Maximum likelihood estimation in a mixture regression model using the EM algorithm

Hideo Hirose, member
Yoshio Komori, nonmember

Department of Control & Engineering Science
Kyushu Institute of Technology
Fukuoka 820-8502 Japan

Abstract

To an extremely difficult problem of finding the maximum likelihood estimates in a specific mixture regression model, a combination of several optimization techniques is found to be useful. These algorithms are the continuation method, Newton-Raphson method, and simplex method. The simplex method finds a globally approximate solution, then a combination of the continuation method and the Newton-Raphson method finds a more accurate solution. In this paper, this combination method is applied to find the maximum likelihood estimates in a Weibull-power-law type regression model, as well as the well-known methods like the EM algorithm, is discussed in this paper.

Keywords: Newton-Raphson, simplex method, power-law, Weibull distribution, Weibull-power-law, EM algorithm

1. Introduction

Maximum likelihood estimation often needs numerical iterative algorithms because of sensible nonlinearity of the likelihood function. A number of algorithms exist for finding for the optimum points of the likelihood function. Well-known algorithms are the Newton-like methods (the Newton-Raphson, quasi-Newton including BFGS which is one of the positive definite secant update methods (Dennis and Schnabel (1996), e.g.), and modified Newton), the conjugate gradient method (Fletcher-Reeves), polytope evaluation (simplex method by Nelder and Mead (1965), by Torczon (1991), and by Lewis et al. (2000)), stochastic optimizations (Markov chain Monte Carlo and simulated annealing) and the EM algorithm family. In a specific type of Weibull-power-law mixture regression model (Hirose, 1997a,b), finding accurate maximum likelihood parameter estimates is extremely difficult. Actually, the

results in Hirose (1997a,b) which were obtained just by the simplex method are found to be inaccurate in higher dimension problems. A combination of the simplex method and the continuation method, which includes the Newton-Raphson method, is quite useful for such a problem. Although the simplex method and the Newton-Raphson method are well known, the continuation method is not so familiar to statistical researchers; a brief explanation of the method is first given in this paper. The idea of the combination of several algorithms is considered elsewhere, e.g., Brooks and Morgan (1994). Why we have chosen a combination of the simplex and continuation methods is as follows: First, the simplex method is handy and we had an experience of use in the very same problem (Hirose, 1997a,b). We can make use of the information. Second, the simplex method is powerful in finding a solution in a wider range of the parameter space. Third, we knew that the continuation method can find a more accurate solution than the simplex method does, if a path of the continuation method to the final stage is appropriately provided (see section 2); the continuation method will use the Newton-Raphson method at the final stage. Fourth, the code of the continuation method is the same as the Newton-Raphson method in principle; we do not need to develop the additional code much.

In mixture models, the EM algorithm by Dempster, Laird, and Rubin, (1977) is commonly used in maximum likelihood parameter estimation (see also Everitt and Hand (1981), McLachlan and Besford (1988), and Titterington et al. (1985)). The most advantageous aspect of this algorithm is its monotonic convergence property. Moreover, the solution to the maximization step in the algorithm often exists in closed form. Thus, the EM algorithm is strongly recommended when there exists a unique solution and the maximization step is solved in closed form because of the capability of its automatic solution finding procedure . Unfortunately, the problem which is dealt with in this paper is rather complex; each maximization step cannot be solved in a closed form. For such a problem the generalized EM algorithm (McLachlan and Krishnan, 1997) can be applied. However, these EM algorithms are found to be inappropriate to solve the problem which is dealt with in this paper, because a merit of the EM algorithm that the closed form computation in the expectation step cannot be applied. This paper deals with such a problem, and does not discuss general mixture models, although comments for the EM algorithm to the problem are given.

This paper consists of six sections. Section 2 describes a general idea of the continuation method, and a concrete method to solve the likelihood equations is introduced. Section 3 explains the mixture regression model of the Weibull-power-law. Section 4 shows the combination of the simplex and continuation methods which includes the Newton-Raphson method; the combination method is applied to the specific problem for the mixture Weibull-power-law. The difficulty of the problem, the EM algorithm applied to the problem, and merits and demerits are discussed in section 5. Section 6 gives concluding remarks.

2. Continuation method

The Newton-Raphson method often fails in finding the solutions unless carefully selected initial values are used, although it has a nice property of second order convergence. It is superior to other

methods in view of the local convergence property. On the other hand, the simplex method has the global convergence property and its formulation is easy, but the rate of convergence is extremely slow. The term of "global convergence" is not strictly used here; we mean that it has a property of wider domain of attractions (see 2.3 and 5.1). The term "global solution" is usually used when there are multiple local solutions. We will not get involved in finding a global solution too deeply here; some method for it can be seen in Hirose (1998). The continuation method has both nice properties of global convergence (compared with the Newton-Raphson method) and the quick convergence (compared with the simplex method) in our experience.

