3부 : 베이지안 분석

김용대[1]

서울대학교 통계학과[1]

목차

들어가며

- **확률변수** : 일정한 확률에 따라 값이 결정되는 변수
- **확률분포** : 확률변수의 값을 결정하는 확률을 나타내는 함수
  - **확률질량함수** : 확률변수가 이산적인 값을 가질 때, 각각의 값에 확률을 부여
    - $X \sim Poisson(\lambda)$:

$$
\begin{aligned}
p(k) &= \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, \ldots \\
P(a < X < b) &= \sum_{a<k<b} p(k)
\end{aligned}
$$

  - **확률밀도함수** : 확률변수가 연속적인 값을 가질 때, 구간에 확률 밀도를 부여
    - $X \sim Normal(\mu, \sigma^2)$:

$$
f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad P(a < X < b) = \int_a^b f(x)dx
$$

- **확률분포의 모수** : 확률분포를 결정하는 값 혹은 벡터
  - 푸아송분포에서는 λ, 정규분포에서는 (μ, σ^2)

들어가며

통계적 분석 방법

1. 자료 $X_1, \ldots, X_n$을 확률변수들의 관측값으로 간주

2. 확률변수들이 따르는 모형(확률분포) P_θ(θ는 모수) 설정

3. 관측된 자료들로 모수 θ를 추측

   – 빈도주의 : θ가 고정된 값으로 가정하고, 이를 추정

   $$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P_\theta(X_1, \ldots, X_n)$$

   – 베이지안 : θ를 확률변수라 가정하고, 이것의 분포를 추정

   $$P(\theta | X_1, \ldots, X_n)$$

목차

조건부 확률

- **조건부 확률** $P(A|B)$: 사건 B가 발생했을 때, 사건 A가 발생할 확률

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

조건부 확률

Q. 어떤 사람이 특정 질병을 가지고 있을 때, 검사를 통해서 제대로 판정받을 확률이 95%이다. 이 검사는 1%의 확률로 실제로 건강한 사람이 그 질병을 가지고 있다고 잘못 판정한다. 어떤 사람이 검사를 통해 질병을 가지고 있다고 판정 받았을 때, 실제로 그 질병을 가지고 있을 확률은 얼마인가? (실제 인구의 0.5%가 해당 질병을 가지고 있다)

$A = \{$질병을 가지고 있음$\}$
$B = \{$검사에서 질병이 있다고 판정$\}$
$P(A) = 0.005$, $P(B|A) = 0.95$, $P(B|A^c) = 0.01$

$$P(A|B) = ?$$

조건부 확률

$$
\begin{aligned}
P(A \cap B) &= P(B|A)P(A) = 0.00475 \\
P(B) &= P(A \cap B) + P(A^c \cap B) \\
&= P(B|A)P(A) + P(B|A^c)\{1 - P(A)\} = 0.0147 \\
P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{0.00475}{0.0147} \approx 0.323
\end{aligned}
$$

베이즈 정리

- **베이즈 정리**

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
P(\theta|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|\theta)P(\theta)}{P(X_1,\ldots,X_n)} \\
&\propto P(X_1,\ldots,X_n|\theta)P(\theta)
\end{aligned}
$$

- **사전분포** $P(\theta)$
  θ의 확률분포로 자료를 관측하기 전에 분석자의 θ에 대한 불확실성을 나타냄

- **우도함수** $P(X_1,\ldots,X_n|\theta)$
  θ가 자료를 잘 설명하는지를 나타냄

- **사후분포** $P(\theta|X_1,\ldots,X_n)$
  자료를 관측한 후에 분석자의 θ에 대한 불확실성을 나타냄

베이지안 추론

- 사후평균

$$\mathbb{E}(\theta|X_1,\ldots,X_n) = \int \theta P(\theta|X_1,\ldots,X_n)d\theta$$

- MAP(Maximum a Posteriori) 추정량

$$\underset{\theta}{\operatorname{argmax}}\, P(\theta|X_1,\ldots,X_n)$$

- 사후분산

$$\mathrm{Var}(\theta|X_1,\ldots,X_n) = \int \theta^2 P(\theta|X_1,\ldots,X_n)d\theta - \mathbb{E}(\theta|X_1,\ldots,X_n)^2$$

- Credible interval (L, U)

$$\int_L^U P(\theta|X_1,\ldots,X_n)d\theta = 0.95$$

베이지안 예측

- 사후예측분포

$$P(\tilde{X}|X_1, \ldots, X_n) = \int P(\tilde{X}|\theta)P(\theta|X_1, \ldots, X_n)d\theta$$

- 베이지안 추론과 예측 모두 사후분포를 통해 이루어짐

목차

여러가지 분포

- 이산확률분포
  - 이항분포

$$p(k;\theta) = \frac{N}{k!(N-k)!}\theta^k(1-\theta)^{N-k} \quad \text{for } k = 0, 1, \ldots, N$$

