베이지안 통계 추론

이 호 석
뉴미디어학과 공과대학 호서대학교
hslee@office.hoseo.ac.kr

On the Bayesian Statistical Inference

Ho Suk Lee
New Media Dept. College of Engineering Hoseo University

요 약

본 논문은 베이지안 통계 추론에 대하여 논의한다. 논문은 베이지안 추론, Markov Chain과 Monte Carlo 적분, MCMC(Markov Chain Monte Carlo) 기법, Metropolis-Hastings 알고리즘, Gibbs 샘플링, Maximum Likelihood Estimation, EM 알고리즘, 상실된 데이터 보완 기법, BMA(Bayesian Model Averaging) 순서로 논의를 진행한다. 이러한 통계적 기법들은 대용량의 데이터를 처리하는 생물학, 의학, 생명 공학, 과학과 공학, 그리고 일반 데이터 조사와 처리 등에 사용되고 있으며, 최적의 추론 결과를 이끌어 내는데 중요한 방법을 제공하고 있다. 그리고 마지막으로 PC(Principal Component) 분석 기법에 대하여 논의한다. PC 분석 기법도 데이터 분석과 연구에 많이 활용된다.

Abstract

This paper discusses the Bayesian statistical inference. This paper discusses the Bayesian inference, MCMC (Markov Chain Monte Carlo) integration, MCMC method, Metropolis-Hastings algorithm, Gibbs sampling, Maximum likelihood estimation, Expectation Maximization algorithm, missing data processing, and BMA (Bayesian Model Averaging). The Bayesian statistical inference is used to process a large amount of data in the areas of biology, medicine, bioengineering, science and engineering, and general data analysis and processing, and provides the important method to draw the optimal inference result. Lastly, this paper discusses the method of principal component analysis. The PCA method is also used for data analysis and inference.

1. 서 론

베이지안 통계 추론은 서로 독립적인 무작위 변량의 생성에 기반을 두고 있다[1]. 여기서 무작위 변량의 통계적 분포는 미리 알려져 있는 경우도 있고 그렇지 않은 경우도 있다. 그러나 일반적으로 균일 분포를 사용하는 경우가 많으며 정규 분포를 사용하는 경우도 있다. 이외에도 무작위 변량에 대한 분포 f(x) 가, 이미 잘 알려져 있고 비교적 실험하기 용이한 다른 분포와 관계가 있는 경우에는, 용이한 다른 분포를 적절한 방법으로 활용하여 분포 f(x)로 무작위 변량을 시뮬레이션 하는 것과 동일한 효과를 내도록 할 수 있다. 즉, 이러한 경우에는 분포 f(x) 의 함수 형태만을 고려하여 실제 시뮬레이션을 수행할 수 있는 용이한 다른 분포 g(x) 를 다음과 같은 수식으로 생각하고 분포g*(x)* 를 활용하여 시뮬레이션을 할 수 있다.

$$f(x) \le Mg(x) \qquad (1)$$

다음은 이와 같은 방법을 활용한 알고리즘이다[2][3].

(알고리즘 1 ; Accept-Reject 알고리즘)

1. Generate X ~ g, U ~ U[0,1].
2. Accept Y=X if ($U \le \dfrac{f(x)}{Mg(x)}$)
3. Go to 1 otherwise.

2. 본 론

2.1 Bayesian 추론

Bayesian 정리는 다음과 같이 표현될 수 있다[4][5].

$$p(\theta|y) = l(\theta|y)p(\theta) \qquad (2)$$

여기서, $l(\theta|y)$는 $p(y|\theta)$를 y의 함수로 보지 않고 θ 의 함수로 해석한 것이다. 즉, y가 주어졌을 때 θ 에 대한 가능성(likelihood) 함수로 간주한 것이다[4][5][6]. 일반화된 식은 다음과 같다.

$$p(\theta|y_1,y_2,\ldots,y_m) \propto p(\theta|y_1,y_2,\ldots,y_{m-1})l(\theta|y_m) \qquad (3)$$

즉, Bayesian 정리는 근본적으로 경험으로부터 학습하는 과정을 나타낸다고 할 수 있다. 즉, Bayesian 정리는 새로운 데이터에 의한 지식 $l(\theta|y_m)$ 을 이전의 지식 $p(\theta|y_1,\ldots,y_{m-1})$ 에 더하여 새로운 지식 $p(\theta|y_1,\ldots,y_m)$ 을 구성하는 과정이라고 할 수 있다[2].

Bayesian 추론은 일반적으로 Bayesian 학습 네트워크를 구성하여 진행하다[7][8][9]. 그러나 일반 통계 추론의 경우에도 그래프 혹은 네트워크를 구성하여 수행하는 경우도 있다[10].

