A COMPARATIVE STUDY ON THE METHODS OF NUMERICAL ANALYSIS FOR MAXIMUM LIKELIHOOD ESTIMATION

Douke, Hideyuki
Department of Information Systems Engineering, Kyushu Tokai University

https://doi.org/10.5109/13418

出版情報：Bulletin of informatics and cybernetics. 24 (3/4), pp.185-196, 1991-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 24, No. 3~4, 1991

A COMPARATIVE STUDY ON THE METHODS OF NUMERICAL ANALYSIS FOR MAXIMUM LIKELIHOOD ESTIMATION

By

Hideyuki DOUKE*

Abstract

This paper comparatively studies adaptability of several methods of numerical analysis to obtain the maximum likelihood estimates of parameters in normal mixture distributions, on the basis of MLE-SYS system on a personal computer. The comparative study is done for seventeen unconstrained and five constrained methods in MLE-SYS system.

1. Introduction

Estimating parameters of a mixture normal distribution by moment method was first discussed by Pearson [1]. Rao [2] also proposed maximum likelihood estimation by scoring method. Several other methods of maximum likelihood estimation (MLE) have studied by Hasselblad [3],[4], Wolfe [5], Day [6], Orcard and Woodbury [7]. Several kinds of iterative algorithms in numerical analysis have been examined, but unsolved difficulties and obscurities still exist for the algorithms, e.g. non-convergency, local optimality, inadequate initial values, unsuitable timing for the convergence, frequent occurrence of improper solutions. Recently, Everitt [9] studied comparatively on six different methods of numerical analysis to obtain the maximum likelihood estimates for normal mixture distributions.

The purpose of this paper is to study comparatively the adaptability of seventeen unconstrained methods and five constrained methods in numerical analysis by the use of MLE-SYS system. The MLE-SYS system [8] has been developed originally on a personal computer by a research group of RIFIS at Kyushu University, and enables us to dissolve interactively the above difficulties by the useful functions, e.g. choice of a method of numerical analysis, restrictions of solutions, convergence conditions, intermediate changes to another analytical method and to different initial values, graphic presentation of contour map, and so on.

* Department of Information Systems Engineering, Kyushu Tokai University, Kumamoto 862, Japan

2. Maximum Likelihood Estimation of Normal Mixture Distributions

A mixture of m normal univariate distributions is as follows,

$$f(x)=\sum_{k=1}^{m} p_k f_k(x) \tag{2.1}$$

where $\sum_{k=1}^{m} p_k=1$, $f_k(x)=\exp(-(x-\mu_k)^2/2\sigma_k^2)/(2\pi)^{-1/2}\sigma_k$, $k = 1, 2, \ldots, m$. Letting n independent observations be $x_1, x_2, \ldots, x_n$, the log-likelihood function is given as follows,

$$\log L=\sum_{i=1}^{n} \log G_i \tag{2.2}$$

where $G_i = \sum_{k=1}^{m} p_k\sigma_k^{-1} E_{ki}$, $E_{ki} = \exp(-A_{ki}^2/2)$ and $A_{ki} = (x_i - \mu_k)/\sigma_k$, $k = 1, 2, \ldots, m$. To obtain the maximum likelihood estimates, the partial derivatives of log-likelihood function with respect to p, μ_k, and σ_k^2 are shown as,

$$\partial \log L/\partial p_k=\sum_{i=1}^{n} E_{ki}/G_i\sigma_k,$$

$$\partial \log L/\partial \mu_k=\sum_{i=1}^{n} p_k E_{ki}A_{ki}/G_i\sigma_k^2,$$

$$\partial \log L/\partial \sigma_k^2=\sum_{i=1}^{n} p_k E_{ki}(A_{ki}^2 - 1)/2G_i\sigma_k^3.$$

Equating to zero, the equations can not be solved explicitly, then the maximum likelihood estimates must be obtained by methods of numerical analysis.

3. The Features of MLE-SYS and Involved Methods of Numerical Analysis

MLE-SYS system has been developed on a personal computer to obtain interactively the maximum likelihood estimates by the iterative methods of numerical analysis. The system provides the following summarized features.