  - 푸아송분포

$$p(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, \ldots$$

  - 기하분포

$$p(k;\theta) = (1-\theta)^{k-1}\theta \quad \text{for } k \in \mathbb{N}$$

여러가지 분포

- 연속확률분포
  - 정규분포

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } x \in \mathbb{R}$$

  - 지수분포

$$f(x;\lambda) = \lambda e^{-\lambda x} \quad \text{for } x \in \mathbb{R}^+$$

  - 감마분포

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \quad \text{for } x \in \mathbb{R}^+ \\ \text{where} \quad \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1}e^{-x}dx \end{aligned}$$

  - 베타분포

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } x \in \mathbb{R}^+$$

이항분포-베타사전분포

- 확률모형

$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} Binomial(\theta)$$

- 사전분포

$$\theta \sim Beta(\alpha, \beta)$$

- 사후분포

$$\begin{aligned}
P(\theta | X_1, \ldots, X_n) &\propto P(\theta; \alpha, \beta) P(X_1, \ldots, X_n | \theta) \\
&= P(\theta; \alpha, \beta) \prod_{i=1}^{n} P(X_i | \theta) \\
&\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \prod_{i=1}^{n} \theta^{X_i} (1-\theta)^{N-X_i} \\
&= \theta^{\alpha + \sum_{i=1}^{n} X_i - 1} (1-\theta)^{\beta + nN - \sum_{i=1}^{n} X_i - 1} \\
\theta | X_1, \ldots, X_n &\sim Beta\left(\alpha + \sum_{i=1}^{n} X_i, \beta + nN - \sum_{i=1}^{n} X_i\right)
\end{aligned}$$

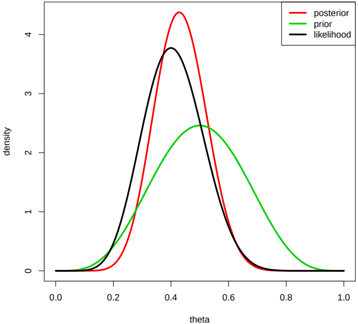
이항분포-베타사전분포 예제

- 동전의 앞면이 나올 확률 θ를 추정하고자 함 ($N = 1$)
- 앞면을 8번, 뒷면을 12번 관측 ($n = 20$, $\sum_{i=1}^{n} X_i = 8$)
- 사전분포 : $Beta(5, 5)$
- 사후분포 : $Beta(13, 17)$

```
a=13; b=17
a/(a+b) # mean

## [1] 0.4333333

a*b/((a+b)^2*(a+b+1)) # variance

## [1] 0.007921147

c(qbeta(0.025,a,b),qbeta(0.975,a,b)) # credible interval

## [1] 0.2644553 0.6106372
```

이항분포-베타사전분포 예제

```
x=seq(from=0,to=1,length.out=1000)
prior=dbeta(x,5,5)
post=dbeta(x,13,17)
loglik=dbeta(x,9,13)
plot(x,post,type="l",col=2,xlab="theta",ylab="density",lwd=3)
lines(x,prior,lwd=3,col=3)
lines(x,loglik,lwd=3,col=1)
legend("topright", c("posterior","prior", "likelihood"),
       lwd=rep(3,4), col=c(2,3,1))
```



켤레사전분포

- 이항분포 확률모형에서 사전분포가 베타분포일 때 사후분포 역시 베타분포이다.
- 이처럼 사후분포가 사전분포와 같은 종류의 분포일 때의 사전분포를 **켤레사전분포**라 부른다.
- 켤레사전분포를 사용하면 사후분포의 계산이 용이한 장점이 있다.

지수분포-감마사전분포

- 확률모형

$$X_1, \ldots, X_n | \lambda \overset{iid}{\sim} Exponential(\lambda)$$

- 사전분포

$$\lambda \sim Gamma(\alpha, \beta)$$

- 사후분포

$$\begin{aligned} P(\lambda | X_1, \ldots, X_n) &\propto P(\lambda; \alpha, \beta) \prod_{i=1}^{n} P(X_i | \lambda) \\ &\propto \lambda^{\alpha-1} e^{-\beta\lambda} \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \\ &= \lambda^{\alpha+n-1} e^{-(\beta+\sum_{i=1}^{n} X_i)\lambda} \\ \theta | X_1, \ldots, X_n &\sim Gamma\left(\alpha + n, \beta + \sum_{i=1}^{N} X_i\right) \end{aligned}$$

푸아송분포-감마사전분포

- 확률모형

$$X_1, \ldots, X_n | \lambda \overset{iid}{\sim} Poisson(\lambda)$$

- 사전분포

$$\lambda \sim Gamma(\alpha, \beta)$$

- 사후분포

$$\begin{aligned}
P(\lambda | X_1, \ldots, X_n) &\propto P(\lambda; \alpha, \beta) \prod_{i=1}^{n} P(X_i | \lambda) \\
&\propto \lambda^{\alpha-1} e^{-\beta\lambda} \prod_{i=1}^{n} \lambda^{X_i} e^{-\lambda} \\
&= \lambda^{\alpha + \sum_{i=1}^{n} X_i - 1} e^{-(\beta+n)\lambda}
\end{aligned}$$

$$\theta | X_1, \ldots, X_n \sim Gamma\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n\right)$$

