N$  observations  $z_1, z_2, \dots, z_N$ .

The estimator we consider here is the maximum likelihood estimator, or mle, defined as the argmax of the log likelihood function:

$$\hat{\theta}_{mle} = \operatorname{argmax}_{\theta \in \Theta} L(\theta) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^N \ln f(z_i|\theta).$$

We study the following questions:

1. (CONSISTENCY)

Under what conditions is the mle consistent for  $\theta_0$ ? More formally, under what conditions is it true that for all  $\varepsilon > 0$ ,

$$\lim_{N \rightarrow \infty} Pr(|\hat{\theta}_{mle} - \theta_0| > \varepsilon) = 0.$$

2. (ASYMPTOTIC NORMALITY)

Under what additional conditions is it true that

$$\sqrt{N}(\hat{\theta}_{mle} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, -\left[E \frac{\partial^2}{\partial \theta \partial \theta'}(Z_i, \theta_0)\right]^{-1}\right).$$

3. (COMPUTATION)

How do we calculate the mle?

For simple examples the answers to the above questions are clear. In general the answers are not trivial. Here are three simple examples that illustrate the range of possibilities.

1. Let  $Z_i \sim \mathcal{N}(\theta_0, 1)$ , and  $\Theta = (-\infty, \infty)$ . In that case the log likelihood function is

$$L(\theta) = -N \cdot \ln(2\pi) - \sum_{i=1}^N (z_i - \theta)^2 / 2.$$

The mle is  $\hat{\theta}_{mle} = \bar{z}$ , which is obviously consistent for  $\theta_0 = E(Z_i)$  and the exact distribution of  $\sqrt{N}(\bar{z} - \theta_0)$  is  $\mathcal{N}(0, 1)$ . There is no problem calculating the mle or establishing its properties.

2.  $Z_i = (Y_i, X_i)$ , with  $X_i$  having an arbitrary distribution with finite mean and variance, and  $Y_i | X_i = x \sim \mathcal{N}(x'\mu_0, \sigma_0^2)$ . Again it is easy to find the mle's,  $\hat{\mu}_{mle} = (\sum x_i x_i')^{-1} (\sum x_i y_i)$ , and  $\hat{\sigma}_{mle}^2 = \sum (y_i - x_i' \hat{\mu}_{mle})^2 / N$ , or to establish their properties.
3. Again  $Z_i = (Y_i, X_i)$ , with  $X_i$  having an arbitrary distribution with finite mean and variance, and  $Pr(Y_i = 1 | X_i = x) = \exp(x'\theta_0) / (1 + \exp(x'\theta_0))$ . There is for this logit model no analytic solution for the mle, and establishing its properties is indirect.

The duration example we will work with is also nonlinear. Again let  $Z_i = (Y_i, X_i)$ , and assume the conditional density of  $Y_i$  given  $X_i = x$  is exponential, with arrival or hazard rate  $\exp(x'\beta_0)$  (implying that the conditional mean of  $Y_i$  is  $\exp(-x'\beta_0)$ ):

$$f(y|x, \beta_0) = e^{x'\beta_0} \exp(-ye^{x'\beta_0}),$$

for positive  $y$  and zero elsewhere. This is an extension of the exponential distribution allowing the arrival rate to depend on covariates. Recall that an exponential distribution with arrival rate  $\lambda$  has a mean of  $1/\lambda$  and a variance of  $1/\lambda^2$ . In economics this distribution, and especially its extension with the arrival rate depending on covariates to allow for individual differences, has been widely used to model employment and unemployment spells since the work by Lancaster (1979).

The (conditional on  $x$ ) log likelihood function is

$$L(\beta) = \sum_{i=1}^N \ln f(y_i | x_i, \beta) = \sum_{i=1}^N x_i' \beta - y_i \cdot \exp(x_i' \beta),$$

with first derivative

$$\frac{\partial L}{\partial \beta}(\beta) = \sum_{i=1}^N x_i \cdot (1 - y \cdot \exp(x'_i \beta)),$$

and second derivative

$$\frac{\partial^2 L}{\partial \beta \partial \beta'}(\beta) = - \sum_{i=1}^N x_i x'_i \cdot y \cdot \exp(x'_i \beta),$$

There is, as in the logit case, no analytic solution for the mle. We can therefore not establish the properties of the mle directly either.

How do we compute the mle? A number of numerical methods exist for this type of problem. For ease of comparison with later optimization problems and the material in the reader we reformulate this as minimizing minus the log likelihood function - this obviously does not affect the substance of the problem. So, the objective function is  $Q(\beta) = -L(\beta)$ , and we are looking for the minimand of this function.