Bayesian 추론은 지식 학습과도 관련이 있지만 인과 관계(causality) 발견에도 적용할 수 있다. 인과 관계 발견과 관련된 문제에는 모델 선택과 모델 평균화, 이전(prior) 값 결정과 이전 구조 결정, 그리고 상실된(missing) 데이터와 감춰진(latent) 변량 문제 등이 있다[9].

만약 n개의 변량이 있을 때, 가능한 모든 모델의 개수는 n! 이다. 따라서 이것을 계산적으로 모두 조사해 보는 것은 매우 많은 시간을 필요로 한다. 따라서 이 문제를 해결하기 위하여 모델 선택이라는 방법을 사용한다. 이 개념은 여러 개의 모델 중에서 좋은 모델을 선택하여 사용하는 것이다. 또 다른 방법으로는 모델 평균화 방법이 있다. 이 방법은 여러 모델 중에서 가능성이 있는 몇 개의 모델을 선정하고, 이 모델들을 사용하여 계산을 수행하고 평균을 구하는 것이다[9].

2.2 Markov Chain과 Monte Carlo 적분

다음은 유한 Markov chain의 중요한 정의와 정리들이다[11][12].

(정의1 ; 유한상태 Markov chain) P는 $\{P_{i,j}\ i,j = 1,\ldots,k\}$를 요소로 가지는 $k \times k$ 상태변화 매트릭스라고 하자. 그러면 만약, 모든 n에 대하여, 모든 $i,j \in \{1,2,\ldots,k\}$이고 또한 모든 $i_0,i_1,\ldots,i_{n-1} \in \{1,2,\ldots,k\}$이라면, 유한 상태 $S=\{s_1, s_2, \ldots, s_k\}$를 가지는 무작위(랜덤) 프로세스 $(X_o, X_1, \ldots)$는 P를 동질성(homogeneous) 상태 변화 매트릭스로 가지는 Markov chain 이라고 한다. 즉,

$$\begin{aligned} P(X_{n+1} = s_j | X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}, X_n = s_i) \\ = P(X_{n+1} = s_j | X_n = s_i) \\ = P_{i,j} \end{aligned} \qquad (4)$$

(정리 1) 유한 상태 $S=\{s_1, s_2, \ldots, s_k\}$, 초기 분포 $\mu^{(0)}$, 그리고 상태 변화 매트릭스 P를 가지는 Markov chain $(X_o, X_1, \ldots)$에 대하여, 어떤 시간 n에 있어서의 $\mu^{(n)}$ 은

$$\mu^{(n)} = \mu(0)P^n \qquad (5)$$

을 만족한다.

Markov chain은 Bayes 추론 그리고 Monte Carlo 방법과 함께 공학, 의학, 생물학, 사회 과학 등 여러 분야에 응용된다[3][9][11][12][13][14][15][16][17].

Bayesian 추론은 다음과 같은 방정식으로 표현될 수 있다. 이 방정식은 직접 계산하기 어려운 경우가 많으며, 이러한 경우는 시뮬레이션 방법이 유용하다[3][12][13].

$$\min_\delta \int_\Theta L(\theta,\delta)\pi(\theta)f(x|\theta)d\theta \qquad (6)$$

시뮬레이션을 위한 Monte Carlo 적분식은 일반적으로 다음과 같

이 나타난다.

$$E_f[h(X)] = \int_{\chi} h(x)f(x)dx \quad (7)$$

위 식에 대하여 분포 f(x) 로부터 표본 $\{X_1, X_2, \ldots, X_m\}$을 구하고 다음 식을 이용하면 근사값을 구할 수 있다.

$$h_m = \frac{1}{m}\sum_{j=1}^{m} h(x_j) \quad (8)$$

여기서 h_m은 $E_f[h(X)]$ 에 수렴한다. 즉,

$$E_f[h(X)] \approx h_m \quad (9)$$

이다. 이런 방법을 고려하는 이유는, h(X)로부터 샘플 $\{x_t\}$를 독립적으로 구하는 것이 어려운 경우가 있기 때문이다. 왜냐하면 분포 함수 f(x)가 적분하기에 어려운 비표준 함수 형태인 경우가 많기 때문이다. 그러나 $\{x_t\}$가 반드시 독립적일 필요는 없으며, 사실 대략적으로 보면 $\{x_t\}$는 f(x)가 정의되어 있는 부분에서 값을 샘플링 할 수 있는 어떤 함수에 의하여서도 정확한 비율로 생성되기만 하면 된다. 이러한 과정을 실현시킬 수 있는 방법이 f(x)를 안정된 분포로 가지는 Markov chain 이다. Markov chain에서는 $\{x_t\}$는 시간이 지남에 따라 근사적으로 분포 h(X)에 의한 샘플이 된다. 따라서 어느 정도의 burn-in 시간이 지나면, $\{x_t ; t=m+1, \ldots, n\}$은 근사적으로 분포 h(X)에 의한 샘플이 된다. 다음은 burn-in 시간이 지난 후의 어고딕(ergodic) average 를 나타낸다[14][15][16][17].

$$\bar{f} = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X_t) \quad (10)$$

2.3 Markov Chain Monte Carlo

MCMC(Markov Chain Monte Carlo)는 안정적인 분포를 가지는 Markov chain을 구축하고 이를 기반으로 Monte Carlo 시뮬레이션을 구현한 것이다.

