確率密度函數의 逐次母數推定方法

84311

韓 榮 烈* ○朴 鎭 秀*
漢陽大學校* 淸州大學校**

Recursive Parameter estimation algorithm of the Probability density function

Young Yeul HAN * *Jin Soo PARK* **
Hanyang University * *Cheongju University* **

ABSTRACT We propose a new parameter estimation algorithm that converge with probability one and in mean square, if the mean is the function of parameter of the probability density function. This recursive algorithm is applicable also even the parameters we estimate are multiparameter case. And the results are shown by the computer simulation.

1. 序 論

$x_1, x_2, \cdots, x_n$ 의 確率密度函數는 알려져 있으나 母數(Parameter) θ는 알려지지 않은 確率密度函數로 부터의 標本値(sample)이고, θ가 實數의 벡터 $\theta=(\theta_1, \theta_2, \cdots, \theta_k)$일때 確率密度函數는 K個의 母數를 가지고 있다고 말할 수 있다. 여기서 $x_1, x_2, \cdots, x_n$가 推定하려는 母數를 포함한 確率密度函數 $f(x;\theta)$의 標本値일때 θ의 推定値 $\hat{\theta}$는 $x_1, x_2, \cdots, x_n$의 函數로써 다음과 같이 表示할 수 있다.

$$\hat{\theta} = g(x_1, x_2, \cdots, x_n)$$

이러한 母數를 推定하는 方法으로써 가장 많이 使用되고 있는 方法은 最確度推定(maximum likelihood estimate)方法과 모멘트推定方法이 있다.[1] 最確度推定方法은 標本値 $x_1, x_2, \cdots, x_n$가 어떤 母數의 값을 갖는 確率密度函數로 부터 標本한 것인가를 알아내는 方法으로 다음과 같이 定義되는 確度函數를 使用하고 있다.

$$L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i;\theta) \quad \cdots\cdots\cdots (1)$$

n는 標本의 數이고 θ는 K個의 母數를 表示하고 있다. 대부분의 確率密度函數는 1個 혹은 2個의 母數를 포함하고 있으며 그 以上 포함하는 경우는 드물다. 例로써 正規確率密度函數는 平均値(mean) μ와 分散(variance) σ^2의 2個의 母數를 갖는다. 따라서 最確度推定은 다음 K個의 방정식을 만족시키는 $\theta_1, \theta_2, \cdots, \theta_k$의 式이다.

$$\frac{\partial L(\theta; x_1, x_2, \cdots, x_n)}{\partial \theta_1} = 0$$
$$\frac{\partial L(\theta; x_1, x_2, \cdots, x_n)}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial L(\theta; x_1, x_2, \cdots, x_n)}{\partial \theta_k} = 0 \quad \cdots\cdots\cdots (2)$$

$L(\theta; x_1, x_2, \cdots, x_n)$과 $\log L(\theta; x_1, x_2, \cdots, x_n)$은 θ의 同一値의 最大値를 가지므로 確度函數의 代數를 取하여 풀어도 같은 式을 얻는다.

모멘트推定方法은 $f(x;\theta)$가 K個의 母數를 갖는 確率密度函數일때 $f(x;\theta)$의 r次 모멘트는 다음式으로 定義된다.

$$U_r = E[X^r] \quad \cdots\cdots\cdots (3)$$

$E[\cdot]$는 기대함수이다. $x_1, x_2, \cdots, x_n$가 確率密度函數 $f(x;\theta)$의 n個의 標本일때 j次 標本모멘트(sample moment)는

$$M_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j \quad \cdots\cdots\cdots (4)$$

이다. 그러므로 다음 K個의 방정식으로 부터 $\theta_1, \theta_2, \cdots, \theta_k$의 推定値를 구할 수 있다.

$$M_j = U_j, \quad j = 1, 2, \cdots, k \quad \cdots\cdots\cdots (5)$$

이와같이 最確度推定方法이나 모멘트推定方法에서는 θ_j, $j=1,2,\cdots,k$에 對하여 각기 推定式이 주어져 있으나, 本論文에서는 母數가 平均値의 函數일때 stochastic approximation方法을 使用하여 母數를 逐次的으로 推定하는 方法에 對하여 論하

고, 平均値가 2個 以上의 母數의 函數일때 동시에 2個의 母數를 推定하는 方法을 提示하고 實例를 들어본다.

