응용통계연구 제 7권 1호
1994년 3월 pp. 121 - 129

문자 연산자에 의한 최대흡사추정에 관한 연구[1)]

최지훈[2)]

요약

기호연산이 가능한 소프트웨어인 MATHEMATICA를 이용함으로써 최대흡사 추정량의 구체적 형태를 구할 수 있음을 보였다. 다변량 정규분포로부터의 확률 표본을 이용한 최대 흡사 추정량을 구하는 과정을 예시하였다.

1. 머리말

최대 흡사 추정량(maximum likelihood estimator)를 구하는데 흔히 사용하는 방법은, 널리 알려져 있는 바와 같이, 다음과 같이 도입되고있다. 여기서 흡사라고 한 것은 likelihood의 뜻을 의미한다. 우리 나라 용어로는 우도 (**尤度**)라고 하고 있으나, 필자는 이 **尤**자에 대해서 너무나 뜻 없는 일본식 한자이므로, 우도 대신 가장 뜻을 잘 나타내는 흡사(**恰似**)로 쓰기로 한다. 참고로 중국에서는 사연(**似然**)이라고 하고 있다.

x의 밀도함수라고 하고, 이 $f(x)$는 m개의 미지모수 $\lambda_1, \cdots, \lambda_m$에 의해 결정되는데, 이들 모수를 크기 n인 확률표본 $x_1, \cdots, x_n$에 의해 구하고자 한다.$f(x)$가 모집단의 원소의 특성 x의 밀도함수라고 하고,이 $f(x)$는 m개의 미지 모수 $\lambda_1,\ldots,\lambda_m$에 의해 결정되는데 , 이들 모수를 크기 n인 확률 표본 $x_1,\ldots,x_n$에 의해 구하고자한다.

그러면 흡사함수 (likelihood function)를

$$L = f(x_1)f(x_2)\cdots f(x_n)$$

라고 하고 모수 $\lambda_1, \cdots, \lambda_m$의 최대 흡사 추정량은 m개의 연립 방정식

$$\frac{\partial L}{\partial \lambda_i} = 0 \qquad (i = 1,2,\cdots,m) \tag{1.1}$$

을 풀어서 구하는 것이라고 되어 있다. 따라서, 최대 흡사 추정 방법의 원칙은, 관찰한 표본을 취하는 확률을 최대로 하는 미지의 모수를 추정량으로 택하는 것이라고 할 수 있다.

이에 덧 붙여 흔히 볼 수 있는 것으로 L와 $\log L$은 $\lambda_1, \cdots, \lambda_m$의 같은 값에 대해서 최대값을 갖게 되므로 연립 방정식 (1.1)을

1) 본 연구는 1992년도 인하대학교 연구비 지원에 의하여 수행되었음
2) (402-751) 인천직할시 남구 용현동 253 인하대학교 통계학과.

$$\frac{\partial \log L}{\partial \lambda_i} = 0 \qquad (i = 1,2, \cdots, m) \tag{1.2}$$

로 바꾸어 보다 편리한 모양으로 만들어 계산하는 것으로 되어 있다.

단일 변수가 독립적이고 동일한(iid) 정규 분포를 갖는 경우의 최대 흡사 추정량을 구하는 것을 MATHEMATICA를 이용해서 (1.2)를 적용한 기호연산을 Cabrera(1989)가 보인바 있다.

그러나 이제 이것을 다변량의 경우로 직접 적용하면 (1.1) 이나 (1.2) 어느 것이든, 이런 식의 연립 방정식을 직접 푸는 것은 매우 복잡하고 어려운 문제로 되고 만다. 그래서 오로지 최대 흡사 방법의 원칙인 관찰한 표본을 취하는 확률을 최대로 하는 미지의 모수를 추정량으로 한다는 사실을 이용하여 대수학적으로 구하고있다.

그 예를 들어 보면 다음과 같다.

p변량 정규 분포를 하는 모집단에서 추출된 확률표본에서 평균 벡터 μ와 공분산 행렬 Σ의 최대 흡사 추정량을 구하는 경우를 보자. 모집단 $N(\mu,\Sigma)$에서 추출된 확률 표본 $x_1, x_2, \cdots, x_n$의 흡사함수 L은 $n>p$라고 할 때,

$$\begin{aligned} L &= \prod_{\alpha=1}^{n} N(x_\alpha \mid \mu,\Sigma) \\ &= \frac{1}{(2\pi)^{\frac{1}{2}pn} \mid \Sigma \mid^{\frac{1}{2}n}} \\ &\quad \exp\left\{ -\frac{1}{2} \sum_{\alpha=1}^{n} (x_\alpha - \mu)' \Sigma^{-1} (x_\alpha - \mu) \right\} \end{aligned} \tag{1.3}$$