(a) Easy entering; For the input of underlying likelihood function or the logarithm in FORTRAN language, substitution of parts of the function are available at the same time, and furthermore if user wants, numerical differentiation of derivatives is available. (b) Variety of the methods of numerical analysis; Since many methods of numerical analysis have been studied with the respective advantages under various circumstances, the system provides twenty-two representative methods. Therefore, one can apply any method registered in the system. (c) Controllable changes of methods and initial values in numerical analysis; After applying a method of numerical analysis, if unsatisfied, the method is complementarily changeable to the other one. Initial values are also able to be devised concretely by using contour maps displaying the values of likelihood function. Thus, it will be possible to expect that MLE-SYS has attainability to the maximum value of likelihood function through the iterative process of calculation. (d) Easy operation; After user's input of likelihood or log-likelihood function, the compilation

and linkage are automatically done by the computer, without user's experience. User's requests for numerical analysis will be satisfied by simple indications in the menus and messages. (e) Easy registration of new analytical method; Regarding registration of new method of numerical analysis, only the subroutine is needed. The command is to be inserted in the menu of analytical methods in the system. (f) Machine independence and System maintenance; MLE-SYS is fully written in FORTRAN 77 language, and is available on any personal computer with Japanese MS-DOS and English MS-DOS, e.g. NEC PC-9801, TOSHIBA J-3100, IBM PC/AT.

3.1. Unconstrained methods

Representative and ordinary methods to be applied to MLE, seventeen unconstrained and five constrained methods, are provided in MLE-SYS system.

Let x and $f(x)$ be n-dimensional parameter vector and the object function. The descent methods for minimizing $f(x)$ [10] generate a sequence $x^0, x^1, \ldots$ of n dimensional points sucn as $f(x^{k+1}) < f(x^k)$, where x^{k+1} is $x^k + \alpha^k d^k$, $k = 0, 1, \ldots$.

i. Random search
   A point is searched as the minimum value of the object function is located among points indicated by random numbers.
ii. Steepest descent method
   A large number of steps are usually needed and this method is inefficient in many cases, although many text-books introduce.
iii. Newton-Raphson method
   Since many nonlinear functions can be approximated by quadratic forms, the method is frequently applied for obtaining the optimal point. If the Hessian matrix is positive definite for all points, the algorithm will converges rapidly.
iv. Levenberg-Marquardt method
   For non-positive definite Hessian matrix, the method converts the Hessian matrix into a positive definite by adding a positive definite matrix.
v. Variable metric method
   The inverse Hessian matrix is approximated by the following methods.
   (a) Fletcher-Powell method (b) Broyden-Fletcher-Shanno method
   (c) Pearson method (d) McCormick method
vi. Powell method
   The conjugate directions are generated without any gradient vector and Hessian matrix.
vii. Conjugate gradient methods
   Gradient vectors $g(x^k)$ at x^k are applied for generating conjugate directions, and the direction vector is given by $d^k = -g(x^k) + \beta^k d^{k-1}$, $\beta^k \in R^1$.
   The following methods are introduced in the system.
   (a) Fletcher-Reeves method (b) Hestenes-Stiefel-Daniel method
   (c) Sorenson-Wolfe method (d) Polak-Ribiere-Polyak method
viii. Direct search methods
   Without any derivatives, the following methods are introduced.

(a) Cyclic Coordinate method (b) Hooke-Jeeves method
(c) Rosenbrock method

3.2. Constrained method

i. Penalty method
This solution for constrained optimization is to minimize f(x), subject to $g_i(x) \leq 0$, $i = 1, 2, \ldots, m$, and $h_j(x)=0$, $j=1, 2, \ldots, r$. This converts the problem into an unconstrained optimization problem by an augmented object function with a penalty term on the object function, $F(x, r^k)=f(x)+r^k P(x)$, where $P(x)$ is a penalty function, and $r^{k+1} > r^k$, for $k = 1, 2, \ldots$.

ii. Barrier method
The method proceeds toward the constraint boundary from the inside of feasible region, and then the new object function is $F(x, r^k)=f(x)+B(x)/r^k$, $B(x)=\sum_{i=1}^{m} 1/(-g_i(x))^{\alpha}$, $(\alpha > 0)$.

iii. Multiplier method
The augmented object function is applied with the connection between the penalty functions and the Lagrange functions.