2. 逐次母數推定알고리즘

Y_n, $n = 1, 2, \cdots$이 數値의 系列(sequence)일때

$$Y_n = F_n(\theta_1) + Z_n, \quad n = 1, 2, \cdots \qquad (6)$$

$F_n(\theta_1)$은 θ_1의 函數이고, θ_1은 推定하려는 母數이다. Z_n은 랜덤 變數(random variable)로 $F_n(\theta_1)$에 雜音(noise) 이 加해진다. Z_n의 平均値가 零이고, 有限한 分散의 값을 가지고 있는 独立的이고 同一한 分布를 가진 確率密度函數로 부터의 標本値일때 Y_n의 平均値는 다음과 같다.

$$E[Y_n] = E[F_n(\theta_1)] + E[Z_n] = F_n(\theta_1) \qquad (7)$$

式(7)은 Z_n의 平均値가 零이란 것만 알려지고 Z_n의 確率密度函數는 定義되지 않아도 成立한다. 다음 推定하려는 母數 θ_1이 하나이고 $\{\hat{\theta}_{1,n}\}$이 다음 逐次関係式으로 定義되는 推定値의 系列일때 $(n+1)$번의 θ_1의 推定値 $\theta_{1,n+1}$은 다음으로 주어진다.[2]

$$\hat{\theta}_{1,n+1} = \hat{\theta}_{1,n} + [a_n(\hat{\theta}_{1,1}, \hat{\theta}_{1,2} \cdots \hat{\theta}_{1,n})] + [Y_n - F_n(\hat{\theta}_{1,n})]$$

$$a_n(\hat{\theta}_{1,1}, \hat{\theta}_{1,2}, \cdots, \hat{\theta}_{1,n}) = \frac{\dot{F}_n(\hat{\theta}_{1,n})}{\sum_{k=1}^{n} \dot{F}_k^2(\hat{\theta}_{1,n})} \qquad (8)$$

$\dot{F}_n(\hat{\theta}_{1,n})$은 θ_1에 对한 $F_n(\theta_1)$의 導函數로 n番推定値를 의미한다. 만일 θ_1의 참값에 对한 사전지식을 가지고 있고,이때 값을 $\theta_{1,0}$라 하면 a_n으로 다음 값을 使用할 수 있다.

$$a_n(\hat{\theta}_{1,1}, \hat{\theta}_{1,2} \cdots \hat{\theta}_{1,n}) = \frac{\dot{F}_n(\hat{\theta}_{1,0})}{\sum_{k=1}^{n} F_k^2(\hat{\theta}_{1,0})} \qquad (9)$$

Albert는 $\hat{\theta}_{1,n}$이 참값θ_1에 对하여 確率 1 (probability one)과 mean square로 수렴함을 증명하였다.[2]

$$P_r[\lim_{n\to\infty}(\hat{\theta}_{1,n} - \theta_1) = 0] = 1 \qquad (10)$$

$$\lim_{n\to\infty} E[(\hat{\theta}_{1,n} - \theta_1)^2] = 0 \qquad (11)$$

이러한 方法을 stochastic approximation이라 부른다.이 式은 앞에서 論議한 바와 같이 Z_n의 確率密度函数가 定義되지 않아도 成立한다.

만일 $F_n(\theta_1)$이 常數이고, 平均値가 零인 Z_n의 確率密度函數를 알고 있다. 가정하면 Y_n은 平均値가 $F_n(\theta_1)$인 確率分布를 가진 標本値이다. 즉 미지의 母數가 θ_1일때 平均値가 θ_1의 函數이면 우리는 stochastic approximation 方法을 使用하여 θ_1을 推定할 수 있다. $F_n(\theta_1)$이 θ_1의 線形函數일때 初期推定値 $\hat{\theta}_{1,1} = 0$로 놓으면

$$F_n(\theta_1) = C\theta_1$$

$$a_n = \frac{1}{nC} \qquad (12)$$

되고, C는 常數이고 θ_1는 推定하려는 母數이다. 그러면

$$\hat{\theta}_{1,2} = \frac{Y_1}{C}$$

$$\hat{\theta}_{1,3} = \frac{(Y_1 + Y_2)}{2C}$$

$$\vdots$$

$$\hat{\theta}_{1,n+1} = \frac{\sum_{k=1}^{n} Y_k}{\sum_{k=1}^{n} C} \qquad (13)$$