가 된다. 즉 이 흡사함수의 벡터 $x_1, x_2, \cdots, x_n$ 은 표본 관찰 값으로서 고정된 값이고, L은 μ와 Σ의 함수로 본다. 그러면 이 흡사함수의 로그는 다음 (1.4)와 같다.

$$\begin{aligned} \log L &= -\frac{np}{2} \log (2\pi) \\ &\quad -\frac{n}{2} \log \mid \Sigma \mid \\ &\quad -\frac{1}{2} \sum_{\alpha=1}^{n} (x_\alpha - \mu) \Sigma^{-1} (x_\alpha - \mu) \end{aligned} \tag{1.4}$$

잘 알려진 바와 같이 $\log L$은 L의 증가함수 이므로 $\log L$의 최대치는 L의 최대치가 된다. 그래서 logL을 최대로 하는 μ, Σ를 구하는데는 $\log L$을 다음과 같이 변형한다.

$$\begin{aligned} \log L = &-\frac{pn}{2} \log(2\pi) - \frac{n}{2} \log \mid \Sigma \mid \\ &-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1} \left(\sum_{\alpha=1}^{n} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \right) \right] \\ &-\frac{1}{2} n (\bar{x} - \mu)' \Sigma^{-1} (\bar{x} - \mu) \end{aligned} \tag{1.5}$$

이것은 다변량 통계학에서는 자주 사용하고있는 표현이다.

먼저 $\log L$을 최대로 하는 μ의 값을 구한다.
Σ가 정치이므로 Σ-1 도 역시 정치 이고

$$n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) \geq 0$$

이므로 $\mu = \bar{x}$ 일 때에만 0이 된다. 이래서 $\log L$을 최대로 하는 μ는 $\bar{x}$라는 것을 알아 낸다.
다음은 $\log L$을 최대로 하는 Σ를 구한다. 이 때에는 다음과 같은 보조 정리를 이용한다.

Lemma (Anderson(1971)) A가 차수 p의 양 정치인 행렬이고

$$f(\Sigma) = -n \log n \mid \Sigma \mid - \mathrm{tr}\Sigma^{-1}A$$

를 최대로 하는 양 정치 행렬 Σ가 존재하면 그것은 $\Sigma = 1overnA$애서 $f(\Sigma)$를 최대로 하고, 그 때의 값은

$$f(1overnA) = pn \log n - n \log \mid A \mid - pn$$

이다.
이 Lemma를 이용하면

$$\Sigma = (1/n)\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

일 때가 $\log L$을 최대로 하는 것을 알 수 있다. 위에서 구한 $\mu=\bar{x}$ 는

$$\mu = \bar{x} = (1/n)\sum_{\alpha=1}^{n} x_\alpha$$

인 것은 쉽게 알 수 있다.

또 이와 유사한 방법이 Johnson and Wichern(1982)에서도 찾아 볼 수 있다. 이들 두 방법은 어느 것이나 e의 멱의 식이 음의 부호를 가지고 있으므로 이 음부호를 고려하지 않는 값을 최소로 하면 $L(\mu,\Sigma)$의 값을 최대로 한다는 조건을 이용하여 대수학적으로 이 조건을 만족시키는 μ와 Σ의 최대 흡사 추정량을 구하고 있다.
물론 이러한 방법이 더 세련된 방법이기는 하지만, 아주 고지식한 방법이라고 할 수 있는, 로그 흡사 함수를 벡터 μ와 행렬 Σ의 각 원소에 관해서 편미분한 식을 0으로놓은 연립방정식 (1.1) 이나 (1.2)를 직접 풀어 보자는 것이 이 논문의 목적이다.

2. MATHEMATICA의 필요한 연산

다변량 통계학에서 자료행렬과, 평균 벡터, 분산-공분산 행렬을 취급하는데는 일정한 형식이 있는 것이 아니고 약속하기 나름 이므로 여기서도 먼저 MATHEMATICA에서 사용 하게 될 자료행렬, 평균 벡터, 분산-공분산 대칭 행렬 을 다음과 같이 정의 한다.
일반적으로 p개의 변수에 관한 n개의 관찰 단위 $y_1, y_2, \cdots y_n$는 각각 p차원 벡터이고, 자료 행렬을 보통은 다음과 같은 p x n 행렬 로 나타 내고 있다.

$$\begin{pmatrix} y_{11}, & y_{12}, & \cdots & y_{1n} \\ y_{21} & y_{22}, & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{pmatrix} \tag{2.1}$$