$$F(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{r} \mu_i h_i(x) + s_1 \sum_{i=1}^{m} g_i^2(x) + s_2 \sum_{i=1}^{r} h_i^2(x),$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$, $\mu = (\mu_1, \mu_2, \ldots, \mu_r)$ and $s_1, s_2 > 0$ are Lagrange multipliers.

iv. The gradient projection method
The algorithm is basing on (a) If x_i is in the interior of the feasible region, take the step which minimize $f(x)$ by applying the unconstrained method, (b) If x_i is on the boundary, treat some of the active constraints as equality constraints, and continue to minimize $f(x)$ along these constraints until Kuhn-Tucker condition is satisfied.

v. Sequential quadratic programming
The method is concerned with finding a Kuhn-Tucher point and is to solve a sequence of the quadratic programming to minimize $g(x_k)d+d^TB^kd/2$, subject to $u(x^k)+\nabla u(x^k)d \geq 0$, and $h(x^k)+\nabla h(x^k)d = 0$, where $\nabla u(x^k)$ and $\nabla h(x^k)$ are the gradient vectors of $u(x^k)$ and $h(x^k)$ respectively, B^k is the Hessian matrix of the Lagrangian, depending on the scheme of variable metric method. The procedure is (a) The d^k and Largange multipliers are obtained with solving a quadratic programming method by a symmetric and positive definite matrix B^k, (b) Set $x^{k+1}=x^k+\alpha^k d^k$ for an α^k, satisfying $min\ f(x^k+\alpha d^k)$, (c) B^{k+1} is updated by x^{k+1} and go to (a), until x^{k+1} is satisfied by a convergence criterion.

4. Comparative Studies

The numerical studies are demonstrated by applying the above mentioned methods in order to obtain the maximum likelihood estimates of parameters in a mixture of two

normal univariate densities as follows,

$$f(x) = pf_1(x)+(1-p)f_2(x) \tag{4.1}$$

where $f_k(x) = \exp(-(x-\mu_k)^2/2\sigma_k^2)/(2\pi)^{-1/2}\sigma_k^{-1}$, $k = 1, 2.$

After generating the uniform random numbers in (0,1), sets of observations were obtained by applying (4.1), when the following three sets of parameters are given,

$$\theta_a = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_1^2) = (0.4, 0.0, 3.0, 0.5, 1.0),$$
$$\theta_b = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_1^2) = (0.4, 0.0, 3.0, 1.0, 2.0),$$
$$\theta_c = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_1^2) = (0.2, 0.0, 3.0, 1.0, 2.0).$$

In the iterative process, the calculated parameters by the above described unconstrained methods are always checked for satisfying following convergence criterion,

$$|\theta_i^{k+1} - \theta_i^k| < \varepsilon, \quad \textit{and} \quad \varepsilon = 10^{-4}, \quad k = 0, 1, \ldots,$$

where θ_i^k is an estimate of i-th parameter in five parameters at k-th iteration. Thus, if any one in five parameters at each iteration is not satisfied on the criterion, the iteration is continued. On the other hand, the iterations by the constrained methods applies the following convergence criterion, because the object function have a big penalty on the constraint boundary,

$$|\log f(\theta^{k+1}) - \log f(\theta^k)| < \varepsilon, \quad \textit{and} \quad \varepsilon = 10^{-2}, \quad k = 0, 1, \ldots,$$

where θ^k is five-dimensional parameter vector at k-th iteration and $\log f(\theta^k)$ is a value of object function at θ^k. The uses of their criteria are to find the location of a global maximum, but the iterative calculations frequently lead to a local maximum in many cases. In the iterative process, if it is clear that the search attains to a local maximum in a few iterations without satisfying the convergence criterion, the iteration is automatically stopped.

The calculated parameters in each iteration are checked for exceeding the constraint boundary of parameters. If calculated parameters have a improper solution, the iteration is automatically stopped.