된다. C가 1일때는 θ_1의 推定値는 標本平均値(sample mean)과 같다. 만일 確率密度函數의 平均値가 θ_1과 θ_2의 函數일때 처음 $\hat{\theta}_{1,1}$을 計算하고 다음 $\hat{\theta}_{2,1}$을 計算할때 $\hat{\theta}_{1,1}$의 값을 使用하여 推定하면 同時에 2個의 母數를 probability one 과 mean square로 수렴하는 추정식을 얻는다.

K個의 母數를 θ_i, $i = 1, 2 \cdots k$, 라 하고, θ_i를 逐次的으로 推定하는 推定値를 $\hat{\theta}_{i,j}$, $j = 1, 2, \cdots n$이라 할때 K個의 母數를 추정하는 알고리즘을 열거하면 다음과 같다.

(1) 推定하려는 母數의 事前근사치를 알고 있으면 初期値를 근사치로 놓고 그렇지 않으면 $\hat{\theta}_{i,1} = 0$로 놓는다.

(2) 다음 $\hat{\theta}_{2,1}$을 다음식을 利用하여 구한다.

$$\hat{\theta}_{1,2} = \hat{\theta}_{1,1} + [a_1(\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \cdots, \hat{\theta}_{k,1})] + [Y_1 - F_1(\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \cdots, \hat{\theta}_{k,1})]$$

$$a_1(\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \cdots, \hat{\theta}_{k,1}) = \frac{\dot{F}_1(\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \cdots, \hat{\theta}_{k,1})}{\dot{F}_1^2(\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \cdots, \hat{\theta}_{k,1})}$$

마찬가지로 $\hat{\theta}_{2,1}, \hat{\theta}_{3,1}, \cdots, \hat{\theta}_{k,1}$을 이미 구한 $\hat{\theta}_{i,1}$의 값을 利用하여 구한다.

(3) $\hat{\theta}_{i,n}$의 값을 구한다.

$$\hat{\theta}_{i,n} = \hat{\theta}_{i,n+1} + [a_n(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}, \cdots, \hat{\theta}_{kn})] + [Y_n - F_n(\hat{\theta}_{i,n})]$$

$$a_n(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}, \cdots, \hat{\theta}_{k,n}) = \frac{\dot{F}_n(\hat{\theta}_{i,n})}{\sum_{k=1}^{n} \dot{F}_n(\hat{\theta}_{i,n})}$$

(4) $\hat{\theta}_i$, $i = 1, 2, \cdots, k$가 수렴할때까지 逐次的으로 計算한다.

3. Computer Simulation과 결과

다음은 例를 들어 本論文에서 제시한 確率密度函數의 母數를 推定하는 方法을 利用하여 보여주고 있다. 正規分布와 그의 平均値는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2\ 2\sigma^2}, \quad -\infty < x < \infty$$

$$E[X] = \mu \qquad (14)$$

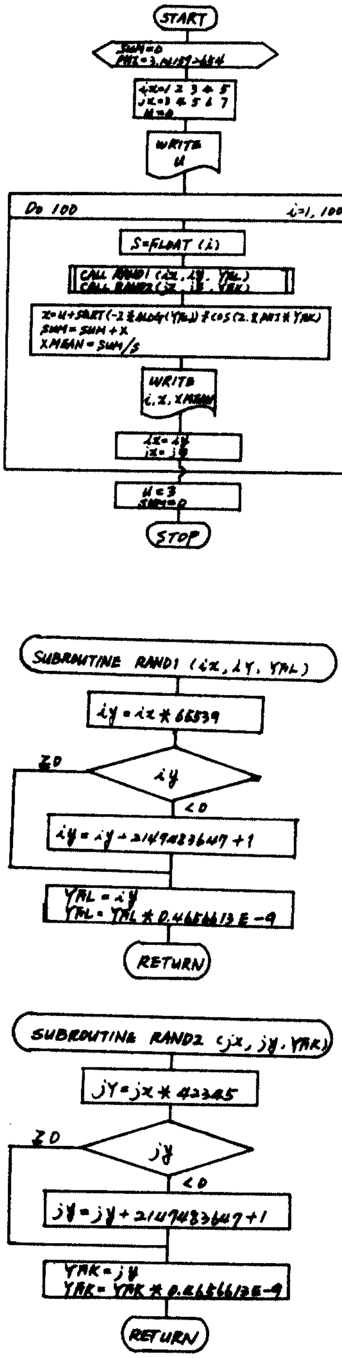
正規랜덤數発生器(Normal random number generator)로는