그러나 MATHEMATICA의 연산의 편의상 예를 들어 p=2, n=3 즉 두 변수의 세 관찰 단위의 자료 행렬을

$$\begin{matrix} y[1][1] & y[2][1] \\ y[1][2] & y[2][2] \\ y[1][3] & y[2][3] \end{matrix} \tag{2.2}$$

라는 형태의 배열로 나타 내기로 한다. 즉 (2.1)의 행렬이 전치 된 모양이다.
이러한 배열을 만들려면 먼저 MATHEMATICA에서 Array 함수를 이용해서 두 변수 y[1], y[2]의 리스트 {y[1],y[2]}를 만들고, 각 변수에 세 개 씩의 관찰치를 갖는 행렬을 만든 후, 이를 전치 하면 된다. 전치하기 전의 모양은 (2.1)의 형식과 같은 배열이 된다. 이 MATHEMATICA의 명령문은 다음과 같다.

```
rv=Array[y,2]
dmat=Transpose[Table[Array[rv[[i]],3],{i,2}]]
```

자료 행렬을 (2.1)과 같은 행렬의 형식에서 (2.2)와 같이 놓는데는 흡사함수를 MATHEMATICA로 정의 하기 편하게 하기 위한 것이다. 즉 첫번 째 관찰 단위에 관한 두 변수의 관찰 치의 벡터 {y[1][1],y[2][1]}는 dmat[[1]]라고 부르면 되기 때문이다.

다음은 평균과 분산에 관해서 미리 그 모양을 알아두자.
먼저 평균이다. 첫번 째 변수 y[1]의 평균 m[1]은 다음과 같이 MATHEMATICA 언어로 나타낼 수 있다.

```
m[1]=Sum[y[1][i],{i,3}]/3
```

이것을 실행 시키면

$$m[1]=(y[1][1] + y[1][2] + y[1][3])/3 \tag{2.3}$$

와 같이 나타난다. 이 표현은 후에 가서 다변량 정규분포의 최대 흡사 추정에 의한 평균의 표

현과 일치하는 것을 확인 하게 된다.

분산과 공분산에 관해서도 미리 그 표현 식을 나타내어 두자. 먼저 두 다변량 변수 y[1]과 y[2]의 분산 s[1,1]과 s[2,2]는 각각

```
s[1,1] = ((y[1][1] - (y[1][1] + y[1][2] + y[1][3])/3)^2 +
          (y[1][2] - (y[1][1] + y[1][2] + y[1][3])/3)^2 +
          (y[1][3] - (y[1][1] + y[1][2] + y[1][3])/3)^2)/3
```

이것을 MATHEMATICA의 명령문의 하나인 Simplify를 사용하면

```
Simplify[[s[1,1]] = (2*(y[1][1]^2 - y[1][1]*y[1][2]
                      + y[1][2]^2 - y[1][1]*y[1][3]                    (2.4)
                      - y[1][2]*y[1][3] + y[1][3]^2))/9
```

가 된다. 마찬가지로 s[2,2] 는

```
s[2,2] = ((y[2][1] - (y[2][1] + y[2][2] + y[2][3])/3)^2 +
          (y[2][2] - (y[2][1] + y[2][2] + y[2][3])/3)^2 +
          (y[2][3] - (y[2][1] + y[2][2] + y[2][3])/3)^2)/3
```

Simplify 명령을 쓰면,

```
s[2,2] = (2*(y[2][1]^2- y[2][1]*y[2][2]
       + y[2][2]^2 - y[2][1]*y[2][3]                                   (2.5)
       - y[2][2]*y[2][3] + y[2][3]^2))/9
```

다음 공분산 s[1,2]는

```
s[1,2] = ((y[1][1] - (y[1][1] + y[1][2] + y[1][3])/3)*
          (y[2][1] - (y[2][1] + y[2][2] + y[2][3])/3) +
          (y[1][2] - (y[1][1] + y[1][2] + y[1][3])/3)*
          (y[2][2] - (y[2][1] + y[2][2] + y[2][3])/3) +
          (y[1][3] - (y[1][1] + y[1][2] + y[1][3])/3)*
          (y[2][3] - (y[2][1] + y[2][2] + y[2][3])/3))/3
```

이를 Simplify하면

```
s[1,2] = (2*y[1][1]*y[2][1] - y[1][2]*y[2][1]
       - y[1][3]*y[2][1] - y[1][1]*y[2][2]
       + 2*y[1][2]*y[2][2] - y[1][3]*y[2][2]                           (2.6)
       - y[1][1]*y[2][3] - y[1][2]*y[2][3]
       + 2*y[1][3]*y[2][3])/9
```

를 얻는다. 이들 (2.3)에서 (2.6) 까지의 식이 후에 계산되어 나오는 결과와 일치하는 것을 확인하면 된다.