정규분포-정규사전분포

- 확률모형

$$X_1, \ldots, X_n | \mu \overset{iid}{\sim} Normal(\mu, \sigma^2) \quad (\sigma^2 \ : \ \text{fixed})$$

- 사전분포

$$\mu \sim Normal(\mu_0, \sigma_0^2)$$

- 사후분포

$$P(\mu | X_1, \ldots, X_n) \propto P(\mu; \mu_0, \sigma_0) \prod_{i=1}^{n} P(X_i | \mu; \sigma)$$

$$\propto \exp\left\{ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right\} \exp\left\{ -\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2} \right\}$$

$$= \exp\left(-\frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{2\sigma_0^2} - \frac{n\mu^2 - 2\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i^2}{2\sigma^2} \right)$$

정규분포-정규사전분포

- 사후분포

$$
\begin{aligned}
& P(\mu|X_1, \ldots, X_n) \\
\propto\ & \exp\left\{ -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n X_i}{\sigma^2}\right) \right\} \\
\propto\ & \exp\left\{ -\frac{(\mu - \mu_n)^2}{2\sigma_n^2} \right\} \\
& \text{where } \sigma_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \quad \text{and} \\
& \qquad \mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\frac{\sum_{i=1}^n X_i}{n} \\
& \mu|X_1, \ldots, X_n \sim Normal(\mu_n, \sigma_n^2)
\end{aligned}
$$

정규분포-감마사전분포

- 확률모형

$$X_1, \ldots, X_n | \lambda \overset{iid}{\sim} Normal(\mu, 1/\lambda) \quad (\mu \ : \ \text{fixed})$$

- 사전분포

$$\lambda \sim Gamma(\alpha, \beta)$$

- 사후분포

$$\begin{aligned}
P(\lambda | X_1, \ldots, X_n) &\propto P(\lambda; \alpha, \beta) \prod_{i=1}^{n} P(X_i | \lambda; \mu) \\
&\propto \lambda^{\alpha-1} e^{-\beta\lambda} \lambda^{\frac{n}{2}} \exp\left\{ -\frac{\lambda \sum_{i=1}^{n} (X_i - \mu)^2}{2} \right\} \\
&= \lambda^{\alpha + \frac{n}{2} - 1} \exp\left[-\left\{ \beta + \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2} \right\} \lambda \right]
\end{aligned}$$

$$\lambda | X_1, \ldots, X_n \sim Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2} \right)$$

정규분포-정규감마사전분포

- 확률모형

$$X_1, \ldots, X_n | \mu, \lambda \overset{iid}{\sim} Normal(\mu, 1/\lambda)$$

- 사전분포

$$\lambda \sim Gamma(\alpha, \beta)$$
$$\mu|\lambda \sim Normal\left(\mu_0, (\kappa_0\lambda)^{-1}\right)$$

- 사후분포

$$P(\mu, \lambda | X_1, \ldots, X_n) \propto P(\mu|\lambda; \mu_0, \kappa_0) P(\lambda; \alpha, \beta) \prod_{i=1}^{n} P(X_i|\mu, \lambda)$$
$$\propto \lambda^{\frac{1}{2}} \exp\left\{-\frac{\kappa_0\lambda(\mu-\mu_0)^2}{2}\right\} \lambda^{\alpha-1} e^{-\beta\lambda} \lambda^{\frac{n}{2}} \exp\left\{-\frac{\lambda\sum_{i=1}^{n}(X_i-\mu)^2}{2}\right\}$$
$$= \lambda^{\alpha+\frac{n-1}{2}} \exp\left[-\left\{\beta + \frac{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{n}(X_i-\mu)^2}{2}\right\}\lambda\right]$$

- 사후분포 $P(\mu, \lambda | X_1, \ldots, X_n)$에서의 표본추출은 어려우나, 이것의 조건부분포 $P(\mu|\lambda, X_1, \ldots, X_n)$와 $P(\lambda|\mu, X_1, \ldots, X_n)$는 각각 정규분포와 감마분포이므로 표본추출이 쉽다.

목차

몬테카를로 방법

- **몬테카를로 방법(Monte Carlo method)**은 난수를 이용하여 함수의 값을 확률적으로 계산하는 알고리즘을 부르는 용어이다.
- 수학이나 물리학 등에 자주 사용되며, 계산하려는 값이 닫힌 형식으로 표현되지 않거나 복잡한 경우에 근사적으로 계산할 때 사용된다.
- 스타니스와프 울람이라는 수학자가 예측하기 어려운 결과는 반복실험을 통해서 예상 확률을 알아낼 수 있다는 생각으로부터 발전시켰으며 도박으로 유명한 모나코의 도시 몬테 카를로의 이름을 본따 명명하였다.
- ‘카드놀이에서 승률을 알아내는 방법 중 가장 좋은 방법은 게임을 여러 번 해 보는 것이고, 게임 횟수가 늘어날수록 추정은 더 정확해진다.’

몬테카를로 방법

- 베이지안 분석에서는 사후분포와 관련된 수치들을 정확하게 계산하는 것이 어렵거나 불가능한 경우가 많다.
- 사후분포에서 모수의 표본을 추출할 수 있다면, 사후분포와 관련된 수치들을 몬테카를로 방법을 이용하여 근사할 수 있다.