2.1 Generic formulation

The continuation method can find the solutions of q nonlinear equations whenever trivial solutions are available (Allgower and Georg (1990), Watson (1986), and Watson et al. (1987, 1997), and Hirose (1994)). Let $f : R^q \to R^q$ be a smooth function, and we want to find the solution of $f(\theta) = 0$, where $\theta = (\theta_1, \theta_2, ..., \theta_q)^T$. A function $h : R^q \times [0, 1] \to R^q$ defined such that

$$h(\theta, 0) = g(\theta), \quad h(\theta, 1) = f(\theta), \tag{1}$$

is also smooth, where $g : R^q \to R^q$ is a trivial smooth map having known zero points. For instance, h is often defined as,

$$h(\theta, t) = tf(\theta) + (1 - t)\{f(\theta) - f(\theta^{(0)})\}, \tag{2}$$

where $\theta^{(0)}$ is a solution when $t = 0$. In much simpler cases (Watson, 1986), it is defined as,

$$h(\theta, t) = tf(\theta) + (1 - t)(\theta - \theta^{(0)}). \tag{3}$$

This paper adopts the former deformation. Then a zero point of (2) at $t = 1$, which becomes a solution of $f(\theta) = 0$, can be obtained by pursuing continuously a set $\{\theta(t) : 0 \le t \le 1 \mid h(\theta(t), t) = 0\}$ which is a curve in R^q, parameterized by t.

By differentiating $h = 0$ with respect to t, a system of differential equations

$$\frac{d}{dt}\theta(t) = -\left(h_\theta(\theta(t), t)\right)^{-1} h_t(\theta(t), t) \tag{4}$$

is obtained. Applying an Eulerian method to solve (4), a successive scheme,

$$\theta^{(i+1)} = \theta^{(i)} - \delta \cdot (J^{(i)})^{-1} f(\theta^{(0)}), \quad i = 0, 1, ..., \tag{5}$$

will find a solution, where $J^{(i)}$ denotes a Jacobian matrix of $f(\theta)$ at $\theta^{(i)}$ and δ is a small number. The sequence of $\theta^{(i)}$ should be located on the solution curve of $h = 0$; this procedure is called the *naive* continuation method. Although the scheme is similar to the Newton-Raphson scheme expressed by:

$$\theta^{(i+1)} = \theta^{(i)} - (J^{(i)})^{-1} f(\theta^{(i)}), \quad i = 0, 1, ..., \tag{6}$$

the continuation method has a global convergence property in principle, unlike the Newton-Raphson method. The procedure in (5) does not use $f(\theta^{(i)})$, although a parameter δ, like a step-length control in modified Newton-Raphson method, is used; (5) is different from the Newton-like scheme in that sense.

However, the naive continuation method will need many iteration steps to obtain the accurate solution because of the Eulerlian scheme. Moreover, it may fail due to turning points, bifurcation or $\theta(t)$ being unbounded; an example of these turning points is shown in Fig.1. To circumvent these inconveniences, arclength s and curve $c(s)$, which consist of zero points $h^{-1}(0)$, are introduced, since s is monotone increasing. The continuation formulation is then denoted as $h(c(s))$. By differentiating $h = 0$ with respect to s,

$$h'(c(s)) \cdot \dot{c}(s) = 0 \tag{7}$$

is obtained, where $\dot{c}(s) = dc/ds$. To reduce one free parameter, a constraint

$$\|\dot{c}(s)\| = 1, \tag{8}$$

is imposed, where $\|\cdot\|$ denotes an Euclidian norm. With an assumption that $\mathrm{rank}(h'(c(s))) = q$, an augmented Jacobian matrix,

$$A(s) = \begin{pmatrix} h'(c(s)) \\ \dot{c}(s)^T \end{pmatrix}, \tag{9}$$

becomes nonsingular, because $h'(c(s))$ is orthogonal to $\dot{c}(s)$ (see (7)). Thus, the direction of traversing the curve $c(s)$ should be determined by a constraint,

$$\det(A(0)) \cdot \det(A(s)) > 0. \tag{10}$$

The starting direction of the curve $c(s)$ is defined such that $t(s) > 0$ in ordinary continuation methods, but it would be convenient to determine the starting direction in such a manner that $\log L(\theta^{(1)}) > \log L(\theta^{(0)})$, as the need is to search for local maxima, and not for local minima.

Using (7)-(10), increments $(dc(s))^{(i)}$ are obtained by solving a system of linear equations (7). Then, a new point $(\check{c}(s))^{(i)}$ is obtained by $((c(s))^{(i)} + (dc(s))^{(i)})$. This step is called the *predictor* step. However, this point is not necessarily on the curve $c(s)$. Some correction process is needed for finding a point $(c(\tilde{s}^{(i)}))$ so that it is located on the curve $c(s)$. To find the point $(c(\tilde{s}^{(i)}))$ for the correction, we make the vector $((c(\tilde{s}))^{(i)} - (\check{c}(s))^{(i)})$ perpendicular to the vector $(dc(s))^{(i)}$; this is known to Riks-Wempner method. The point $(c(\tilde{s}))^{(i)}$ can be obtained by solving $h((c(\tilde{s}))^{(i)}) = 0$ with the restriction just mentioned above. This step is called the *corrector* step. The procedure which consists of these two steps is called the *predictor-corrector* continuation method. An illustration of the predictor and corrector points is given in Fig.2.

As is seen, the Newton-Raphson method to solve $q + 1$ dimensional nonlinear equations is already used in each corrector step. The final solution can also be obtained by using the Newton-Raphson procedure of (6); its starting point $\theta^{(k_s)}$ is obtained by the interpolation of (11),

$$\theta^{(k_s)} = \theta^{(k)} + \left(\frac{1 - t^{(k)}}{t^{(k+1)} - t^{(k)}} \right) (\theta^{(k+1)} - \theta^{(k)}), \tag{11}$$

where $c(s)^{(k)}$ and $c(s)^{(k+1)}$ are such that $t^{(k)} < 1$ and $t^{(k+1)} > 1$.