MCMC 방법이 성공적이 되기 위해서는 앞 절에서 설명한 ergodic average가 최초의 분포가 어떤 분포이었든지 간에 수렴하여야 한다. 이 수렴 조건을 충족시키는 최소의 조건은 Markov chain의 비축소성(irreducibility)이다. 즉, Markov chain은 모든 가능한 상태에 반드시 도달할 수 있어야 한다. 상태변화 매트릭스 P가 분포 f(x)에 대하여 비축소성이 될 충분 조건은 P^n 이 분포 f(x)에 대하여 양수 함수를 가지는 것이다. 순환(recurrence)은 어떤 출발점에서 출발하던지 Markov chain이 모든 chain을 방문할 수 있다는 의미이다. 대체적으로 비축소적이고 비주기적인 Markov chain은 수렴한다.

다음의 예는 MCMC 방법을 적용하여 X_{i+1} 샘플을 구하는 방법이다[13].

(예 1) 함수 f(x)를 實數에 대한 확률분포라고 하고 f(x)=cg(x)라고 하자. 여기서 c>0이다. 물론, 정의상 $\int f(x)dx = 1$ 이므로 c= 1/$\int g(x)dx = 1$ 이다. 그러나 이러한 적분 계산이 가능하지 않을 수도 있고 혹은 c값을 알 필요가 없을 수도 있다. 이제 X_0 를 임의의 값이라고 하고, $X_0, X_1, \ldots, X_i$ 가 주어져 있을 때, X_{i+1} 값을 샘플링하는 방법을 다음과 같이 생각해 보자. 첫째로, W를 샘플링하자. 즉, $W \sim N(X_i, b^2)$, 여기서 b>0이고 상수이다. 그리고,

$$r = \min\left\{\frac{g(W)}{g(X_i)}, 1\right\} \quad (11)$$

라고 하고, U값을 추출하자. 즉, U~Uniform[0,1]. 그리고 다음과 같이 설정한다.

$$X_{i+1} = \begin{cases} W \text{ if } U < r \\ X_i \text{ if } U \geq r \end{cases} \quad (12)$$

이러면 $X_0, X_1, \ldots, X_i$ 는 약한 조건하에서 안정된 분포 f(x) 로부터 추출된 어고딕(ergodic) Markov chain 이라고 할 수 있고, X_{i+1}은 분포 f(x) 로부터 추출된 것으로 간주할 수 있다. 즉, 무작위 변수 $X_0, X_1, \ldots, X_i$ 가 분포 f(x) 로부터 추출되고 Markov chain 성격을 가지면, 다루기 용이한 분포 g(x) 이용하여 분포 f(x) 를 가지는 변수 X_{i+1}을 근사적으로 구할 수 있다. 실험에서 상실된(missing) 데이터를 근사적으로 구할 때에도 이 방법을 사용하기도 한다.

2.4 Metropolis-Hastings 알고리즘

이 알고리즘은 객관적 혹은 목표 분포라는 확률 분포를 가지고 수행된다. 다음에 제안 확률 분포 q(y|x) 를 선택한다. 알고리즘은 제안 확률 분포 q(・|x) 가 명시적으로 존재하고 시뮬레이션하기 쉬운 형태이거나 혹은 대칭일(symmetric) 때 구현하기도 용이하다. 좀 더 일반적인 조건은 x에 독립적으로 비율 f(y)/q(y|x) 가 상수로 존재하는 경우이다. 즉, 목표 확률 분포와 제안된 확률 분포를 가지는 Metropolis-Hastings 알고리즘은 MCMC 방법으로서 Markov chain ($X^{(t)}$)을 생성한다[14][15][16][17].

(알고리즘 2 ; Metropolis-Hastings 알고리즘)

Objective(target) 분포 f를 가진 $x^{(t)}$ 가 주어졌을 때.

1. Generate $Y_t \sim q(y|\, x^{(t)})$
2. Accept $X^{(t+1)} = Y_t$ 확률값 $\rho(x^{(t)}, Y_t)$ or
   $X^{(t+1)} = x^{(t)}$ 확률값 $1-\rho(x^{(t)}, Y_t)$

여기서 $\rho(x,y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$ 이며,

분포 q 는 instrumental(proposal) 분포이다.