4.1. Comparative studies on unconstrained methods

(1) Results for θ_a

When the θ_a is compared with θ_b and θ_c, it is clear that the configuration of two normal density functions has a separated position clearly rather than the other sets. Ten samples each containing fifty observations were generated from (4.1) and θ_a. The iterations start from the following three sets of initial values.

$$\theta_1^0 = (0.4, 0.0, 3.0, 0.5, 1.0),$$
$$\theta_2^0 = (0.4, 0.0, 3.0, 1.0, 0.5),$$
$$\theta_3^0 = (0.5, 1.0, 2.0, 1.0, 0.5).$$

For many samples, depending on the initial values θ_1^0, θ_2^0 and θ_3^0, each method leads to various maximum likelihood solutions. However, many methods are classified

in some groups which have the similar maximum likelihood estimates although the estimates among their groups have the different solutions. Table 1 shows the comparison of the converged rates of all methods, e.g. the averages and ranges of iteration for ten samples without counting the case of improper solutions. It is clear that the speed of convergence depends on each initial value or method. Figure 1 shows the mixture density which corresponds to the estimates obtained from the Steepest descent method. For a sample, all methods from θ_1^0 have the following results.

i. Steepest descent method, Fletcher-Powell method, Broyden-Fletcher-Shanno method, Pearson method, Mccormick method, Fletcher-Reeves method, Hestenes-Stiefel-Daniel method, Sorensen-Wolfe method and Polak-Libiere-Polyak method have the following estimates from all initial values,
$p = 0.38$, $\mu_1 = -0.16$, $\mu_2 = 2.56$, $\sigma_1^2 = 0.30$, $\sigma_2^2 = 1.08$, $log\ L = -88.10$
and Levenberg-Marquardt method has similar estimates to the above estimates. Thus, it is clear that they lead to higher likelihood than other methods, and it usually gives the same maximum likelihood estimates regardless the initial values. Comparing the estimates with the values of θ_a, all methods estimate in common a low values for the given $\mu_2 = 3.0$, $\sigma_1^2 = 0.5$. Steepest descent method needs very many iterations, thus it shows that the method increases very slowly with 10^{-88} order.

ii. Newton-Raphson method gives different estimates to the above estimate after only one iterations,
$p = 0.42$, $\mu_1 = -0.23$, $\mu_2 = 2.50$, $\sigma_1^2 = 0.11$, $\sigma_2^2 = 0.73$, $log\ L = -93.92$*.
Also, Newton-Raphson method with linear search after 3 iterations has estimates as follows,
$p = 0.42$, $\mu_1 = -0.10$, $\mu_2 = 2.74$, $\sigma_1^2 = 0.32$, $\sigma_2^2 = 0.83$, $log\ L = -88.41$*,
where solutions with asterisks mean that the methods attain to a local maximum without satisfying the convergence criterion and finally come to stop. For a few cases, improper solutions appear on the Newton-Raphson method. Also the Newton-Raphson method and the Newton-Raphson method with linear search usually attain to a local maximum of lower likelihood after a few iterations on θ_1^0, θ_2^0. They occasionally need very many iterations on θ_3^0, they attain to a local maximum.

iii. Cyclic-Coordinate method and Rosenbrock method have the following estimates,
$p = 0.45$, $\mu_1 = 0.00$, $\mu_2 = 3.00$, $\sigma_1^2 = 0.50$, $\sigma_2^2 = 1.00$, $log\ L = -89.75$*.
Hooke-Jeeves method give similar estimates to the above estimates. They take a few iteration but have various estimates depending on initial values.

iv. The Powell method has the following estimates after 6 iterations,
$p = 0.45$, $\mu_1 = 0.00$, $\mu_2 = 3.00$, $\sigma_1^2 = 0.50$, $\sigma_2^2 = 1{,}00$, $log\ L = -89.76$,
The method converges by the convergence criterion after a few iterations and often takes lower likelihood.

v. Random method has the following estimates after 2 iterations,
$p = 0.38$, $\mu_1 = -0.03$, $\mu_2 = 2.53$, $\sigma_1^2 = 0.29$, $\sigma_2^2 = 1.38$, $log\ L = -88.87$*.
Random search takes various solutions depending on the initial values after a

few iterations.