3. 최대 흡사 추정량

3.1 입력 자료

다변량 정규분포의 분포함수의 정의에 앞서, 그 표현 식에서 사용할 변수들을 정의 할 필요가 있다. 다음의 함수를 보면 필요한 자료가 어떻게 정의 되는가를 알 수 있다.

```
DataMuCovMat[rv_,dmat_,mean_,cov_,ups_,p_,n_]:=
        (rv=Array[y,p];
        dmat=Transpose[Table[Array[rv[[i]],n],{i,p}]];
        mean=Array[m,p];
        cov=Array[s,{p,p}];
        Do[s[i,j]=s[j,i],{j,1,p},{i,j,p}];
        ups=Table[s[i,j],{i,p},{j,i,p}];)
```

여기서 특기할 것은 공분산 행렬 cov는 대칭 행렬이고, 후에 가서 구할 분산-공분산은 전 cov 행렬의 원소 전부 대신, 이 행렬의 대각선을 포함한 그 위에 있는 원소만을 추정 하면 되므로 ups라는 상부 삼각 행렬을 정의 하였다.

이제 우리는 예시적으로 p=2, n=3인 경우에 대한 최대 흡사 추정을 해보고자하니 다음과 같이 위의 함수를 부른다.

```
DataMuCovMat[rv,xx,mu,ss,tt,2,3]
```

3.2. 함수의 정의

다음은 다변량 정규분포 함수의 정의이다.

```
mdf[mrvx_,{muv_,cov_},p_]:=
        1/(Det[cov]^(1/2) Sqrt[2 Pi]^p)*
        E^(-(mrvx - muv).Inverse[cov].(mrvx - muv)/2)
```

가령 p=2인 경우의 다변량 정규분포의 식을 보면

```
mdf[rv,{mu,ss},2]
```

라고 했을 때

```
1/(2*E^(((-m[2] + y[2])*(-((s[1, 2]*(-m[1] + y[1]))/
                (-s[1, 2]^2 + s[1, 1]*s[2, 2])) +
            (s[1, 1]*(-m[2] + y[2]))/(-s[1, 2]^2 + s[1, 1]*s[2, 2])) +
        (-m[1] + y[1])*((s[2, 2]*(-m[1] + y[1]))/
            (-s[1, 2]^2 + s[1, 1]*s[2, 2]) -
            (s[1, 2]*(-m[2] + y[2]))/(-s[1, 2]^2 + s[1, 1]*s[2, 2])))/2)*Pi*
```

```
    (-s[1, 2]^2 + s[1, 1]*s[2, 2])^(1/2))
```

라고 된다.

다음은 다변량 정규 분포의 로그 흡사 함수의 정의이다. 이제는 관찰치가 삽입될 차례다. 위에서 정의한 xx에 유의 하면 이것이 자료 행렬인 것을 알 수 있다.

```
LogLikeMdf[mdf_,xx_,theta_,p_]:=
        Sum[Log[mdf[xx[[i]],theta,p]],{i,1,Length[xx]}]
```

3.3. 로그 흡사 함수의 풀이

이것을 모수 {m[1],m[2],s[1,1],s[1,2],s[2,2]}들로 편 미분해서 이들을 0과 같게 놓고 연립방정식을 풀기 위해

```
llfunc=LogLikeMdf[mdf,xx,{mu,ss},2];
  flatyeta=Flatten[{mu,tt}]
Do[diff[i]=D[llfunc,flatyeta[[i]]],{i,Length[flatyeta]}]
Solve[Thread[Equal[Array[diff,2],0]],mu]
```