$$X_1 = \mu + (-2\sigma^2 \log U_j)^{\frac{1}{2}} \cos 2\pi U_{j+1}$$

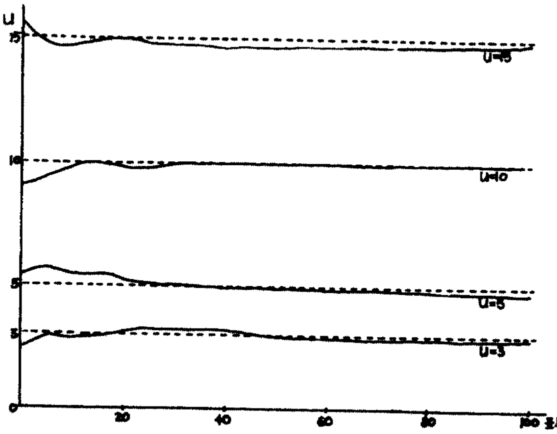
$$X_{j+1} = \mu + (-2\sigma^2 \log U_{j+1})^{\frac{1}{2}} \cos 2\pi U_{j+1}, \quad j = 1, 3, 5, \cdots \qquad (15)$$

를 使用하였다.[3] U_j, U_{j+1}은 IBM/360에 使用하는 Uniform random number인 RANDU을 使用하여 発生하였다.[3]

이때 computer simulation flow chart는 그림 1과 같으며 平均値 $\mu=3,5,10,15$에 대한 simulation 結果는 그림 2와 같다. 平均値 $E[X]=\mu$이므로 本論文에서 提示한 平均値 μ을 推定하는 알고리즘은 最尤度推定方法과 同一하다. 標本値가 많을수록 본래母數에 수렴함을 보여주고 있다.



〈그림 1〉 正規分布函數의 平均値 推定의 flow chart



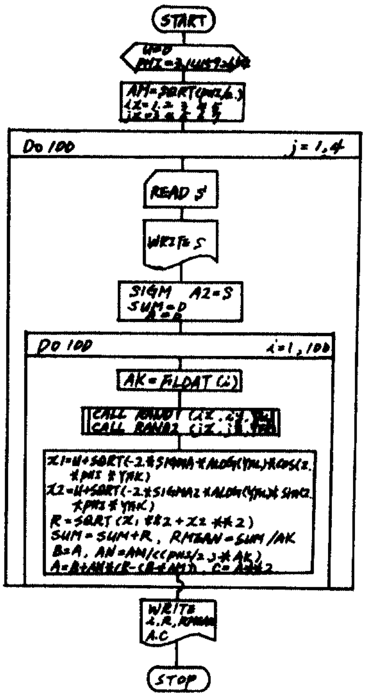
〈그림 2〉 正規分布의 平均値 推定의 수렴

다음은 X_1과 X_2가 正規分布를 가지고 있을때 $Y=(X_1+X_2)^{\frac{1}{2}}$의 確率密度函數를 Rayleigh 確率密度函數라 하며 이때 Rayleigh 確率密度函數와 그의 平均値는 다음과 같다.