Table. 1 A Comparison of convergence rates for θ_a

Method	Initial values			Method	Initial values		
	$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$		$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$
Random	1.8	2.8	2.2	Powell	6.4	6.8	8.5
	(1−3)	(2−4)	(1−5)		(6−8)	(6−8)	(8−9)
Steepest descent	144.1	125.8	165.2	Fletcher Reeves	37.3	36.3	48.1
	(53−266)	(87−164)	(96−288)		(7−61)	(17−61)	(19−80)
Newton Raphson	1.2	1	184	Hestenes	40.1	40.5	50.5
	(1−3)	(1−1)	(1−437)		(21−74)	(18−89)	(43−62)
N.R. with L.S.	3.7	4.3	170.2	Sorensen	36.7	39.5	53.7
	(2−8)	(1−14)	(1−602)		(18−68)	(27−55)	(33−91)
Marquardt	31.9	24.7	40.3	Polak	88.5	94.4	103.2
	(2−80)	(4−77)	(18−88)		(38−133)	(48−151)	(28−176)
Fletcher Powell	28.3	18.6	32.1	Cyclic	2.7	3.3	4.6
	(9−68)	(13−39)	(18−48)		(1−6)	(2−4)	(4−5)
Broyden	38.7	46.8	46.1	Hooke	2.2	2.2	2.2
	(25−62)	(22−66)	(31−72)		(1−4)	(2−3)	(1−4)
Pearson	40.1	42.4	48.3	Rosenbrock	2.7	3.3	4.5
	(21−55)	(24−67)	(37−62)		(1−6)	(2−4)	(4−5)
Mccormick	34.1	35.4	53.4				
	(18−63)	(19−78)	(34−84)				

Entries are the averages for 10 samples. Figures in parentheses give the range.

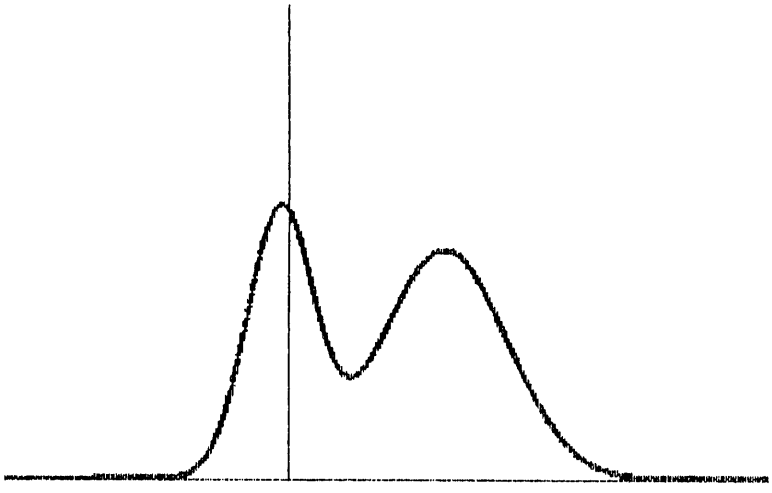

Figure 1

(2) Results for θ_b

The θ_b shows that the configuration of the two normal density functions has close position rather than the θ_a. Again ten samples each containing fifty observations were generated from (4.1) and θ_b. The iterations start from three sets of initial values as follows,

$$\theta_1^0 = (0.4, 0.0, 3.0, 1.0, 2.0),$$
$$\theta_2^0 = (0.4, 0.0, 2.0, 0.5, 1.0),$$
$$\theta_3^0 = (0.6, 1.0, 2.5, 2.0, 1.0).$$

For many samples, each method leads to more different solutions than the different solutions for θ_a. A comparison of convergence rates is given in Table 2. The whole iterations take more times than the case of θ_a. The Steepest descent method and the Newton-Raphson method and the Newton-Raphson method with linear search from θ_3^0 need much more iterations. Figure 2 shows the mixture density which corresponds to the estimates obtained from the Steepest descent method.

A example from θ_1^0 shows as follows,

i. Steepest descent method, Fletcher-Powell method and Pearson method have the following estimates regardless the initial values,
$p = 0.43$, $\mu_1 = 0.17$, $\mu_2 = 2.95$, $\sigma_1^2 = 0.69$, $\sigma_2^2 = 1.16$, $log\ L = -94.21$.

ii. Broyden-Fletcher-Shanno method, Fletcher-Reeves method, Hestenes-Stiefel-Daniel method and Sorensen-Wolfe method give the following estimates regardless the initial values,
$p = 0.43$, $\mu_1 = 0.18$, $\mu_2 = 2.96$, $\sigma_1^2 = 0.69$, $\sigma_2^2 = 1.16$, $log\ L = -94.21$.

iii. Other methods usually lead to the different maximum likelihood estimates depending on the initial values or samples, and also they have lower likelihood. For many cases, improper solutions occasionally appear in some methods, e.g. Newton-Raphson method, Fletcher-Reeves method, Hestenes-Stiefel-Daniel method and so on.