하면 다음과 같은 mu={m[1],m[2]}에 대한 해를 기호 연산의 결과로 제공해 준다.

```
{{m[1] -> -(s[1, 2]*(s[1, 2]*y[1][1] + s[1, 2]*y[1][2] +
            s[1, 2]*y[1][3] - s[1, 1]*y[2][1] -
            s[1, 1]*y[2][2] - s[1, 1]*y[2][3]))/
       (3*(-s[1, 2]^2 + s[1, 1]*s[2, 2])) +
      (s[1, 1]*(s[2, 2]*y[1][1] + s[2, 2]*y[1][2] +
           s[2, 2]*y[1][3] - s[1, 2]*y[2][1] -
           s[1, 2]*y[2][2] - s[1, 2]*y[2][3]))/
       (3*(-s[1, 2]^2 + s[1, 1]*s[2, 2])),
    m[2] -> -(s[2, 2]*(s[1, 2]*y[1][1] + s[1, 2]*y[1][2] +
            s[1, 2]*y[1][3] - s[1, 1]*y[2][1] -
            s[1, 1]*y[2][2] - s[1, 1]*y[2][3]))/
       (3*(-s[1, 2]^2 + s[1, 1]*s[2, 2])) +
      (s[1, 2]*(s[2, 2]*y[1][1] + s[2, 2]*y[1][2] +
           s[2, 2]*y[1][3] - s[1, 2]*y[2][1] -
           s[1, 2]*y[2][2] - s[1, 2]*y[2][3]))/
       (3*(-s[1, 2]^2 + s[1, 1]*s[2, 2])))}
```

이것을 Simplify하면

```
{{m[1] -> (y[1][1] + y[1][2] + y[1][3])/3,
   m[2] -> (y[2][1] + y[2][2] + y[2][3])/3}}
```

라는 결과를 얻는다. 이것은 (2.3)의 식과 일치 하고 있다.

이제 이들 두 변수 y[1]와 y[2]의 각 모평균의 최대 흡사 추정량을 다음과 같이 m[1], m[2]에 대입하여 나머지 세 미지의 모수 s[1,1],s[1,2],s[2,2]를 구한다.

```
Do[subdiff[k]=diff[k]/.m[1]->Sum[y[1][i],{i,3}]/3
                        /.m[2]->Sum[y[2][i],{i,3}]/3,{k,3,5}]

Solve[Thread[Equal[Table[subdiff[k],{k,3,5}],0]],
        Table[flatyeta[[i]],{i,3,5}]]
```

이 결과 다음을 얻는다.

```
{{s[1, 1] -> (2*y[1][1]^2 - 2*y[1][1]*y[1][2] + 2*y[1][2]^2 -
       2*y[1][1]*y[1][3] - 2*y[1][2]*y[1][3] + 2*y[1][3]^2)/9,
   s[2, 2] -> (2*y[2][1]^2 - 2*y[2][1]*y[2][2] + 2*y[2][2]^2 -
       2*y[2][1]*y[2][3] - 2*y[2][2]*y[2][3] + 2*y[2][3]^2)/9,
   s[1, 2] -> (2*y[1][1]*y[2][1] - y[1][2]*y[2][1] - y[1][3]*y[2][1] -
       y[1][1]*y[2][2] + 2*y[1][2]*y[2][2] - y[1][3]*y[2][2] -
       y[1][1]*y[2][3] - y[1][2]*y[2][3] + 2*y[1][3]*y[2][3])/9}}
```

이들은 우리가 (2.4),(2.5),(2.6)에서 보인바 있는 식들이다.

4. 맺는 말

위에서 보인바와 같이, 본 논문은 기존의 결과를 새로운 기호연산의 기능을 갖는 MATHEMATICA라는 소프트 웨어를 이용해서 구한 것에 지나지 않지만, 여러 밀도 함수에 대해서도 동일한 생각의 알고리즘을 이용해서 최대 흡사 추정을 구할 수 있으리라 믿어진다. 특히 흡사 함수를 이용한 분야에 관한 연구를 하는 사람들에게 이러한 알고리즘이 이용되면 무엇인지 도움이 되지 않을까 기대한다.

참 고 문 헌

[1] Anderson,J.W. (1971), *An Introduction to Multivariate Statistical Analysis,* 2nd ed. John Wiley & Sons.

[2] Cabrera, J. (1989), Some Experiments with maximum likelihood estimation using symbolic manupulations, Computer Science and Statistics, *Proceedings of the 21st Symposium on the interface,* 405-409

[3] Johnson, R.A. and Wichern D.W. (1982), *Applied Multivariate Statistical Analysis,* 3rd ed. Prentice-Hall, Inc.

[4] Wolfram, S. (1991), *Mathematica,* 2nd ed., Addison-Wesley, Inc.

A Study on Maximum Likelihood Estimation Using Symbolic Manipulation[3)]

Chi-Hoon Choi[4)]

Summary

Using the software named MATHEMATICA which can evaluate symbolic operation, it is demonstrated to obtain the maximum likelihood estimation form. A sample from Multivariate Normal distribution is adopted to show the process of obtaining the maximum likelihood estimator.

3) This research was supported by Inha University Research Grant, 1992.

4) Dept. of Statistics, Inha University, #253 Yunghyun-Dong, Namku, Inchon, 402-751.