분포 q(y|x)는 일반적으로 $N(x, b^2)$을 선택한다. b>0 이다. 즉, 이 경우에는 샘플링 되는 값은 제안된 분포인 x축의 중앙에 위치한 정규분포로부터 샘플링 된다. 만약, 이 경우에 제안된 분포 q가 대칭이면 위의 $\rho(x,y)$는 다음과 같이 간단한 모양이 된다.

$$\rho(x,y) = \min\left\{\frac{f(y)}{f(x)}, 1\right\} \quad (13)$$

Markov chain의 분포가 f(x)가 되는 것을 증명할 수 있다. 즉, Metropolis-Hastings 알고리즘의 상태변화 커널(transition kernel)은 다음과 같다.

$$P(X_{t+1}|X_t) = q(X_{t+1}|X_t)\alpha(X_t, X_{t+1}) + I(X_{t+1} = X_t)[1 - \int q(Y|X_t)\alpha(X_t, Y)dY] \quad (14)$$

여기서 $I(.)$는 indicator 함수이다. 위 식에서 첫 번째 항은 $Y = X_{t+1}$ 인 경우이고, 두 번째 항은 그 반대의 경우이다. 다음의 관계를 이용하여

$$f(X_t)P(X_{t+1}|X_t) = f(X_{t+1})P(X_t|X_{t+1}) \quad (15)$$

식의 양 쪽을 X_t 로 적분하면,

$$\int f(X_t)P(X_{t+1}|X_t) = f(X_{t+1}) \quad (16)$$

이 된다. 식의 왼쪽은 X_t 의 분포가 f(x)라고 할 경우에 있어서의 X_{t+1}의 marginal 분포를 나타낸다. 따라서 위의 식은 만약에 X_t 의 분포가 f(x)라고 한다면, X_{t+1}의 분포도 f(x)라는 것을 의미한다. 즉, 만약에 하나의 샘플이 안정된 분포로부터 샘플링 되었다면, 그 이후의 샘플들도 동일한 분포로부터 샘플링 될 수 있다는 의미이다[14].

2.5 Gibbs 샘플링

Gibbs 샘플링은 다차원 MCMC 문제를 여러 개의 일차원 MCMC 문제로 변형시키는데 사용할 수 있다. 우선 두 개의 변량으로 구성된 간단한 Gibbs 샘플링 예를 보자[3][12][13][14][15][16][17][18].

(예 2) (X, Y)가 분포 $f_{X,Y}(x,y)$ 를 가진다고 하자. 그리고 조건부 확률 분포 $f_{X|Y}(x|y)$와 $f_{Y|X}(y|x)$로부터 샘플링이 가능하다고 하자. (X_0, Y_0)가 초기 값이라고 하고, $(X_0, Y_0), \ldots, (X_n, Y_n)$을 샘플링 했다고 하자. 그러면, Gibbs 샘플링 알고리즘은 다음의 방식으로 (X_{n+1}, Y_{n+1})을 샘플링 한다[11]. 즉,

$$X_{n+1} \sim f_{X|Y}(x|Y_n)$$
$$Y_{n+1} \sim f_{Y|X}(y|X_{n+1}) \quad (17)$$

이제 일반적인 경우의 Gibbs 샘플링을 보자. 어떤 정수 p>1에 대하여, 무작위 변량 $\mathbf{X} \in \chi$ 가 $\mathbf{X}=(X_1, X_2, \ldots, X_p)$로 나타나고, X_i는 단차원이거나 혹은 다차원이라고 가정한다. 더욱이, 다음과 같이 해당되는(corresponding) 단변량 (univariate) 조건부 확률 $f_1, f_2, \ldots, f_p$ 로부터 시뮬레이션을 할 수 있다고 한다[3].

(알고리즘 3 ; Gibbs 샘플링)

$x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_p^{(t)})$ 가 주어졌을 때, generate

1. $X_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)})$
2. $X_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$

$\vdots$

p. $X_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$

Gibbs 샘플링은 완성(completion) 구축에 의하여 일반화될 수 있다.

(정의 2 ; Completion) : 확률 분포 f(x)가 주어지고, 다른 분포 g(x)가 다음 식을 만족할 때, g(x)를 f(x)의 완성이라고 한다.

$$\int_Z g(x,z)dz = f(x) \quad (18)$$

분포 g(x)는 계산하기 쉬운 분포를 선택하고, Gibbs 알고리즘은 분포 f(x) 대신에 분포 g(x)에 구현된다. 정수 p>1에 대하여, y=(x,z)라고 하고, g(y)=g(y_1, ... , y_p) 에 대한 조건부 확률을 다음과 같이 생각하자.

$$Y_1|y_2,\ldots,y_p \sim g_1(y_1|y_2,\ldots,y_p)$$
$$Y_2|y_1,y_3,\ldots,y_p \sim g_2(y_2|y_1,y_3,\ldots,y_p)$$
$$\vdots$$
$$Y_p|y_1,y_3,\ldots,y_{p-1} \sim g_p(y_p|y_1,y_2,\ldots,y_{p-1}) \quad (19)$$

그러면, $Y^{(t+1)}$에 대한 Gibbs 샘플링은 다음 알고리즘으로 제시된다[10].