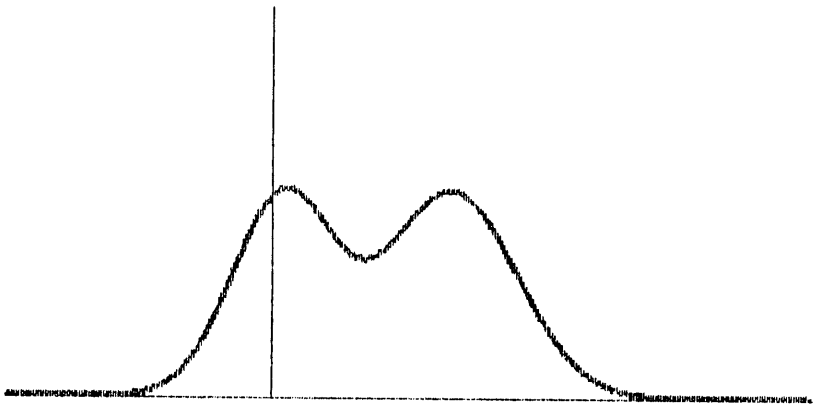

Figure 2

Table. 2 A Comparison of convergence rates for θ_b

Method	Initial values			Method	Initial values		
	$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$		$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$
Random	1.6	2.5	2.3	Powell	7.3	7.9	8.0
	(1–3)	(1–5)	(1–5)		(6–9)	(7–8)	(8–8)
Steepest descent	310.2	240.1	379.1	Fletcher Reeves	70.8	47.5	73.5
	(80–522)	(19–477)	(181–524)		(57–87)	(17–100)	(39–114)
Newton Raphson	1	5.1	399.6	Hestenes	55.2	68.5	50.8
	(1–1)	(1–7)	(1–652)		(28–82)	(22–109)	(2–72)
N.R. with L.S.	5.2	8.3	289.8	Sorensen	52.2	56.1	64.7
	(1–16)	(3–13)	(1–796)		(40–58)	(26–105)	(43–103)
Marquardt	24.4	54.2	36.6	Polak	92.7	82.8	127.4
	(3–54)	(4–100)	(3–100)		(16–196)	(21–132)	(57–246)
Fletcher Powell	41.3	53.4	42.7	Cyclic	3.4	4.7	4.2
	(13–112)	(8–126)	(9.76)		(2–6)	(3–10)	(4–5)
Broyden	36.1	41.4	46.0	Hooke	2.5	2.1	2.7
	(16–66)	(16–61)	(26–91)		(2–4)	(2–3)	(2–5)
Pearson	47.2	49.6	52.8	Rosenbrock	3.4	9.8	4.2
	(25–81)	(33–61)	(40–81)		(2–6)	(3–56)	(4–5)
Mccormick	69.5	54.2	58.0				
	(13–148)	(23–89)	(13–129)				

Entries are the averages for 10 samples. Figures in parentheses give the range.

(3) Results for θ_c

Here, starting values of the three sets are as follows,

$$\theta_1^0 = (0.2, 0.0, 3.0, 1.0, 2.0),$$
$$\theta_2^0 = (0.2, 0.0, 3.0, 1.0, 1.0),$$
$$\theta_3^0 = (0.4, 1.0, 3.0, 0.5, 1.0).$$

Again, the methods for θ_c have similar results to θ_b. In this case, most of the samples have only a few methods which lead to the similar solutions to one another regardless the initial values. Table 3 shows the comparison of convergence rates of all methods. The result is similar to θ_b except Newton-Raphson method and Newton-Raphson method with linear search from ${\theta_3}^0$. Again Figure 3 shows the mixture density corresponding to the estimates obtained from Steepest descent method. A sample from ${\theta_1}^0$ shows as follows,

i. Broyden-Fletcher-Shanno method, Hestenes-Stiefel-Daniel method, and Sorensen-Wolfe method have the following estimates from all initial values, $p = 0.69$, $\mu_1 = 1.81$, $\mu_2 = 3.54$, ${\sigma_1}^2 = 2.98$, ${\sigma_2}^2 = 0.19$, $log\ L = -92.08$. Fletcher-Reeves method shows similar estimates to the above ones.