(INSERT FIGURES 1 & 2 ABOUT HERE.)

2.2 Specific formulation for the likelihood equation

In solving likelihood equations, a system of nonlinear equations corresponding to $f = 0$ is,

$$f = \left(\frac{\partial \log L}{\partial \theta_a} \right) = 0, \quad (a = 1, ..., q). \tag{12}$$

First, to obtain the direction of the incremental vector at the ith predictor step, $(d\theta^*)^{(i)}$, we temporarily solve the system of linear equations

$$(d\theta^*)^{(i)} = - \left(\left(\frac{\partial^2 \log L}{\partial \theta_a \partial \theta_b} \right)^{(i)} \right)^{-1} \left(\frac{\partial \log L}{\partial \theta_c} \right)^{(0)}, \quad (a, b, c = 1, ..., q). \tag{13}$$

This corresponds to the computation of $-\delta \cdot (J^{(i)})^{-1} f(\theta^{(0)})$ in (5) with $\delta = 1$. By setting $(dt^*)^{(i)} = 1$, adjusted increments at step i, $(d\theta)^{(i)}$, and $(dt)^{(i)}$ are,

$$\begin{aligned} (d\theta)^{(i)} &= (d\theta^*)^{(i)}/u^{(i)}, \\ (dt)^{(i)} &= 1/u^{(i)}, \end{aligned} \tag{14}$$

where, $u^{(i)} = |\sum_{a=1}^{q} ((d\theta_a^*)^{(i)})^2 + 1|^{1/2}$; $((d\theta_1)^{(i)}, (d\theta_2)^{(i)}, \cdots, (d\theta_q)^{(i)}, (dt)^{(i)})^T$ becomes a unit vector. Then, a predictor point at step i is given by

$$\begin{aligned} \check{\theta}^{(i)} &= \theta^{(i)} + \delta \cdot (d\theta)^{(i)}, \\ \check{t}^{(i)} &= t^{(i)} + \delta \cdot (dt)^{(i)}. \end{aligned} \tag{15}$$

In the corrector step, a point $c(\tilde{s})^{(i)} (= (\theta(\tilde{s})^{(i)}, t(\tilde{s})^{(i)})^T)$ is obtained by solving the system of equations iteratively (with the Newton-Raphson method in $q+1$ dimensional space),

$$\begin{aligned} \begin{pmatrix} \theta(\tilde{s}) \\ t(\tilde{s}) \end{pmatrix}^{(i)}_{(j+1)} &= \begin{pmatrix} \theta(\tilde{s}) \\ t(\tilde{s}) \end{pmatrix}^{(i)}_{(j)} - \left((B(\tilde{s}))^{(i)}_{(j)} \right)^{-1} \\ &\cdot \begin{pmatrix} \left(\frac{\partial \log L}{\partial \theta} \right) + (t-1) \left(\frac{\partial \log L}{\partial \theta} \right)^{(0)} \\ \sum_{a=1}^{q} (\theta_a(\tilde{s}) - \check{\theta}_a(s)) d\theta_a + (t(\tilde{s}) - \check{t}(s)) dt \end{pmatrix}^{(i)}_{(j)}, \\ &(j = 0, 1, ...) \end{aligned} \tag{16}$$

until $\| c(\tilde{s})^{(i)}_{(j+1)} - c(\tilde{s})^{(i)}_{(j)} \|$ becomes small, where $B(s)$ is the augmented matrix which is obtained by using the Hessian deduced from the log-likelihood function expressed as:

$$B(s) = \begin{pmatrix} \left(\frac{\partial^2 \log L}{\partial \theta_a \partial \theta_b} \right) & \left(\frac{\partial \log L}{\partial \theta_c} \right)^{(0)} \\ d\theta & dt \end{pmatrix}, \quad (a, b, c = 1, ..., q). \tag{17}$$

Here, arclength s is implicitly used. The direction of the curve $c(s)$ is determined by using the sign of $\det(B(0)) \cdot \det(B(s))$ which is the same as that of $\det(A(0)) \cdot \det(A(s))$ (see (10)). The $(i+1)$th step point $((\theta^{(i+1)}), t^{(i+1)})$ is $((\theta(\tilde{s})^{(i)}), t(\tilde{s})^{(i)})$.

2.3 Example: Two-parameter Weibull distribution

For illustration, the continuation method is applied to a small example shown in Rockette and Klimko (1974). A two-parameter Weibull model,

$$F(x; \eta, \beta) = 1 - \exp\{-(x/\eta)^{\beta}\}, \tag{18}$$

is considered. The data are, 3.1, 4.6, 5.6, and 6.8, and the maximum likelihood estimates are $\hat{\eta} = 5.54$, and $\hat{\beta} = 4.31$. Fig.3 shows the contour plot of the log-likelihood.

For comparison, the results from the continuation method with $\delta = 0.1$, the simplex method, and the Newton-Raphson method are shown. Four initial guesses are first used for typical cases: $(\eta, \beta) = (8, 2), (8, 6), (4, 2), (4, 6)$. Fig.4 shows traces of parameters for typical successful cases. The Newton-Raphson method is successful only when a initial guess, $(\eta, \beta) = (4, 6)$, is used; the continuation method fails only when a initial guess, $(\eta, \beta) = (8, 2)$, is used, but the simplex method find the solution from any starting point. Fig.4 also shows how smooth the path of the continuation method from the starting point to the final solution is; the simplex method takes many points, but is successful; the Newton-Raphson method leaps from one point to another, but shows quick convergence.