(알고리즘 4 ; Completion Gibbs 샘플링)

$Y^{(t)}$= ($y_1^{(t)}$, $y_2^{(t)}$, ... $y_p^{(t)}$) 가 주어졌을 때, generate

1. $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)},y_3^{(t)},\ldots,y_p^{(t)})$
2. $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)},y_3^{(t)},\ldots,y_p^{(t)})$

$\vdots$

p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^{(t+1)},y_2^{(t+1)},\ldots,y_{p-1}^{(t+1)})$

완성 개념은 상실된(missing) 데이터 모델과 관련이 있다[3][14].

2.6 Maximum Likelihood Estimation

K 개의 확률 분포가 있고 m 개의 변량이 있다고 하자. 즉, $\Theta = (\theta_1,\theta_2,\ldots,\theta_K)$ 이고 X=(x_1, x_2, ... , x_m) 라고 하자. 그러면 $p(x_i|\theta_j)$ 는 θ_j 분포에 대한 변량 x_i 의 조건부 확률이며, 분포 θ_j 에 대한 선택을 또 다른 확률로 나타내고, 만약 변량들이 독립적으로 생성된다면, 전체 변량에 대한 확률은 각 변량의 확률을 곱한 것으로 다음과 같이 표현된다[19].

$$p(X|\Theta) = \prod_{i=1}^{m}\sum_{j=1}^{K} w_j p_j(x_i|\theta_j) \quad (20)$$

그리고 m 개의 변량이 독립적으로 동일한 Gaussian 분포로부터 생성되었다고 한다면, 모든 m개의 변량에 대한 확률은 Gaussian 분포의 곱으로 다음과 같이 표현된다.

$$p(X|\Theta) = \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \quad (21)$$

이 함수를 log 함수를 적용하여 다음과 같이 나타낸다.

$$\log(p(X|\Theta)) = -\sum_{i=1}^{m}\frac{-(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma \quad (22)$$

일반적으로 위의 식을 최대화하는 μ 와 σ 를 구하는 것이 목적이다. 이것을 최대화 가능성 원리 (maximum likelihood principle) 라고 하고 데이터로부터 이러한 값을 구하는 과정을 MLE(Maximum Likelihood Estimation)라고 한다.

2.7 EM 알고리즘

MLE 개념을 확장하여 혼합 확률 분포에 의하여 생성된 변량에 적용하여 혼합 모델의 파라메타를 추정할 수 있다. 이러한 경우에 사용할 수 있는 알고리즘이 EM(Expectation Maximization) 알고리즘이다[2][3]. EM 알고리즘은 E-step과 M-step 2단계로 구성된다. 첫 번째 E-step 에서는 데이터 그리고 데이터에 대한 파라메타의 추정 값을 사용하여 데이터에 대한 분포의 확률을 계산한다. 다음에 M-step에서는 E-step에서 계산된 확률을 기반으로 가능성 기댓값(expectation of likelihood)을 최대화할 수 있는 새로운 파라메타의 추정치를 찾는다. 이 두 과정을 파라메타가 더 이상 변하지 않을 때까지 반복한다. 이런 과정을 반복하여 혼합 분포에 의하여 생성된 데이터에 대한 분포의 파라메타를 추정한다. 다음은 EM 알고리즘이다[2][3][16][17][20][21][22][23][24][25][26]. K-means 알고리즘은 EM 알고리즘의 특수한 형태로 볼 수 있다[24][25][26].

(알고리즘 6 ; EM 알고리즘 1) // Complete 데이터

분포에 대한 초기값 선정.

1. Compute (E-step)
   $Q(\Theta,\theta_j) = E[\log L(\Theta|x_i,\theta_j)]$
2. Maximize (M-step)
   $Q(\Theta,\theta_j)$를 최대화할 수 있는 새로운 예측값 θ_{j+1}을 찾는다.