ii. Steepest descent method gives the following estimates after 450 iterations, $p = 0.19$, $\mu_1 = -0.19$, $\mu_2 = 2.75$, ${\sigma_1}^2 = 0.25$, ${\sigma_2}^2 = 1.24$, $log\ L = -89.36$

The method has a higher likelihood, but has various estimates depending on the initial values.

iii. Fletcher-Powell method has similar result to the above estimates.
$p = 0.19$, $\mu_1 = -0.22$, $\mu_2 = 2.72$, $\sigma_1^2 = 0.22$, $\sigma_2^2 = 1.30$, $log\ L = -89.35$.

iv. Other methods lead to the different maximum likelihood estimates depending on the initial values or samples, and have lower likelihood. For the many cases, improper solutions occasionally appear in some methods.

Table. 3 A Comparison of convergence rates for θ_c

Method	Initial values			Method	Initial values		
	$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$		$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$
Random	2.4 (1–4)	1.7 (1–3)	2.6 (2–5)	Powell	6.9 (6–8)	7.4 (6–8)	8.4 (8–9)
Steepest descent	429.7 (227–812)	301.9 (33–519)	241.4 (170–404)	Fletcher Reeves	73.7 (21–174)	60.1 (20–102)	51.4 (42–65)
Newton Raphson	1 (1–1)	1.5 (1–4)	4.5 (2–8)	Hestenes	83.0 (36–106)	56.4 (2–109)	60.1 (2–101)
N.R. with L.S.	3.0 (1–14)	5.2 (1–15)	6.6 (4–10)	Sorensen	66.7 (28–124)	63.1 (38–106)	52.1 (29–79)
Marquardt	21.7 (5–88)	47.0 (5–114)	35.4 (3–100)	Polak	163.7 (57–234)	143.5 (42–296)	95.8 (36–185)
Fletcher Powell	36.2 (18–68)	50.5 (29–83)	29.2 (9–49)	Cyclic	2.7 (2–5)	3.8 (2–6)	4.3 (3–6)
Broyden	46.2 (31–70)	44.7 (35–65)	43.7 (27–65)	Hooke	2.7 (2–6)	2.1 (2–3)	3.7 (2–5)
Pearson	39.4 (14–61)	48.4 (28–75)	42.9 (25–66)	Rosenbrock	4.5 (2–23)	10.6 (2–73)	5.8 (3–19)
Mccormick	74.1 (24–159)	110.7 (24–637)	63.1 (14–133)				

Entries are the averages for 10 samples. Figures in parentheses give the range.

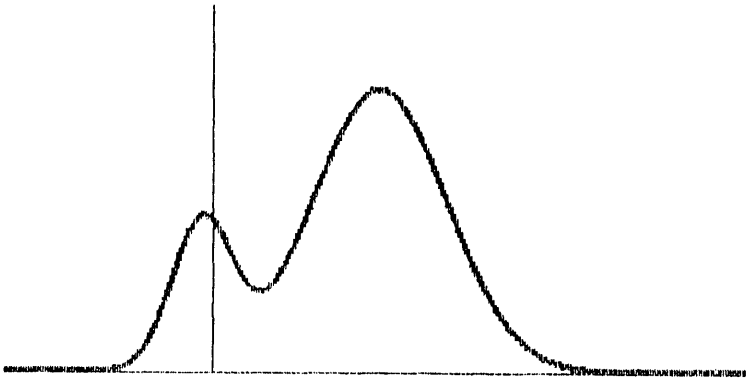

Figure 3

4.2. Comparative studies on constrained methods

Studies on the constrained methods are demonstrated under the constraint $\mu_2 - \mu_1 \leq 2.0$. The starting values are as follows,

$$\theta_1^0 = (0.4, 0.8, 2.2, 0.5, 1.0).$$

For each sample, the constrained methods of numerical analysis lead to the various maximum likelihood solutions, but Penalty method, Barrier method and Multiplier method lead to almost the same maximum likelihood estimates although they have the different solutions for each sample. Table 4 shows the comparison of convergence rates of all methods. It also shows that Gradient Projection method needs fewer iterations than other methods. A sample shows as follows,

i. Penalty method, Barrier method and Multiplier method have the following estimates,
$p = 0.36$, $\mu_1 = 0.38$, $\mu_2 = 2.40$, $\sigma_1^2 = 0.92$, $\sigma_2^2 = 1.12$, $log\ L = -89.22$.