몬테카를로 방법

예. 원주율 문제

- 한 변의 길이가 $2r$인 정사각형을 그리고, 정사각형에 내접하는 원을 그리면 원의 반지름은 r이다.

- 원의 넓이는 πr^2이고, 정사각형의 넓이는 $4r^2$이다. 따라서 원의 넓이를 정사각형의 넓이로 나누면 $\frac{\pi}{4}$가 된다.

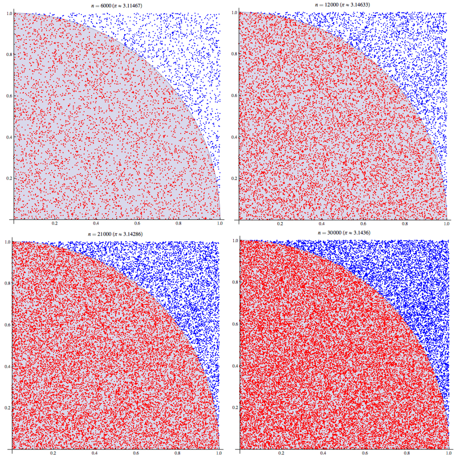
- 정사각형 내부에 무작위로 점을 찍는다. 그리고 점이 원 안에 찍혔는지 여부를 확인한다.

- 점을 많이 찍을수록 전체 점의 수와 원 안에 찍힌 점의 수의 비는 정사각형의 넓이와 원의 넓이의 비에 가까워진다.

몬테카를로 방법

예. 원주율 문제

$$\pi = 4 \times \frac{\text{원의 넓이}}{\text{정사각형의 넓이}} \approx 4 \times \frac{\text{원에 찍힌 점의 수}}{\text{전체 점의 수}}$$

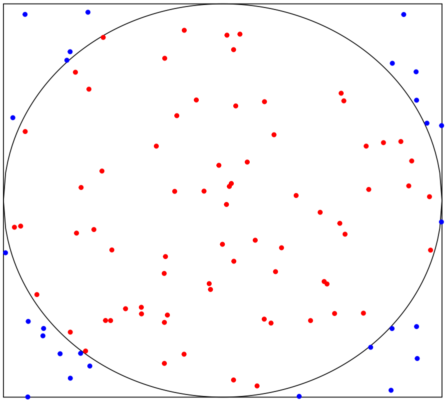


몬테카를로 방법

```r
CalculatePi <- function(n)
{
  count <- 0

  par(mfrow=c(1,1))
  x=seq(-1,1, 0.01)
  y=sqrt(1-x^2)
  plot(x,y,type="l", xlim=c(-1,1), ylim=c(-1,1), axes = F, xlab="", ylab="")
  lines(x,-y,type="l", xlim=c(-1,1), ylim=c(-1,1))
  lines(x=seq(-1,1, 0.01), rep(1, 201))
  lines(x=seq(-1,1, 0.01), rep(-1, 201))
  lines(rep(-1, 201),seq(-1,1, 0.01))
  lines(rep(1, 201),seq(-1,1, 0.01))

  for (i in 1: n) {
    coord <- runif(2, min=-1, max=1)
    if (sqrt(coord[1]^2+coord[2]^2)<=1)
    {
      count <- count+1
      points(coord[1],coord[2], col="red", pch=20)
    }
    else
    {
      points(coord[1],coord[2], col="blue", pch=20)
    }
  }
  return(4*count/n)
}
```

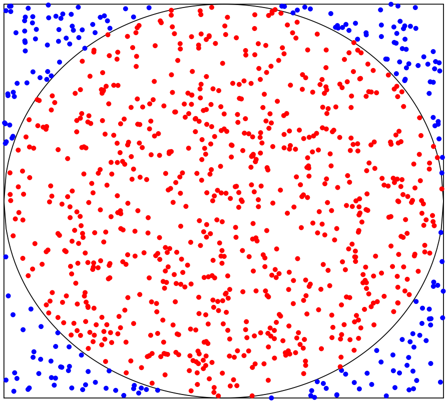
몬테카를로 방법

```r
set.seed(2017)
CalculatePi(100)
```



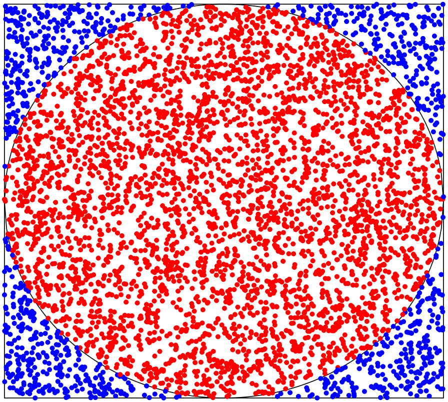
몬테카를로 방법

```
CalculatePi(1000)
```



몬테카를로 방법

```
CalculatePi(5000)
```



몬테카를로 방법

- 사후분포 $P(\theta|X_1, \ldots, X_n)$로부터 표본을 추출할 수 있다고 가정하자.