그러나 EM 알고리즘을 상실된 데이터가 존재하는 경우에도 적용할 수 있다. 즉, EM 알고리즘은 분포 파라메타의 예측과 함께 상실된 데이터의 예측에도 사용할 수가 있다. 왜냐하면 EM 알고리즘은 예측값을 초기값으로 시작하여 E-step과 M-step을 계속 수행하면서 최대값을 구하는 과정이기 때문이다. 즉, y=(x, z) 라고 하자. 여기서, y는 관찰된(observable or incomplete) 데이터, x는 완전한(complete) 데이터, 그리고 z는 상실된(missing or latent) 데이터를 나타낸다. 그리고 $f(x|\theta)$를 x의 확률 밀도 함수, $g(y|\theta)$를 y의 확률 밀도 함수, 그러면,

$$g(y|\theta) = \int_y f(x|\theta)dx \quad (23)$$

이 성립한다. 그리고 참고문헌[18]에서는 (1) 예측되는 분포의 파라메타는 상실된 데이터와는 무관하다고 가정하고, (2) 상실된 데이터는 무작위로 상실되었다고 가정한다. 그러면 EM 개념에 의하여 다음 수식들이

$$Q(\theta|\theta^k) = E[\log L(\theta|x)|y,\theta^k], \quad (24)$$
$$\theta^{k+1} = \arg\max_{\theta\in\Theta} E[\log L(\theta|x)|y,\theta^k] \quad (25)$$

성립한다. EM 알고리즘에 대한 자세한 논의는 참고문헌[2][3][13][16][17][20][21][22][23][24][25]에 제시되어 있다.

그 밖에 Monte Carlo EM에 대하여 알아보자. EM 알고리즘의 문제점은 E-step에서 log likelihood 의 기대값을 계산하여야 한다는 것이다. 이 문제는 상실된 데이터 z_1, ... , z_m 값들을 조건부 확률 $g(y,\theta^k)$ 로부터 시뮬레이션을 통하여 구함으로서 해결할 수 있다. 그리고 다음의 값에 대하여 최대값을 구한다.

$$\hat{Q}^{(t+1)}(\theta|\theta^t) = \frac{1}{m}\sum_{i=1}^{m}\log L(\theta|x) \quad (26)$$

m이 무한대로 접근하면, 위의 양은 $\Theta(\theta|\theta_0,x)$ 에 수렴하며 Monte Carlo EM 알고리즘의 극한값은 원래의 EM 알고리즘 값에 수렴한다.

2.8 상실된 데이터 처리

통계 조사를 위한 데이터 수집에서 데이터가 제대로 측정되지 않는 경우는 많이 발생한다. 이러한 상실된 데이터를 해결하는 방법으로는 다음의 방법이 있다. (1) 측정되지 않은 데이터를 포함한 측정 단위를 고려하지 않는 방법이 있다. 이 방법은 상실된 데이터가 많지 않을 경우에 사용할 수 있다, (2) 상실된 데이터를 채워 넣는 방법이 있을 수 있다. 이 경우에는 측정된 데이터들로부터 평균, 혹은 회귀 값을 계산하여 사용할 수 있다, (3) 샘플 설계 가중치(design weight)를 이용하는 방법이 있다. 가중치(π_i^{-1})와 측정된 데이터 비율($\hat{p}_i^{-1}$)을 사용하여 측정된 값들의 평균을 구한 다음에, 이 평균을 상실된 값으로 사용할 수 있다, 그리고 (4) 모델을 사용하는 방법이 있다. 이 방법은 maximum likelihood 를 이용한 방법이다[25].

또한 단변량(univariate)만을 측정 대상으로 하고, 그 단변량에서 데이터 상실이 발생하는 경우가 있고, 혹은 여러개의 변량을 측정대상으로 하지만, 하나의 변량에서만 데이터 상실이 발생하는 경우도 있다. 혹은 다변량(multivariate)의 경우에 여러 다변량에서 데이터 상실이 발생할 수도 있다.

상실된 데이터 모델은 다음과 같이 가능성(likelihood)이 표시되거나,

$$g(x|\theta) = \int_Z f(x,z|\theta)dz \quad (27)$$

혹은, 최적화될 함수 h(x)가 다음과 같은 기대값 형태로 표시되는 경우라고 할 수 있다.

$$h(x)=E[H(x, Z)] \quad (28)$$

여기서 z 혹은 Z에 의하여 표시되는 변량이 상실된 데이터를 나타낸다. 참고문헌 [28][29]은 주로 사회 조사의 경우에 데이터 수집 과정에서 상실된 데이터를 다루는 방법을 상세히 설명하고 있다. EM 알고리즘을 사용한 상실된 데이터 보완 문제는 참고문헌[18]에 의하여 체계화되고 일반화된 해결 방법이 마련되어, 통계적인 추론을 가능하게 하였다[19]. 그러나 원래의 EM 알고리즘은 복잡한 계산을 수행하여야 하기 때문에 모든 경우에 사용될 수

는 없다. Monte Carlo EM 알고리즘을 사용하여 시뮬레이션에 의하여 상실된 데이터를 보완하여 EM 알고리즘을 사용할 수 있다 [3][16][25].

2.9 BMA(Bayesian Model Averaging)

일반적인 통계 추론에서 모델의 불확실성을 고려하지 않는 경우가 많다. 즉, 분석자들은 모델의 집합으로부터 모델을 선정하여 분석을 진행하며, 모델 선정 과정의 불확실성을 고려하지 않는 경우가 많다. 이러한 경우에 선정된 모델은 과도한 추론을 가능하게 하며 또한 위험한 결정을 내리게 할 수도 있다. BMA(Bayesian Model Averaging)는 이러한 모델 불확실성을 해결할 수 있는 방법을 제공한다[16][17][30][31].