Fig.5 shows the domain of attractions. The domain of attraction D is defined such that $D = \{\theta_{\text{init}} |$ the search is successful using $\theta_{\text{init}}\}$. The successful region of the Newton-Raphson method is smaller than that of the continuation method and of the simplex method. This suggests that the continuation method is superior to the Newton-Raphson method but inferior to the simplex method in a global convergence viewpoint. This tendency does not seem unusual in our experience. Therefore, the procedure to find the maximum likelihood estimates in this paper consists of the following three stages: the simplex method first, the continuation method next, and finally, the Newton-Raphson method. As mentioned before, the predictor-corrector continuation method automatically includes the Newton-Raphson method. The combination method actually consists of the two procedure, the simplex method and the continuation method.

(INSERT FIGURES 3, 4, & 5 ABOUT HERE.)

3. Mixture regression model

Hirose (1997a,b) finds the approximate maximum likelihood estimates in a Weibull-power-law type mixture regression model by using the simplex method. The procedure of solution finding was tedious. Moreover, it is found that some of the solutions are not correct, as will be shown later; the simplex method finds stationary points, but the points are not accurate solutions. Here, light is shed on finding accurate solutions by using the combination of several optimization techniques. Before we apply it to the mixture regression model, a brief explanation of the Weibull-power-law model seems necessary.

The electrical strength of an underground electric power cable gradually deteriorates as time goes on according to its environmental stress, and finally the cable breaks. The relationship between the electrical strength of the cable and the duration that the cable has been exposed to electrical stress can be considered to follow a power-law. This relationship is similar to the relation between the lifetime of insulation and the imposed stress, which is seen in the literature (e.g., Tanaka (1983), Nelson (1990) and Hirose (1993)). According to the individual environmental stress where the cable is exposed, there exist many degradation patterns. Thus, a mixture power-law will be used for such data. For simplicity, we suppose that there are finite number of distinct groups for this mixture regression model. Fig.6 shows actual degradation sampled data. In the figure, degradation data are split into two types of cable; one is 22 and 33 kV type, and the other is 66 and 77 kV type. The data of 22 and 33 kV type are the same as in Hirose (1997a,b), but 66 and 77 kV data are newly introduced here.

(INSERT FIGURE 6 ABOUT HERE.)

According to each circumstance i of the installed cable, we assume that the degradation model can be expressed by a law,

$$v = k_i t^{-m_i}, \quad (i = 1, \cdots, g), \tag{19}$$

where v, t, m_i, and k_i denote the breakdown strength, years in use, a degradation rate, and a constant, respectively; g is the number of groups. This is inversely related to the traditional inverse-power-law, $t = Kv^{-n}$. In addition, we assume that regression lines (19) on a bilogarithmic plot include a common point (t_0, v_0), which requires a restriction such as,

$$v_0 = k_i t_0^{-m_i}, \quad (i = 1, \cdots, g). \tag{20}$$

The physical meaning of strictly positive t_0 can be explained as follows: according to Tanaka and Greenwood (1983), there are two distinct time periods in deterioration process. The first is an incubation period during which nothing appears to happen and the second is a propagation period during which cables deteriorate. Time t_0 can, therefore, be considered to be the incubation period. The degradation patterns from the beginning of use of the cables to time t_0 are considered to be common. After time t_0, the degradation rate will increase according to the location circumstance of the installed cable, which makes the regression line steeper.

We further assume a Weibull distribution (21) for the breakdown time t.

$$W(v; t) = 1 - \exp\{-\big(v/\eta(t)\big)^{\beta}\} \tag{21}$$

Here, $\eta(t)$ is a characteristic breakdown strength, and β, which is independent of t, is a shape parameter. By combining (19), (20), and (21), a power-law degradation model in group i can be formulated as,

$$\begin{aligned} F_i(v; t) &= 1 - \exp\big[-\big\{\big(k_i^{-1} t^{m_i} v\big)^{\beta}\big\}\big] \\ &= 1 - \exp\big[-\big\{\big((t/t_0)^{m_i}(v/v_0)\big)^{\beta}\big\}\big], \quad (i = 1, \cdots, g), \end{aligned} \tag{22}$$

and a mixture model of the Weibull-power-law probability distribution is,

$$F(v;t) = \sum_{i=1}^{g} p_i F_i(v;t), \quad (\sum_{i=1}^{g} p_i = 1, \quad 0 < p_i < 1), \tag{23}$$

where p_i is a proportion of a population G_i which has a distribution function F_i. For identifiability, the restriction, $m_1 < m_2 < \cdots < m_g$, is imposed.

The breakdown test method is a step-stress test, so the sampled data are actually grouped. The log-likelihood function, $\log L(g)$, can be expressed by

$$\log L(g) = \sum_{j=1}^{N} \log\left\{ \sum_{i=1}^{g} p_i \{F_i(v_j;t_j) - F_i(v_j^*;t_j)\} \right\}, \tag{24}$$

where, $v = v_{\rm bd}/v_{\rm sv}$, $v^* = (v_{\rm bd} - v_{\rm st})/v_{\rm sv}$; $v_{\rm sv}, v_{\rm bd}, v_{\rm st}$ express service stress, breakdown stress, and step-up stress, respectively. N is the number of samples. Maximum likelihood estimates of the unknown parameters, β, v_0, t_0, m_i $(i = 1, \cdots, g)$, and p_i $(i = 1, \cdots, g-1)$, can be obtained by maximizing the function (24). Although v_0 and t_0 do not appear in a explicit form, they are included in (24) implicitly; here, v_0 in (22) should be interpreted as $v_0 \leftarrow v_0/v_{\rm sv}$ in (24). The number of unknown parameters is three when $g = 1$, and $2(g+1)$ when $g > 1$. We consider the case that the number of groups, g, is also unknown. To determine the number of groups is a tricky work, but we do not concentrate on how to determine it here. We only focus on how to find the maximum likelihood estimates.