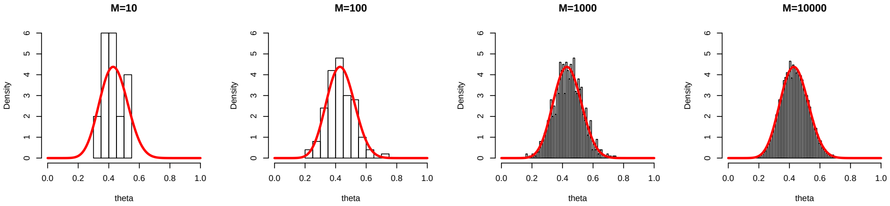
$$\theta^{(1)}, \ldots, \theta^{(M)} \overset{iid}{\sim} P(\theta|X_1, \ldots, X_n)$$

- $\theta^{(1)}, \ldots, \theta^{(M)}$의 경험분포는 표본 크기 M을 늘려감에 따라 $P(\theta|X_1, \ldots, X_n)$로 수렴한다.

- 이항분포-베타사전분포 예제에서 공식을 통해 구했던 사후분포의 평균, 분산, Credible interval을 표본을 이용해 근사할 수 있고, 표본 크기가 커질수록 근사가 정확해진다.

몬테카를로 방법

```r
set.seed(100)
x=seq(from=0,to=1,length.out=1000)
post=dbeta(x,13,17)
theta=rbeta(10000,13,17)
par(mfrow=c(1,4))
hist(theta[1:10],main="M=10",probability=T,xlab="theta",xlim=c(0,1),ylim=c(0,6))
lines(x,post,col=2,lwd=3)
hist(theta[1:100],10,main="M=100",probability=T,xlab="theta",xlim=c(0,1),ylim=c(0,6))
lines(x,post,col=2,lwd=3)
hist(theta[1:1000],50,main="M=1000",probability=T,xlab="theta",xlim=c(0,1),ylim=c(0,6))
lines(x,post,col=2,lwd=3)
hist(theta[1:10000],50,main="M=10000",probability=T,xlab="theta",xlim=c(0,1),ylim=c(0,6))
lines(x,post,col=2,lwd=3)
```



몬테카를로 방법

```r
a=13; b=17
theta=rbeta(100000,a,b)
mean(theta) # mean

## [1] 0.4332692

var(theta) # variance

## [1] 0.007902542

quantile(theta,c(0.025,0.975)) # credible interval

##      2.5%     97.5%
## 0.2649345 0.6105812
```

마코프 체인

- 확률변수들의 수열 $(X^{(t)} : t \in \mathbb{N})$이 아래의 관계를 만족할 때, 이 수열을 **마코프 체인**(Markov chain)이라 부른다.

$$P(X^{(t+1)} = x | X^{(t)} = x_t, \cdots, X^{(1)} = x_1) = P(X^{(t+1)} = x | X^{(t)} = x_t)$$

- 즉, 현재의 상태가 주어졌을 때 미래의 상태와 과거의 상태는 독립이다.
- 예)
  - $X^{(t)} \overset{iid}{\sim} F$
  - AR 모형 : $X^{(t+1)} = \rho X^{(t)} + \epsilon_{t+1}, \quad \epsilon_{t+1} \overset{iid}{\sim} N(0,1)$
  - 랜덤워크 : $X^{(t)} = \sum_{k=1}^{t} Z_k, \quad Z_k \overset{iid}{\sim} F$
- 마코프 체인의 분포를 결정하는 $P(X^{(t+1)} = v | X^{(t)} = u)$를 **전이확률**(transition probability)이라 부른다.
- 전이확률이 t에 의존하지 않는 마코프 체인을 동질 마코프 체인이라 부르고, 전이확률을 다음과 같이 나타낸다.

$$p(v|u) = P(X^{(t+1)} = v | X^{(t)} = u) \;\; \forall t$$

마코프 체인

- $X^{(1)}$이 주어지면 전이확률을 이용하여 $X^{(2)}, X^{(3)}, \ldots$의 확률분포를 계산하는 것이 가능하다.

$$
\begin{aligned}
P(X^{(2)} = x | X^{(1)} = x_1) &= p(x|x_1) \\
P(X^{(3)} = x | X^{(1)} = x_1) &= \sum_{x_2} p(x_2|x_1) p(x|x_2) \\
P(X^{(4)} = x | X^{(1)} = x_1) &= \sum_{x_2} \sum_{x_3} p(x_2|x_1) p(x_3|x_2) p(x|x_3) \\
&\vdots
\end{aligned}
$$

마코프 체인

- 특정한 조건 하에서 전이를 하여도 확률분포가 계속 동일한 평형상태에 도달한다.
- 평형상태에서의 확률분포를 **점근분포**(limiting distribution) 라고 하고, 이 분포는 초기값 $X^{(1)}$에 의존하지 않는다.

$$\pi(x) = \lim_{t \to \infty} P(X^{(t)} = x | X^{(1)} = x_1) \ \ \forall x$$

MCMC

- 몬테카를로 방법은 사후분포에서의 표본추출이 쉽지 않으면 효용도가 떨어지는 문제가 있다.