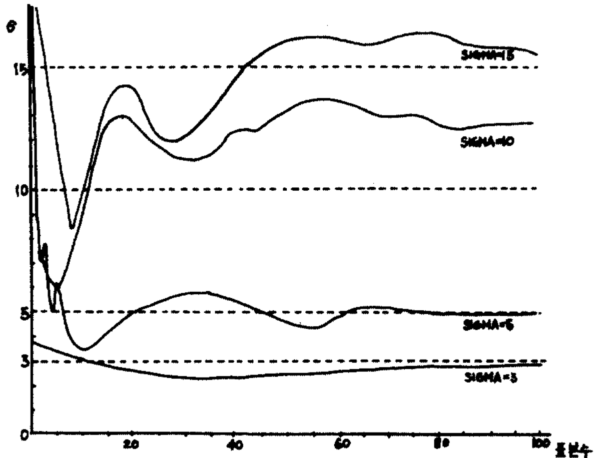
$$f(x)=\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}\quad,\quad x\geq 0$$

$$E[X]=\sqrt{\frac{\pi}{2}}\,\sigma \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (16)$$

simulation의 flow chart는 그림 3과 같으며 $\sigma^2=3,5,10,15$일때 推定値는 그림 4와 같다. Rayleigh 確率密度函數의 母數 σ^2을 推定하는 方法은 本推定方法의 有用性을 제시하고 있다.



〈그림 3〉 Rayleiph 分布函數의 σ 推定의 flow chart

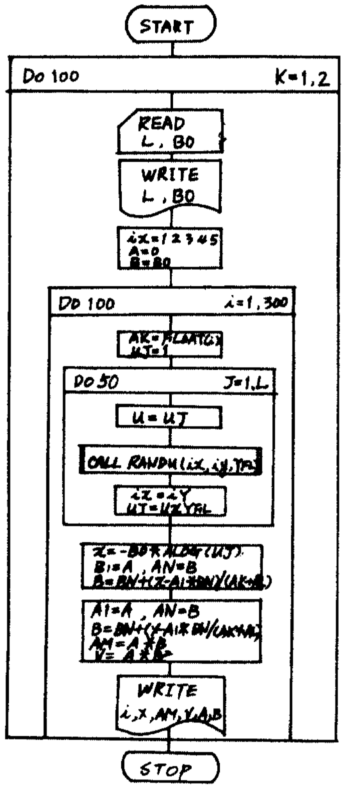


〈그림 4〉 Rayleigh 分布의 σ 推定의 수렴

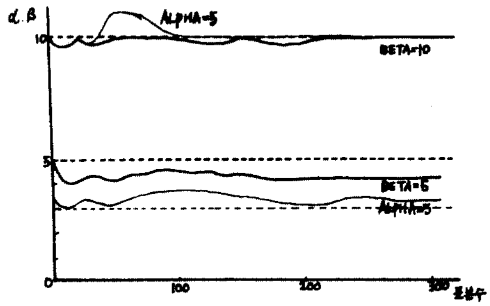
마지막으로 2個의 母數를 가진 Gamma 確率密度函數의 2個 母數를 同時에 推定하는 flow chart 와 그의 結果는 그림 5 와 그림 6 과 같다. Gamma 確率密度函數와 그의 平均値는 다음으로 주어진다.

$$f(x) = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^{\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 \le x,\ \alpha > 0,\ \beta > 0$$

$$E[X] = \alpha\beta$$



〈그림 5〉 Gamma 分布函數의 α 와 β 推定 flow chart



〈그림 6〉 Gamma 分布의 α 와 β 의 수렴

4. 結 論

推定하려는 母數가 平均値의 函數일때 母數를 推定하는 새로운 逐次알고리즘을 제시하였다. 특히 推定하려는 母數가 2個以上일때 逐次的으로 2個以上의 母數를 同時에 推定하는 方法을 제시하고 computer simulatin 에 依하여 結果를 확인하였다. 本 逐次알고리즘은 stochastic approximatin 方法을 利用하고 있으며 probibility one 과 mean square 로 수렴한다.

參 考 文 獻

(1) Mood A.M, F.A.Graybill and D.C.Boes, "Introduction to the theory of statistics," New York, McGraw-Hill, 1974.

(2) Albert, A.E and L.A.Gardner Jr., "Stochastic approximation and nonlinear regression," Cambridge, Mass. M.I.T Press, 1967.

(3) G.S.Fishman, "Concepts and methods in discrete event digital simulation," John Wiley and Sons, 1973.

(4) Kiefer, J and J.Wolfowitz, "Stochastic estimation of the maximum of a regression function," Ann, Math.Stat.pp.462-466, 1952.

(5) Eisenstein B.A, "A self-learning estimator for tracking," IEEE Trans. on Sys.Man and Cybernetics, pp.281-284, April, 1972.