One leading method is Newton–Raphson. The idea is to approximate the objective function  $Q(\beta)$  around some starting value  $\beta_0$  by a quadratic function and find the exact minimum for that quadratic approximation. Redo the quadratic approximation around the minimum of the initial quadratic approximation and find the new minimum. Do this repeatedly and the solution will converge to the minimum of the objective function. Formally, given a starting value  $\beta_0$ , define iteratively

$$\beta_{k+1} = \beta_k - \left[ \frac{\partial^2 Q}{\partial \beta \partial \beta'}(\beta_k) \right]^{-1} \frac{\partial Q}{\partial \beta}(\beta_k).$$

In this case the matrix of second derivatives is

$$\frac{\partial^2 Q}{\partial \beta \partial \beta'}(\beta) = - \frac{\partial^2 L}{\partial \beta \partial \beta'}(\beta) = \sum_{i=1}^N y_i x_i x'_i \cdot \exp(x'_i \beta),$$

which is positive definite if  $\sum x_i x'_i$  is positive definite. Hence the objective function is globally convex and if there is a solution to the first order conditions, it is the unique mle. In this application the Newton–Raphson algorithm works very well.

In practice we often modify this a bit to prevent it from going out of the range of reasonable values early on. Even though this is not a problem theoretically with concave objective functions, in practice this may run into problems with machine precision. The idea is that the order of magnitude of the optimal value is typically not too large, as a result of the definition of the variables. Therefore we normalize the step as follows: Let

$$s_k = - \left[ \frac{\partial^2 Q}{\partial \beta \partial \beta'}(\beta_k) \right]^{-1} \frac{\partial Q}{\partial \beta}(\beta_k),$$

be the original Newton-Raphson step. Then the modified *direction* is

$$d_k = \frac{s_k}{1 + \sqrt{s_k' s_k}},$$

and

$$\beta_{k+1} = \beta_k + d_k.$$

This ensures that the step is never more than one in absolute value.

Another class of algorithms does not require the calculation of the second derivatives. Most of these methods separate out the choice of direction and the choice of steplength. Let  $A_k$  be any positive definite matrix, and consider iterations of the type

$$\beta_{k+1} = \beta_k - \lambda_k \cdot A_k \cdot \frac{\partial Q}{\partial \beta}(\beta_k).$$

The choice  $\lambda_k = 1$  and  $A_k = \frac{\partial^2 Q}{\partial \beta \partial \beta'}(\beta_k)^{-1}$  corresponds to original Newton–Raphson.

Below we discuss some alternatives. These alternatives involve choices for  $A_k$  that are easier to calculate than the inverse of the second derivative, typically combined with line searches for  $\lambda_k$  (discussed below). We often combine these again with normalizing the direction. In this set up the direction is

$$\tilde{d}_k = -A_k \cdot \frac{\partial Q}{\partial \beta}(\beta_k).$$

We modify this again to

$$d_k = \frac{\tilde{d}_k}{1 + \sqrt{\tilde{d}_k' \tilde{d}_k}},$$

to avoid problems with directions that are too large.

They alternative algorithms include:

1. (IDENTITY MATRIX)

$A_k$  is the identity matrix (steepest descent),

2. (INITIAL MATRIX OF SECOND DERIVATIVES)

$$A_k = \frac{\partial^2 Q}{\partial \beta \partial \beta'}(\beta_0)^{-1},$$

based on calculating and inverting the matrix of second derivatives only once, namely at the starting value  $\beta_0$ . This will work well if the third derivatives are close to zero.

3. (BERNDT-HALL-HALL-HAUSMAN)

$$A_k = \left[ \frac{1}{N} \sum_{i=1}^N \frac{\partial \ln f}{\partial \beta}(\beta_k) \frac{\partial \ln f}{\partial \beta}(\beta_k)' \right]^{-1}$$

This choice, specific to the case where the objective function is minus the log likelihood function, exploits the information matrix equality that states that at the true value  $\beta_0$  the expected value of the outer product of the first derivatives is equal to minus the expected value of the second derivatives:

$$-E \left[ \frac{\partial^2 \ln f}{\partial \beta \partial \beta'}(Z, \beta^*) \right] = E \left[ \frac{\partial \ln f}{\partial \beta}(Z, \beta^*) \frac{\partial \ln f}{\partial \beta}(Z, \beta^*)' \right].$$

It is discussed in Berndt, Hall, Hall and Hausman (1974).