If the breakdown data are assumed to be continuously sampled, then the log-likelihood function becomes,

$$\begin{aligned} \log L(g) &= \sum_{j=1}^{N} \log\left\{ \sum_{i=1}^{g} p_i \frac{d}{dv_j}(F_i(v_j;t_j)) \right\} \\ &= \sum_{j=1}^{N} \log\left\{ \sum_{i=1}^{g} p_i \frac{\beta}{v_j} z_{i,j} \exp(-z_{i,j}) \right\}, \end{aligned} \tag{25}$$

where,

$$z_{i,j} = \{k_i^{-1} t_j^{m_i} v_j\}^{\beta} = \{(t_j/t_0)^{m_i} (v_j/v_0)\}^{\beta}.$$

This is not the correct likelihood model for the step-stress test, but the model is useful for finding the initial value for the true model of (24).

4. Procedure of solution search

Hirose (1997b) shows that each solution for $g = 1, \cdots, 6$ can be obtained by using the simplex method, and that the case $g = 4$ is optimal from the AIC, Akaike Information Criterion, (Akaike, 1973) view-point. As mentioned before, some of these solutions, however, were not accurate. In this section, how accurate solutions are found is introduced by using the combination of the simplex and continuation methods. First, the simplex method with simpler model is used for finding initial appropriate solutions; then, the continuation method finds the more accurate solutions by using the results obtained by the

simplex method as the starting points. The Newton-Raphson method is combined with the continuation procedure for finding the more accurate solution.

4.1 Use of simpler model and simpler method first

The larger the sample size, the easier it is to find the solution. In addition, the simpler the model (i.e., the smaller the number of groups in a mixture model), the easier it is to find the solution. Thus, the solution search was begun with the simplest model and with the largest sample size. Although the 22 and 33 kV service voltage type data and the 66 and 77 kV data should be treated separately because design specifications and the manufacturing system differ from each other, the breakdowns show similar patterns in both types as shown in Fig.6. Therefore, the model $g = 1$ with (25) using all the sampled data was dealt with. The number of unknown parameters is 3 in this case. Fig.7 shows the contour plot of the likelihood function in which the Weibull shape parameter β is optimized. The optimal point in the figure seems unclear. However, the simplex method finds the solution of $\hat{\beta} = 2.54$, $\hat{m} = 0.521$, and $\hat{k}' = 8.53 \times 10^{-4}$ (power-law is $v = (k't)^{-m}$, and $k = (k')^{-m} = 39.6$), although an appropriate initial guess was selected by a messy trial-and-error method.

Using these values as initial values (e.g., appropriately selected m_1 and m_2 such that $m_1 < \hat{m} = 0.521 < m_2$, and $p_1 = p_2 = 0.5$), the solution of the mixture model of $g = 2$ is obtained as shown in Table 1. In a similar manner, each solution of the model (25) using all the 22, 33, 66, and 77 kV data is obtained with some effort to find appropriate initial values. The number of iterations is very large; it becomes even larger than 10000 when $g = 6$. All the solutions are given in Table 1.

Then, the 22 and 33 kV type data and 66 and 77 kV type data are treated separately. Here, the computation is done with 22 and 33 kV type data. Also, the data are dealt with as grouped data as in (24). Since the appropriate initial values are already arranged, the procedure of the solution search becomes easier. However, the number of iterations for solution finding is still large. The solutions for 22 and 33 kV data are shown in Table 1 in Hirose (1997b).

(INSERT FIGURE 7 & TABLE 1 ABOUT HERE.)

4.2 Use of the continuation method

The simplex method provides the approximate solutions which are used as the initial guesses for the continuation method. The continuation method then gives the solutions shown in Table 2. Comparing the results obtained by the simplex method (Hirose, 1997b) with those obtained by the continuation method, it is found that the solutions for $g = 1, \ldots, 4$ by the simplex method are the same as those by the continuation method but the solutions for $g = 5, 6$ differ from each other; the continuation method provides the more accurate solutions. As mentioned before, the Newton-Raphson method is used for obtaining the final (accurate) solutions.

The trace of the continuation method procedure when $g = 6$ is shown in Fig.8. The figure demonstrates that the continuation method works well even on an extremely unstable problem; parameters

v_0 and t_0 dramatically change their values during the continuation steps although the corresponding log-likelihoods have almost the same values. Only the continuation method can find the accurate solution.

(INSERT FIGURE 8 & TABLE 2 ABOUT HERE.)

5. Discussions

5.1 Difficulty of the problem in the Newton-Raphson method

The simplex method and the continuation method can search for optimal solutions stably for the problem in this paper. However, the solution search procedure of the Newton-Raphson method alone shows extremely unstable behaviors. The following experiences indicate how difficult the problem is.

(a) Domain of attraction

As the Newton method has the second order convergence property, the convergence rate of the method is quick if the initial value is appropriately set. The domain of attraction, however, is sometimes extremely small when the objective function has many variables. The domain of attraction D is defined such that $D = \{\theta_{\text{init}} |$ using θ_{init}, the Newton method is successful$\}$.