- 사후분포가 점근분포가 되도록 마코프 체인을 만든다면, 마코프 체인을 통해 사후분포에서 표본을 추출하는 것이 가능하다.

- 이러한 방법을 MCMC(Markov Chain Monte Carlo)라 부르고, 대표적인 방법으로 깁스 샘플러(Gibbs Sampler)와 Metropolis-Hastings 알고리즘(M-H Algorithm)이 있다.

MCMC

- 마코프 체인의 앞쪽 확률변수들의 확률분포는 극한분포에 도달했다고 보기 어렵다.
- 전이를 충분히 반복하여 극한분포에 도달하였다고 하더라도, $X^{(t)}$과 $X^{(t+1)}$는 상당히 연관성이 크다.
- 마코프 체인 $(X^{(1)}, \ldots, X^{(T)})$을 얻었을 때, 앞부분의 값들을 버리는 과정을 burn-in이라고 하고, 연관성을 줄이기 위해 일정 간격씩 띄어서 추출하는 과정을 thinning이라 한다.
- 일반적으로 burn-in과 thinning을 모두 사용하고, 여러 개의 마코프 체인을 만들어 추출하기도 한다.

깁스 샘플러

- p차원 모수 $\theta = (\theta_1, \ldots, \theta_p)$에 대해서 $\pi(\theta)$는 표본 추출이 어려우나, 각각의 조건부분포 $\pi(\theta_k|\theta_{-k})$는 표본 추출이 쉬운 경우 적용할 수 있는 MCMC 방법이다.

- $\theta^{(t)}$이 주어졌을 때, $\theta^{(t+1)}$이 아래의 과정을 통해 추출된다.

$$
\begin{aligned}
\theta_1^{(t+1)} &\sim \pi(\theta_1|\theta_2 = \theta_2^{(t)}, \ldots, \theta_p = \theta_p^{(t)}) \\
\theta_2^{(t+1)} &\sim \pi(\theta_2|\theta_1 = \theta_1^{(t+1)}, \theta_3 = \theta_3^{(t)}, \ldots, \theta_p = \theta_p^{(t)}) \\
\theta_3^{(t+1)} &\sim \pi(\theta_3|\theta_1 = \theta_1^{(t+1)}, \theta_2 = \theta_2^{(t+1)}, \theta_4 = \theta_4^{(t)}, \ldots, \theta_p = \theta_p^{(t)}) \\
&\vdots \\
\theta_p^{(t+1)} &\sim \pi(\theta_p|\theta_1 = \theta_1^{(t+1)}, \ldots, \theta_{p-1} = \theta_{p-1}^{(t+1)})
\end{aligned}
$$

정규분포-정규감마사전분포

- 확률모형

$$X_1, \ldots, X_n | \mu, \lambda \overset{iid}{\sim} Normal(\mu, 1/\lambda)$$

- 사전분포

$$\begin{aligned} \lambda &\sim Gamma(\alpha, \beta) \\ \mu|\lambda &\sim Normal\left(\mu_0, (\kappa_0 \lambda)^{-1}\right) \end{aligned}$$

- 사후분포

$$\begin{aligned} & P(\mu, \lambda | X_1, \ldots, X_n) \\ \propto\ & P(\mu|\lambda; \mu_0, \kappa_0) P(\lambda; \alpha, \beta) \prod_{i=1}^{n} P(X_i | \mu, \lambda) \\ \propto\ & \lambda^{\frac{1}{2}} \exp\left\{-\frac{\kappa_0 \lambda (\mu - \mu_0)^2}{2}\right\} \lambda^{\alpha-1} e^{-\beta\lambda} \lambda^{\frac{n}{2}} \exp\left\{-\frac{\lambda \sum_{i=1}^{n} (X_i - \mu)^2}{2}\right\} \\ =\ & \lambda^{\alpha + \frac{n-1}{2}} \exp\left[-\left\{\beta + \frac{\kappa_0 (\mu - \mu_0)^2 + \sum_{i=1}^{n} (X_i - \mu)^2}{2}\right\} \lambda\right] \end{aligned}$$

정규분포-정규감마사전분포

$$P(\lambda|\mu, X_1, \ldots, X_n) \propto \lambda^{\alpha+\frac{n-1}{2}} \exp\left[-\left\{\beta + \frac{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^n (X_i-\mu)^2}{2}\right\}\lambda\right]$$

$$P(\mu|\lambda, X_1, \ldots, X_n) \propto \exp\left[-\frac{\lambda}{2}\left\{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^n (X_i-\mu)^2\right\}\right]$$

$$\propto \exp\left\{-\frac{\lambda(n+\kappa_0)}{2}\left(\mu - \frac{\kappa_0\mu_0 + \sum_{i=1}^n X_i}{n+\kappa_0}\right)^2\right\}$$

- 사후분포의 조건부분포들에서 쉽게 표본을 추출할 수 있으므로 깁스 샘플러를 사용할 수 있다.

$$\lambda|\mu, X_1, \ldots, X_n \sim Gamma\left(\alpha + \frac{n+1}{2}, \beta + \frac{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^n (X_i-\mu)^2}{2}\right)$$