4. (DAVIDON-FLETCHER-POWELL) The final choice for  $A_k$  is due to Davidon, Fletcher and Powell (see for example, Luenberger 1972; Gill, Murray and Wright, 1981). It is similar to Newton-Raphson, but does not require the analytic calculation of the second derivatives. Instead it approximates this iteratively. It involves some additional notation. Let  $\tilde{d}_k$  be the direction:

$$\tilde{d}_k = -A_k \cdot \frac{\partial Q}{\partial \beta}(\beta_k),$$

and  $d_k = \tilde{d}_k / (1 + \sqrt{\tilde{d}_k' \tilde{d}_k})$  the modified direction. Given the optimal choice for the step-length  $\lambda_k$  (the solution to  $\operatorname{argmin}_{\lambda} Q(\beta_k + \lambda \cdot d_k)$ ), the actual step is the product of the step-length and the direction:

$$p_k = \lambda_k \cdot d_k,$$

and the new point is

$$\beta_{k+1} = \beta_k + p_k.$$

Define the change in derivative between iterations as

$$q_k = \frac{\partial Q}{\partial \beta}(\beta_{k+1}) - \frac{\partial Q}{\partial \beta}(\beta_k),$$

and update the matrix  $A_{k+1}$ :

$$A_{k+1} = A_k + \frac{p_k p_k'}{p_k' q_k} - \frac{A_k q_k q_k' A_k}{q_k' A_k q_k}.$$

At the solution  $p_{\infty} = q_{\infty} = 0$ , and  $A_{\infty}$  is equal to the inverse of the second derivative of the objective function.

Many algorithms require that the steplength is chosen optimally. In general, this amounts to solving, at each step, a problem of the type

$$\min_{\lambda} Q(\beta + \lambda \cdot d).$$

Often one has a reasonable idea what the optimal steplength is. For example, if one uses the DFP or NR methods for determining the direction, at least when you are close to the optimum, the optimal steplength should be close to one. Another point to keep in mind is that with most methods for choosing the direction the objective function has negative derivative with respect to  $\lambda$  at  $\lambda = 0$ . Hence there must be a solution with  $\lambda > 0$  that at least decreases the objective function.

One approach is a simple grid search, evaluating the objective function at a number of values for  $\lambda$ , to get some idea where its minimum is. Close to the optimum this is not a very efficient strategy. More efficient methods in that case include:

1. (GOLDEN SECTION) This assumes the function is unimodal. Start with two points that you know include the minimum,  $\lambda_l < \lambda < \lambda_h$ , say  $\lambda_l = 0$  and  $\lambda_h = 1$ , and calculate the function values at those points,  $Q_l = Q(\lambda_l)$  and  $Q_h = Q(\lambda_h)$ . Choose two points in this interval,  $\lambda_{m1} = \lambda_l + 0.392 \cdot (\lambda_h - \lambda_l)$  and  $\lambda_{m2} = \lambda_l + 0.618 \cdot (\lambda_h - \lambda_l)$  (coming from  $2/(1 + \sqrt{5}) = 0.618$  and  $1 - 2/(1 + \sqrt{5}) = 0.392$ . This is where the name golden section comes from. This choice optimal in some sense, see Luenberger, 1972). Now calculate the values of the objective function at those two points,  $Q_{m1} = Q(\lambda_{m1})$  and  $Q_{m2} = Q(\lambda_{m2})$ . There are two possibilities:

- (a)  $Q_{m2} > Q_{m1}$ : in this case you know the minimum is in the range  $[\lambda_l, \lambda_{m2}]$ . In that case choose a new point in this interval, symmetric with respect to the already included point  $\lambda_{m1}$ , and continue that way.
- (b)  $Q_{m2} < Q_{m1}$ : in this case you know the minimum is in the range  $[\lambda_{m1}, \lambda_h]$ . In that case choose a new point in this interval, symmetric with respect to the already included point  $\lambda_{m2}$ , and continue that way.

At each step we have three points,  $\lambda_l < \lambda_m < \lambda_h$ , and know that the interval includes the minimum. We choose a new point  $\lambda_n$ , symmetric with respect to the interior point (so that  $|\lambda_h - \lambda_n| = |\lambda_l - \lambda_m|$  and  $|\lambda_l - \lambda_n| = |\lambda_h - \lambda_m|$ , and use that to discard part of the interval. The way of choosing the points is optimal in some sense, without

relying on smoothness.

To start this algorithm we need two starting points that contain the optimum. Suppose we have one value,  $\lambda_l$ , and we know that  $\lambda > \lambda_l$ . In that case pick  $\lambda_h = \lambda_l + 1$ . If the objective function at  $\lambda_h$  is higher than at  $\lambda_l$ , we know the optimum is in the interval  $[\lambda_l, \lambda_h]$ . If not, choose  $\lambda'_h = \lambda_l + 3 \cdot (\lambda_h - \lambda_l)$ . If the objective function at  $\lambda'_h$  is lower than at  $\lambda_h$ , we know the optimum is in the interval  $[\lambda_l, \lambda'_h]$ . If not, we know  $\lambda > \lambda_h$ . Then set  $\lambda_l = \lambda_h$ ,  $\lambda_h = \lambda'_h$  and continue with  $\lambda'_h = \lambda_l + 3 \cdot (\lambda_h - \lambda_l)$  till you have found an interval that contains  $\lambda$ .