Fig.9 shows the log-likelihood function when $g = 2$, which does not seem to be so deformed around the maximum likelihood estimates. However, the domain of attraction is not so simple as shown in Fig.10; the bright region is the successful region. In the figure, parameter values except the values corresponding to vertical and horizontal axes are set to be the maximum likelihood estimates. Even when a point is close to the maximum likelihood estimates, it is often not in the domain of attraction. The next experiment shows this clearly.

Using a Monte Carlo simulation, the maximum radius R of the q-dimensional sphere which is included in the domain of attraction standardized by $(\theta_{\text{init}} - \theta_{\text{mle}})/\theta_{\text{mle}}$ can be obtained, where θ_{mle} and θ_{init} are maximum likelihood estimates and simulated initial starting parameters, respectively. The center of the sphere is located at the maximum likelihood estimates. Each generated random number corresponds to $|(\theta^i_{\text{init}} - \theta^i_{\text{mle}})/\theta^i_{\text{mle}}|$ uniformly in the q-dimensional sphere, where θ^i_{init} and θ^i_{mle} are i-th components of the initial guess of an unknown parameter and the maximum likelihood estimate, respectively. The algorithm which generates the uniform distribution in a q-dimensional sphere is given by Knuth (1981).

When $g = 4$, the simulated maximum radius R becomes 0.022 based on 20000 generated random numbers; the random numbers are simulated so as to hold the restriction, $r = \sqrt{\sum_{i=1}^{10} \left((\theta^i_{\text{init}} - \theta^i_{\text{mle}})/\theta^i_{\text{mle}}\right)^2} < 0.2$. That is, the Newton method successfully obtains the solution whenever the initial guess is located such that $r < 0.022$. This radius seems so small that the initial starting point of the Newton method cannot be guessed easily. This fact indicates how difficult the initial value search is.

(INSERT FIGURE 9 & 10 ABOUT HERE.)

(b) Monte Carlo search

From a simulation mentioned above, searching for an initial starting point for the Newton method seems extremely difficult. This is also confirmed by the next simulation.

Since the maximum likelihood estimates are not known a priori, initial starting points are selected uniformly from a q-dimensional rectangle such that $1 \leq \beta \leq 10$, $10 \leq v_0 \leq 60$, $1 \leq t_0 \leq 5$, $0 \leq m_i \leq 1 (i = 1, ..., 4)$, $0 \leq p_i \leq 1 (i = 1, ..., 3)$.

Even by using 1000 simulated starting points for the Newton method, none of the points are successful in searching for the optimal point. More than 10000 randomly selected points are required for finding successful initial starting points. This also demonstrates how difficult the initial value selection is for the Newton method.

(c) Failures in solution search

The Newton method is successful in searching for the solution when $g \leq 5$ if the initial starting point is the result given by the simplex method, but it fails to find the solution in the case of $g = 6$ even when the initial values obtained by the simplex method are used. This indicates how sensitive the initial value selection is to the problem in this paper.

5.2 EM algorithm experiment

The EM algorithm is often used in mixture models, and it performs well in normal mixture models as shown in Everitt (1984) and Atkin and Atkin (1996). Everitt (1984) shows a comparison of performance of various algorithms, in which the EM algorithm behaves best and Newton-Raphson, using exact first and second derivatives, follows the EM algorithm. The simplex method sometimes finds alternative solutions in his examples.

The EM algorithm can also be applied to the problem in this paper. However, it does not perform well for this problem, unlike for the examples in Everitt (1984). The domains of attraction by various methods are given in Fig.11, in which $g = 2$ and the case that V_0 and m_2 are variables in Fig.10 is taken. In the figure, it can be seen that the simplex method provides the largest successful region and that the Newton-Raphson method provides the smallest region; the successful region by the continuation method is located between them. The result using the EM algorithm disappoints us. It seems better than the result by the Newton-Raphson method to some extent, but worse than the result by the continuation method. The EM algorithm for this problem requires Newton-Raphson iterations in the M-step, which makes the computation unstable.

This suggests that the performance depends on the problem. This is true for the EM algorithm. In the Appendix, the E-step and M-step of the EM algorithm derivations are shown. For the examples in Everitt (1984), the programming of the EM algorithm is easy but the programming of the Newton-Raphson method is not so easy. On the contrary, for the problem in this paper the programming of the continuation method and Newton-Raphson method is less complex than that of the EM algorithm.

(INSERT FIGURE 11 ABOUT HERE.)

5.3 Merits and demerits of the continuation method

The continuation procedure is more stable in solution finding than that of the Newton-Raphson method, but it requires more iteration steps than the Newton-Raphson method. In our experience, the number of iterations in the Newton-Raphson method is at most 20 in many maximum likelihood solution search applications; when the number of iterations exceeds this number, it suggests that the method fails. In the continuation method, the number of iterations of the predictor step is controlled by δ which is 0.02 to 0.2 according to applications. In the corrector step, the computing time for the Newton-Raphson process is dominant, but only few iteration steps for each corrector step are needed. Thus, the continuation method combined with the Newton-Raphson method requires 100 iterations, at most, in many applications. The computing cost in higher dimension only depends on linear equation solving cost, which is proportional to order of q^3 (Kincade and Cheney, 1991). We often use $\delta = 0.1$ first. If the computation fails, then we use smaller δ.