BMA를 다음과 같이 수식으로 표현하여 설명할 수 있다. 만약 Δ가 중요한 변량이라고 하면 데이터 D가 주어졌을 때, Δ에 대한 이후(posterior) 확률은 다음과 같다.

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|M_k, D)p(M_k|D) \tag{29}$$

여기서 M_k는 모델이다. 그리고 모델 M_k의 이후 확률은 Bayesian 정리에 의하여 다음과 같다.

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{n=1}^{K} p(D|M_n)p(M_n)} \tag{30}$$

그리고,

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \tag{31}$$

은 적분식으로 표시된 모델 M_k의 likelihood 식이며, θ_k 는 모델 M_k의 파라메타이다. $p(\theta_k|M_k)$는 모델 M_k 에 대한 파라메타 θ_k 의 이전 확률이며, $p(D|\theta_k, M_k)$는 likelihood 이다. 그리고 $p(M_k)$는 모델 M_k의 이전 확률이다. 그러면, $\Delta|D$ 의 평균과 분산은 다음과 같다.

$$E[\Delta|D] = \sum_{k=1}^{K} E[\Delta|D, M_k]p(M_k|D) \tag{32}$$

$$\begin{aligned} Var[\Delta|D] &= \sum_{k=1}^{K} (Var[\Delta|D, M_k] + E[\Delta|D, M_k]^2)p(M_k|D) \\ &\quad - E[\Delta|D]^2 \end{aligned} \tag{33}$$

즉, 이런 식으로 모든 모델들에 대하여 평균을 구하는 것이 더 우수한 평균적인 예측 능력을 보여준다는 것이 제시되었다.

BMA를 구현하기 위해서는 위 (31) 식을 계산해야 하는데, 고려해야만 하는 모델의 수가 너무 많으면 계산이 어려워진다. 계산에 필요한 모델의 수를 감소시키는 방법으로는, (1) 전체 모델 중에서 데이터에 의하여 지원되는 모델만을 고려하는 방법이 있고, (2) Markov Chain Monte Carlo 방법인 Metropolis-Hastings 알고리즘을 사용하여 수식 (31)을 근사적으로 계산하는 방법이 있다. 또 다른 문제는 수식 (33) 계산의 어려움이 있다. 이 문제를 해결하기 위하여 Laplace 근사, MLE 근사 등을 사용한다[30].

2.10 PC 분석

주요 요소(Principal Component) 분석은 많은 수의 서로 관련되어 있는 변수 집합을 차원(dimensionality)이 적은 다른 변수 집합으로 변환시켜서 표현과 분석이 용이하도록 하는 것이다. 이 변환은 원래의 변수 집합이 가지고 있는 변수들의 변이(variation)를 되도록 모두 가지고 있어야 하며, 새로운 변수들은 되도록 서로 관련이 적어야하며, 변수들 간에는 일종의 순서가 있어서 처음 몇개의 변수들이 변환되기 이전에 원래의 변수들이 가지고 있었던 변이들을 모두 가지고 있어야 한다[33][35].

다음은 공분산(covariance)을 사용한 PC 분석 예를 보여준다 [33]. 공분산을 S 라고 하면, S는 다음의 식에 의하여 대각선 (diagonal) 매트릭스 L로 변환될 수 있다.

$$U^t S U = L \tag{34}$$

대각선 매트릭스 L의 원소 $l_1, l_2, ..., l_n$ 은 eigenvalue라고 한다. 매트릭스 U의 열 벡터인 원소 $u_1, u_2, ..., u_n$는 eigenvector 라고 한다. 대각선 매트릭스를 구성하는 원소인 eigenvalue는 다음 식을 통하여 구할 수 있다.

$$|S - lI| = 0 \tag{35}$$

매트릭스 U를 구성하는 eigenvector는 다음 두 식으로 구할 수 있다.

$$(S - lI)t_i = 0 \tag{36}$$

그리고

$$u_i = \frac{t_i}{\sqrt{t_i^t t_i}} \tag{37}$$

그러면,

$$|u_i| = 1, \quad u_i u_j = 0 \tag{38}$$

이다. u_i 와 u_j 는 데이터들의 기본(basis) 벡터가 되어, 데이터들의 주요 요소(principal component)가 된다. 이 밖에도 correlation을 사용하여 PC를 정의할 수도 있다[30][31]. 참고문헌 [34]는 PC 분석 기법을 개념을 학습하는 문제에 적용한 예를 제시하고 있다.