2. (NEWTON'S METHOD) Start with initial guess  $\lambda = \lambda_0$ . Calculate the objective function,  $Q(\lambda_0)$ , its derivative,  $Q'(\lambda_0)$  and its second derivative  $Q''(\lambda_0)$ . Fit a quadratic function to these values and calculate the minimizing value  $\lambda_1$ . The solution is

$$\lambda_{k+1} = \lambda_k - \frac{Q'(\lambda_k)}{Q''(\lambda_k)}.$$

Repeat till convergence. This is just a scalar version of the algorithm we discussed already for the vector case.

3. (QUADRATIC FIT) Start with three initial points  $\lambda = (\lambda_l, \lambda_m, \lambda_h)$ , with  $\lambda_l < \lambda_m < \lambda_h$ . If 1 is an initial guess for the optimal step length, as it would often be in Newton-Raphson and Davidon-Fletcher-Powell, one might choose  $\lambda_l = 0$ ,  $\lambda_m = 1/2$  and  $\lambda_h = 1$ . Calculate the objective function at the three points,  $Q_l = Q(\lambda_l)$ ,  $Q_m = Q(\lambda_m)$ ,  $Q_h = Q(\lambda_h)$ . Fit a quadratic function to these three points, solve for its minimum to get a new point  $\lambda_n^0$ . The quadratic function that fits these three points is

$$q(\lambda) = a\lambda^2 + b\lambda + c,$$

with

$$Q_l = q(\lambda_l) = a\lambda_l^2 + b\lambda_l + c,$$

$$Q_m = q(\lambda_m) = a\lambda_m^2 + b\lambda_m + c,$$

$$Q_h = q(\lambda_h) = a\lambda_h^2 + b\lambda_h + c,$$

leading to the solution

$$q(\lambda) = Q_l \cdot \frac{(\lambda - \lambda_m) \cdot (\lambda - \lambda_h)}{(\lambda_l - \lambda_m) \cdot (\lambda_l - \lambda_h)} + Q_m \cdot \frac{(\lambda - \lambda_l) \cdot (\lambda - \lambda_h)}{(\lambda_m - \lambda_l) \cdot (\lambda_m - \lambda_h)} \\ + Q_h \cdot \frac{(\lambda - \lambda_l)(\lambda - \lambda_m)}{(\lambda_h - \lambda_l)(\lambda_h - \lambda_m)},$$

and the minimum is

$$\lambda_n = \frac{1}{2} \frac{(\lambda_m^2 - \lambda_h^2)Q_l + (\lambda_h^2 - \lambda_l^2)Q_m + (\lambda_l^2 - \lambda_m^2)Q_h}{(\lambda_m - \lambda_h)Q_l + (\lambda_h - \lambda_l)Q_m + (\lambda_l - \lambda_m)Q_h}.$$

Now we discard one of the three original points and replace it with  $\lambda_n$  in such a way that we have three points that are closer to the minimand  $\hat{\lambda}$  than the original three points. There are four possibilities depending on the value of  $\lambda_n$ .

- (a)  $\lambda_n$  is larger than the other  $\lambda$ 's: discard the smallest of the original values,  $\lambda_l$ , and start over with  $\lambda = (\lambda_m, \lambda_h, \lambda_n)$ .
- (b)  $\lambda_n$  is smaller than the other  $\lambda$ 's: discard the largest of the original values,  $\lambda_h$ , and repeat with  $\lambda = (\lambda_n, \lambda_l, \lambda_m)$ .
- (c)  $\lambda_l < \lambda_n < \lambda_m$ : repeat with  $\lambda = (\lambda_l, \lambda_n, \lambda_m)$ .
- (d)  $\lambda_m < \lambda_n < \lambda_h$ : repeat with  $\lambda = (\lambda_m, \lambda_n, \lambda_h)$ .

Compared to the first method, this method has the advantage of not requiring derivatives of the objective function, and thus fits well the the DFP algorithm. It does also rely on unimodality, and one should start with values  $\lambda_l$  and  $\lambda_h$  such that the solution is in the interval  $[\lambda_l, \lambda_h]$ .

## REFERENCES

BERNDT, E., B. HALL, R. HALL, AND J. HAUSMAN, (1974), "Estimation and Inference in Nonlinear Structural Models", *Annals of Social Measurement*, Vol. 3, 653-665..

KARLIN, S., (1962), "Stochastic Models and Optimal Policy for Selling an Asset", *Studies in Applied Probability and Management Science*, Arrow, Karlin and Scarf (eds.), Stanford, Stanford University Press.

LUENBERGER, D. (1972), *Introduction to Linear and Nonlinear Programming*, Addison Wesley, Reading Massachusetts.

GILL, P., W. MURRAY, AND M. WRIGHT, (1981), *Practical Optimization*, Harcourt Brace and Company, London