The programming cost of the continuation method is almost the same as that in the Newton-Raphson method. They both require second derivatives of the likelihood function. However, packages such as Mathematica and Maple are helpful for obtaining these derivatives. These symbolic manipulation packages output programme codes, such as FORTRAN, but it is not recommended to use these output codes directly. They require much computing time.

6. Concluding Remarks

This paper introduces a novel approach for an extremely difficult problem of finding the maximum likelihood estimates in the Weibull-power-law mixture regression model. The method is the combination of the simplex method and the continuation method which includes the Newton-Raphson method. The method works well especially for high dimensional parameter spaces where conventional methods such as the simplex method and the Newton-Raphson method fail to find the optimal points. The EM algorithm which is often used in mixture models does not provide satisfactory results to the problem in this paper.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle", in *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, pp.267-281, 1973.

[2] E.L. Allgower and K. Georg, *Numerical Continuation Methods*, Springer-Verlag, 1990.

[3] M. Atkin and I. Atkin, "A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions", *Statistics and Computing*, vol.6 pp.127–130, 1996.

[4] S.P. Brooks and B.J.T. Morgan, "Automatic starting point selection for function optimization", *Statistics and Computing*, vol.4 (1994), pp.173–177.

[5] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society, Series. B*, vol.39, pp.1–38, 1977.

[7] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, 1996.

[8] B.S. Everitt, "Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithm", *The Statistician*, vol.33, pp.205–215, 1984.

[9] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, 1981.

[10] H. Hirose, "Estimation of The Threshold Stress in Accelerated Life Testing", *IEEE Transactions on Reliability*, vol.42, pp.650–657, 1993.

[11] H. Hirose, "Parameter estimation in the extreme-value distributions using the continuation method", *Transactions of Information Processing Society of Japan*, vol.35, pp.1674–1681, 1994.

[12] H. Hirose, "Mixture model of the power law", *IEEE Transactions on Reliability*, vol.46, pp.146–153, 1997a.

[13] H. Hirose, "Lifetime assessment by intermittent inspection under the mixture Weibull power law model with application to XLPE cables", *Lifetime Data Analysis*, vol.3, pp.179–189, 1997b.

[14] H. Hirose, "Parameter estimation for the 3-parameter gamma distribution using the continuation method", *IEEE Transactions on Reliability*, vol.47 pp.188–196, 1998.

[15] D. Kincade, and W. Cheney, *Numerical Analysis*, Brooks/Cole Publishing Company, Pacific Grove, California, 1991.

[16] D.E. Knuth, *The Art of Computer Programming, Vol.2*, Addison-Wesley, 1981.

[17] G.J. McLachlan and K.E. Basford, *Mixture Models*, Marcel Dekker, 1988.

[18] R.M. Lewis, V. Torczon, and M.W. Trosset, "Direct search methods: then and now", *Journal of Computational and Applied Mathematics*, vol.124, pp.191–207, 2000.

[19] G.J. McLachlan and G.J. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.

[20] J.A. Nelder and R. Mead, "A simplex method for function minimization", *Computer Journal*, vol.7, pp.308-313, 1965.

[21] W.B. Nelson, *Accelerated Testing*, Wiley, 1990.

[22] H. Rockette, C. Antle and L.A. Klimko, "Maximum likelihood estimation with the Weibull model", *Journal of the American Statistical Association*, vol.69, pp.246–249, 1974.

[23] T. Tanaka and A. Greenwood, *Advanced Power Cable Technology, Vol. I & II*, CRC Press, 1983.

[24] D.M. Titterington, A.F.M. Smith and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York, 1985.

[25] V. Torczon, "On convergence of the multidirectionnal search algorithm", *SIAM Journal on Opimization*, vol.1, pp.123–145, 1991.

[26] L.T. Watson, S.C. Billups, and A.P. Morgan "HOMPACK: A suite of codes for globally convergent homotpy algorithm", *ACM Transactions on Mathematical Software*, vol.13, pp.281–310, 1987.

[27] L.T. Watson, "Numerical linear algebra aspects of globally convergent homotopy methods", *SIAM Review*, vol.28, pp.529–545, 1986.

[28] L.T. Watson, M. Sosonkina, R.C. Melville, A.P. Morgan, and H.F. Walker "HOMPACK90: A suite of Fortarn 90 codes for globally convergent homotpy algorithm", *ACM Transactions on Mathematical Software*, vol.23, pp.514–549, 1997.

Appendix: EM algorithm implementation

Suppose that the complete data $y = (y_1, y_2, \cdots, y_N)$ are obtained, and π_{ji} expresses the probability that y_j belongs to group i. Define the log-likelihood function for complete data as,

$$\begin{aligned}\log L_c(\theta) =& \log \prod_{j=1}^{N}\prod_{i=1}^{g}(p_i f_i(y_j; v_0, t_0, \beta, m_j))^{\pi_{ji}} \\ =& \sum_{j=1}^{N}\sum_{i=1}^{g}\pi_{ji}\big\{\log p_i - \beta m_i \log(t_0/t_j) - \big((t_0/t_j)^{-m_i}/v_0\big)^{\beta} y_j\beta\big\} \\ &+ n(\log\beta - \beta\log v_0) + (\beta-1)\sum_{j=1}^{N}\log y_i.\end{aligned} \tag{26}$$

Since y_j is observed such that $v_j^* \le y_j \le v_j$, the expectation of the log-likelihood in k-step is,

$$\begin{aligned}E_{\theta^{(k)}}[\log L_c | v_j^* \le y_j \le v_j] =& N(\log\beta - \beta\log v_0) \\ &+ \sum_{j=1}^{N}\sum_{i=1}^{g}\big\{\log p_i - \beta m_i \log(t_0/t_j)\big\}E_{\theta^{(k)}}[\pi_{ji} | v_j^* \le y_j \le v_j] \\ &+ (\beta-1)\sum_{j=1}^{N}E_{\theta^{(k)}}[\log y_j | v_j^* \le y_j \le v_j] \\ &- \sum_{j=1}^{N}\sum_{i=1}^{g}\big((t_0/t_j)^{-m_i}/v_0\big)^{\beta}E_{\theta^{(k)}}[\pi_{ji} y_j^{\beta} | v_j^* \le y_j \le v_j].\end{aligned} \tag{27}$$

This is the E-step.