3. 결 론

본 논문에서는 다양한 분야에서 활용되는 시뮬레이션 방식의 Bayesian 추론에 대하여 논의하였다. Bayesian 추론은 Markov Chain Monte Carlo 시뮬레이션 기법과 결합되어 응용력이 높아지면서 많은 분야에 적용되어 좋은 효과를 내게 된 것으로 보인다. 잘 준비된 데이터를 대상으로 이러한 방법을 적용하면 좋은 결과를 내게 될 것으로 보인다.

참고 문헌

[1]Athanasios Papoulis, S. U. Pillai, Probability, Random Variables and Stochastic Processes(4th ed.), McGraw-Hill, 2002.
[2]Robert V. Hogg, Joseph W. McKean, Allen T. Craig, Introduction to Mathematical Statistics(6th ed.), Prentice-Hall, 2005.
[3]Christian P. Robert, George Casella, Monte Carlo Statistical Methods(2nd ed.), Springer-Verlag, 2004.
[4]George E. P. Box, George C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, 1973.
[5] Andrew Gelman, John B. Carlin, Hal S. Stern, Donald B. Rubin, Bayesian Data Analysis, Chapman & Hall, 1997.
[6] Thomas A. Severini, Likelihood Methods in Statistics, Oxford University Press, 2004.
[7] Richard E. Neapolitan, Learning Bayesian Networks, Prentice-Hall Inc., 2004.
[8]Kevin B. Korb, Ann E. Nicholson, Bayesian Artificial Intelligence, Chapman and Hall/CRC Press, 2004.
[9]Clark Glymour, Gregory F. Cooper, Computation, Causation, and Discovery, AAAI Press/The MIT Press, 1999.
[10] David Edwards, Introduction to Graphical Modelling(2nd ed.), Springer Verlag, 2000.
[11] Olle Haggstorm, Finite Markov Chains and Algorithmic Applications, Cambridge University Press, 200.
[12]Pierre Bremaud, Markov Chains Gibbs Fields, Monte Carlo Simulation, and Queues, Springer Verlag, 1999.
[13] Larry Wasserman, All of Statistics A Concise Course in Statistical Inference, Springer Verlag, 2004.
[14]W. R. Gilks, S. Richardson, D. J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman & Hall, 1996.
[15] B. Walsh, "Markov Chain Monte Carlo and Gibbs Sampling," Lecture Notes for EEB 581, April 2004.
[16]Geof H. Givens, Jennifer A. Hoeting, Computational Statistics, John Wiley & Sons, 2005.
[17]Ming-Hui Carlo, Qi-Man Shao, Joseph G. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer-Verlag, 2000.
[18] Stuart Geman, Donald Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, No. 6, pp.721-741, November 1984.
[19]Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006.
[20] A. P.Dempster, N. M. Laird, D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, pp.1-38, 1977.
[21] Michiko Watanabe, Kazunori Yamaguchi, The EM Algorithm and Related Statistical Models, Marcel Dekker, Inc., 2004.
[22] Martin A. Tanner, Tools for Statistical Inference, Lecture Notes in Statistics, Vol. 67, Springer-Verlag, 1992.
[23] David J. C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.
[24] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Inference, Springer-Verlag, 2001.
[25]Roderick J. A. Little, Donald B. Rubin, Statistical Analysis with Missing Data(2nd ed.), John Wiley & Sons, 2002.
[26] Geoffrey J. McLachlan, Thriyambakam Krishnan, The EM algorithm and extensions, Wiley-Interscience, 1996.
[27]Vicki A. Freedman, Douglas A. Wolf, "A Case Study on the Use of Multiple Imputation," Demography, Vol. 32, No. 3, pp.459-470, August 1995.
[28]Eric Schulte Nordholt, "Imputation : Methods, Simulation Experiments and Practical Examples," International Statistical Review, Vol. 66, No. 2, pp.157-180, August 1998.
[29]Greg C. G. Wei, Martin A. Tanner, "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," Journal of the American Statistical Association, Vol. 85, No. 441, pp.699-704, September, 1990.
[30]Jennifer A. Hoeting, et al., "Bayesian Model Averaging : A Tutorial," Statistical Science, Vol. 14, No. 4, pp.382-417, 1999.
[31] Jennifer A. Hoeting, "Methodology for Bayesian Model Averaging : An Update," www.stat.colostate.edu/~jah/.
[32]I.T. Jolliffe, Principal Component Analysis (2nd ed.), Springer-Verlag, 2002.
[33]J. Edward Jackson, A User's Guide to Principal Components, John-Wiley & Sons, 1991.
[34] Philippe, G. Schyns, Luc Rodet, "Concept Learning", in The Handbook of Brain Theory and Neural Networks (Michael Arbib editor), The MIT Press, 1995.
[35] Richard Duda, Peter E. Hart, David G. Stork, Pattern Classification(2nd ed.), Wiley-Interscience, 2000.