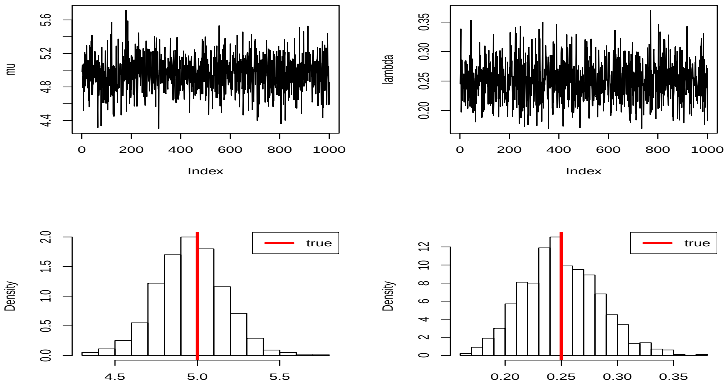
$$\mu|\lambda, X_1, \ldots, X_n \sim Normal\left(\frac{\kappa_0\mu_0 + \sum_{i=1}^n X_i}{n+\kappa_0}, \frac{1}{\lambda(n+\kappa_0)}\right)$$

정규분포-정규감마사전분포

```
set.seed(100)
X=rnorm(100,5,2) # Normal(5,4)
Gibbs=function(X, alpha, beta, m0, k0, iter, burnin, thinning){
  mus=numeric(iter)
  lambdas=numeric(iter)
  Xsum=sum(X)
  n=length(X)
  mu=1
  for(i in 1:iter){
    lambdas[i]=lambda=rgamma(1,alpha+(n+1)/2,
                beta+(k0*(mu-m0)^2+sum((X-mu)^2))/2)
    mus[i]=mu=rnorm(1,(k0*m0+Xsum)/(n+k0),1/sqrt(lambda*(n+k0)))
  }
  mus=mus[-(1:burnin)] # burn-in
  mus=mus[1:((iter-burnin)/thinning)*thinning] # thinning
  lambdas=lambdas[-(1:burnin)]
  lambdas=lambdas[1:((iter-burnin)/thinning)*thinning]
  list(mu=mus,lambda=lambdas)
}
res=Gibbs(X,5,1,0,1,20000,10000,10)
mu=res$mu
lambda=res$lambda
```

정규분포-정규감마사전분포

```r
par(mfrow=c(2,2))
plot(mu,type="l",main="")
plot(lambda,type="l",main="")
hist(mu,20,main="",probability=T,xlab="mu")
abline(v=5,col=2,lwd=3)
legend("topright", c("true"), lty=1, lwd=3, col=2)
hist(lambda,20,main="",probability=T,xlab="lambda")
abline(v=1/4,col=2,lwd=3)
legend("topright", c("true"), lty=1, lwd=3, col=2)
```



정규분포-코시사전분포

- 확률모형

$$X_1, \ldots, X_n | \theta \overset{iid}{\sim} Normal(\theta, 1)$$

- 사전분포

$$P(\theta) = \frac{1}{\pi(1 + \theta^2)} \quad \text{for } \theta \in \mathbb{R}$$

- 사후분포

$$\begin{aligned} P(\theta | X_1, \ldots, X_n) &\propto P(\theta) \prod_{i=1}^{n} P(X_i | \theta) \\ &\propto \frac{1}{1 + \theta^2} \exp\left\{ -\frac{\sum_{i=1}^{n} (X_i - \theta)^2}{2} \right\} \\ &\propto \frac{1}{1 + \theta^2} \exp\left\{ -\frac{n(\theta - \bar{X})^2}{2} \right\} \quad \text{where } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \end{aligned}$$

- 알려진 형태의 분포가 아니기 때문에 직접 표본을 추출하기가 어렵다.

Metropolis-Hastings 알고리즘

- 표본을 추출하고자 하는 분포를 $\pi(\theta)$라 하자.
- 베이지안 분석에서는 $\pi(\theta) = P(\theta|X_1, \ldots, X_n)$이 된다.
- M-H 알고리즘에서는 **제안분포**(proposal distribution) $q(\cdot|\cdot)$을 정해주어야 한다.
- 제안분포는 일종의 전이확률로 볼 수 있지만, 임의로 선택했으므로 이 마코프체인의 극한분포는 $\pi(\theta)$가 아니다.
- M-H 알고리즘은 극한분포가 $\pi(\theta)$가 되도록 제안분포를 개량하여 전이확률을 만드는 방법이다.

Metropolis-Hastings 알고리즘

- $\theta^{(t)}$가 주어졌을 때, $\theta^{(t+1)}$이 아래의 과정을 통해 추출된다.
- 이를 반복하여 $\theta^{(1)}, \ldots, \theta^{(T)}$를 얻을 수 있고, burn-in과 thinning을 통해 MCMC 표본을 선택한다.