Next, to construct the M-step, we differentiate (26) with respect to θ. Then,

$$p_i = \frac{1}{N}\sum_{j=1}^{N}E_{\theta^{(k)}}[\pi_{ji} | v_j^* \le y_j \le v_j], \tag{28}$$

and

$$v_0 = \Big\{\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{g}\big(t_0/t_j\big)^{-\beta m_i}E_{\theta^{(k)}}[\pi_{ji} y_j^{\beta} | v_j^* \le y_j \le v_j]\Big\}^{1/\beta}, \tag{29}$$

are obtained in closed forms by solving $\partial E_{\theta^{(k)}}[\log L_c | v_j^* \le y_j \le v_j]/\partial p_i = 0$ and $\partial E_{\theta^{(k)}}[\log L_c \ | v_j^* \le y_j \le v_j]/\partial v_0 = 0$.

By $\partial E_{\theta^{(k)}}[\log L_c | v_j^* \le y_j \le v_j]/\partial t_0 = 0$ and $\partial E_{\theta^{(k)}}[\log L_c | v_j^* \le y_j \le v_j]/\partial m_i = 0$,

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{g}m_i E_{\theta^{(k)}}[\pi_{ji} | v_j^* \le y_j \le v_j] \\ &= \frac{\sum_{j=1}^{N}\sum_{i=1}^{g}\big(t_0/t_j\big)^{-\beta m_i}m_i E_{\theta^{(k)}}[\pi_{ji} y_j^{\beta} | v_j^* \le y_j \le v_j]}{\sum_{j=1}^{N}\sum_{i=1}^{g}\big(t_0/t_j\big)^{-\beta m_i}E_{\theta^{(k)}}[\pi_{ji} y_j^{\beta} | v_j^* \le y_j \le v_j]}\end{aligned} \tag{30}$$

and

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{g}\log(t_0/t_i)E_{\theta^{(k)}}[\pi_{ji}|v_j^*\le y_j\le v_j]\\&\quad=\frac{\sum_{j=1}^{N}\sum_{i=1}^{g}\left(t_0/t_j\right)^{-\beta m_i}\log(t_0/t_i)E_{\theta^{(k)}}[\pi_{ji}y_j^{\beta}|v_j^*\le y_j\le v_j]}{\sum_{j=1}^{N}\sum_{i=1}^{g}\left(t_0/t_j\right)^{-\beta m_i}E_{\theta^{(k)}}[\pi_{ji}y_j^{\beta}|v_j^*\le y_j\le v_j]}\end{aligned}\tag{31}$$

are obtained. With (30) and (31) and by $\partial E_{\theta^{(k)}}[\log L_c|v_j^*\le y_j\le v_j]/\partial\beta=0$,

$$\begin{aligned}&\frac{1}{\beta}+\frac{1}{N}\sum_{j=1}^{N}E_{\theta^{(k)}}[\log y_j|v_j^*\le y_j\le v_j]\\&\quad=\frac{\sum_{j=1}^{N}\sum_{i=1}^{g}\left(t_0/t_j\right)^{-\beta m_i}E_{\theta^{(k)}}[\pi_{ji}y_j^{\beta}\log y_j|v_j^*\le y_j\le v_j]}{\sum_{j=1}^{N}\sum_{i=1}^{g}\left(t_0/t_j\right)^{-\beta m_i}E_{\theta^{(k)}}[\pi_{ji}y_j^{\beta}|v_j^*\le y_j\le v_j]}\end{aligned}\tag{32}$$

is obtained. By solving (30), (31), and (32) simultaneously, $t_0^{(k+1)}$, $\beta^{(k+1)}$, and $m_i^{(k+1)}$ are obtained; the number of unknown parameters is $g+2$. This is the M-step. The stopping criterion is such that $|\log L^{(k+1)}-\log L^{(k)}|$ is small, and $|\theta_a^{(k+1)}-\theta_a^{(k)}|$ is small, where

$$L(\theta)=\prod_{j=1}^{N}\sum_{i=1}^{g}p_i\big[\exp\big\{-\big(\frac{v_i}{\eta_i(t_j)}\big)^{\beta}\big\}-\exp\big\{-\big(\frac{v_i^*}{\eta_i(t_j)}\big)^{\beta}\big\}\big].\tag{33}$$

In computing the expectations, integrations, $\int_a^b u^s e^{-u}du$, $\int_a^b(\log u)u^s e^{-u}du$, and $\int_a^b(\log u)^2u^s e^{-u}du$ appear. This makes the numerical computation complex, and the accuracy of the EM algorithm is lost to some extent. In addition, the iterative procedure for obtaining the M-step solution is unstable as long as the Newton-like method is used. Thus, the EM-algorithm is not recommended for the problem in this paper.