ii. Gradient projection method gives as follows,
$p = 0.39$, $\mu_1 = 0.40$, $\mu_2 = 2.40$, $\sigma_1^2 = 0.88$, $\sigma_2^2 = 1.11$, $log\ L = -89.13$.
Thus, the Gradient Projection method for some samples have higher likelihood than other methods.

iii. Sequential quadratic programming method has the following estimates.
$p = 0.39$, $\mu_1 = 0.74$, $\mu_2 = 2.18$, $\sigma_1^2 = 0.66$, $\sigma_2^2 = 1.11$, $log\ L = -94.17$.
The Sequential quadratic programming method often obtains the estimates which do not satisfy the constraint.

Table 4. A Comparison of convergence rates for θ_a

Method	Initial values $\theta_1^{(0)}$	Method	Initial values $\theta_1^{(0)}$
Penalty	10.6 (2–29)	Gradient Projection	6.3 (2–10)
Barrier	9.2 (3–21)	Sequential quadratic programming	7.7 (3–12)
Multiplier	10 (3–21)		

Entries are the averages for 10 samples. Figures in parentheses give the range.

5. Considerations

The studies are to compare the adaptability of the seventeen unconstrained methods and five constrained methods in order to obtain the maximum likelihood estimates of the parameters in the mixture of two univariate normal densities. Thus, it is to say generally that the best methods were Variable metric method, Conjugate

gradient method and Steepest descent method, because they obtain higher likelihood than other methods regardless the initial values. Newton-Raphson method and Newton-Raphson method with linear search usually attain to a local maximum with lower likelihood after having a few iterations in many cases. Powell method and Direct search method take a few iterations, but they do not always have higher likelihood. Thus, it seems that they are not suitable in this research.

On the other hand, regarding the methods on constrained problem, it seems that the most suitable method is Gradient projection method. However, for a few samples, Penalty method has higher likelihood than Gradient projection method.

References

[1] PEARSON, K.: *Contribution to the mathematical theory of evolution*, Philosophical Transactions, A **185** (1894), 71–110.

[2] RAO, C. R.: *The utilization of multiple measurements in problems of biological classification*, J. Roy. Statist. Soc. B, **10** (1948), 159–193.

[3] HASSELBLAD, V.: *Estimation of prameters for a mixture of normal distributions*, Technometrics, **8** (1966), 431–444.

[4] HASSELBLAD, V.: *Estimation of finite mixtures of distributions from the exponential family*, J. Amer. Statist. Assoc. **64** (1969), 1459–1471.

[5] WOLFE, J. H.: *Pattern clustering by multivariate mixture analysis*, multiva. Behav. Res. **5** (1970), 329–350.

[6] DAY, N. E.: *Estimating the components of a mixture of normal distributions*, Biometrika **56** (1969), 463–474.

[7] ORCHARD, T. and WOODBURY, M. A.: *A missing information principle; theory and applications*, Proc. 6th Berkeley Symp. Vol. 1. Berkeley, University of California Press, (1972), 697–715.

[8] DOUKE, H. and ASANO, CH.: *Statistical methods and data analysis*, Studies on Numerical Solutions for Parameter Estimation of a Multivariate Exponential Distribution, Edited by N. Niki, Scientist Inc. Tokyo (1990).

[9] EVERITT, B. S.: *Maximum likelihood estimation of the parameters in a mixture of two univariate normal densities; a comparison of different algorithms*, The Statistician, **33** (1984), 205–216.

[10] KONNO, H. and YAMASITA, H.: *Nonlinear programming*, Nikkagiren, Publishing Co. Tokyo, (1987).

Received October 4, 1990
Communicated by Ch. Asano