1. 제안분포에서 $\theta^{(t+1)}$의 후보 $\tilde{\theta}$를 추출

$$\tilde{\theta} \sim q(\cdot|\theta^{(t)})$$

2. 수락확률 $\alpha(\theta^{(t)}, \tilde{\theta})$ 계산

$$\alpha(\theta^{(t)}, \tilde{\theta}) = \min\left\{ \frac{\pi(\tilde{\theta})q(\theta^{(t)}|\tilde{\theta})}{\pi(\theta^{(t)})q(\tilde{\theta}|\theta^{(t)})}, 1 \right\}$$

3. θ_{n+1} 결정

$$\theta^{(t+1)} = \begin{cases} \tilde{\theta} & \text{w.p. } \alpha(\theta^{(t)}, \tilde{\theta}) \\ \theta^{(t)} & \text{o.w.} \end{cases}$$

Metropolis-Hastings 알고리즘

- M-H 알고리즘의 전이확률은 $p(v|u) = \alpha(u, v)q(v|u)$이다.
- 알고리즘에서 $\frac{\pi(\tilde{\theta})}{\pi(\theta^{(t)})}$만 필요로하기 때문에 $\pi(\theta)$에 비례하는 표준화되지 않은 분포만 알아도 충분하다. (베이즈 정리 사용 용이)
- θ가 고차원일 때 α가 너무 작아서 $\pi(\theta)$로 수렴하기까지 시간이 매우 오래걸릴 수 있다.

정규분포-코시사전분포

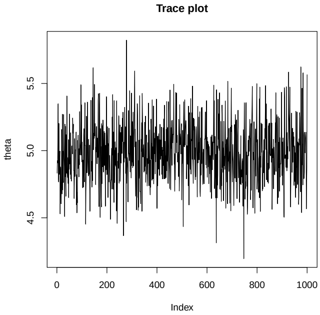
- 사후분포 : $P(\theta|X_1, \ldots, X_n) \propto \frac{1}{1+\theta^2} \exp\left\{-\frac{n(\theta-\bar{X})^2}{2}\right\} =: \pi(\theta)$
- 제안분포 : $q(\cdot|u) = Normal(u, 1)$
- 수락확률 : $\alpha(\theta^{(t)}, \tilde{\theta}) = \min\left\{\frac{\pi(\tilde{\theta})q(\theta^{(t)}|\tilde{\theta})}{\pi(\theta^{(t)})q(\tilde{\theta}|\theta^{(t)})}, 1\right\}$

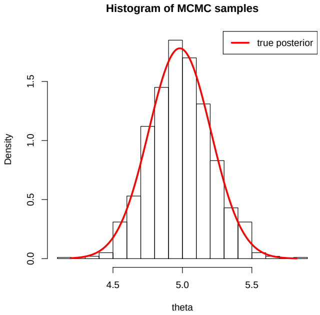
정규분포-코시사전분포

```
n=20; Xbar=5 # observations
pi=function(theta,n,Xbar){ # log posterior
  -n*(theta-Xbar)*(theta-Xbar)/2-log(1+theta*theta)
}
q=function(theta,u){ # log proposal
  dnorm(theta,u,log=T)
}
MH=function(theta1, n, Xbar, iter, burnin, thinning){
  tilde=rnorm(iter-1)
  theta=numeric(iter)
  theta[1]=theta1
  u=runif(iter-1)
  for(i in 1:(iter-1)){
    tilde[i]=tilde[i]+theta[i] # candidate
    alpha=exp(pi(tilde[i],n,Xbar)+q(theta[i],tilde[i])
              -pi(theta[i],n,Xbar)-q(tilde[i],theta[i]))
    if(u[i]<alpha) theta[i+1]=tilde[i]
    else theta[i+1]=theta[i]
  }
  theta=theta[-(1:burnin)] # burn-in
  theta=theta[1:((iter-burnin)/thinning)*thinning] # thinning
}
set.seed(100)
theta=MH(1,n,Xbar,20000,10000,10) # run M-H algorithm
```

정규분포-코시사전분포

```r
par(mfrow=c(1,2))
plot(theta,type="l",main="Trace plot")
C=integrate(function(x) exp(pi(x,n,Xbar)),-Inf,Inf)$value
x=seq(from=min(theta),to=max(theta),length.out=1000)
post=exp(pi(x,n,Xbar))/C
hist(theta,15,main="Histogram of MCMC samples",probability=T,xlab="theta")
lines(x,post,col=2,lwd=3)
legend("topright", c("true posterior"), lty=1, lwd=3, col=2)
```





베이지안 회귀분석

- 확률모형

$$Y_i|X_i, \beta, \tau \sim Normal\left(X_i^T\beta, \frac{1}{\tau}\right) \quad \text{for } i = 1, \ldots, n$$

- 사전분포

$$\beta \sim MVN(\beta_0, \Sigma_0), \quad \tau \sim Gamma(u, v)$$

- 사후분포

$$\begin{aligned}
&P(\beta, \tau|\mathbf{X}, \mathbf{Y}) \\
\propto\ & P(\beta; \beta_0, \Sigma_0)P(\tau; u, v)\prod_{i=1}^{n} P(Y_i|X_i, \beta, \tau) \\
\propto\ & \exp\left\{-\frac{1}{2}\left(\beta^T\Sigma_0^{-1}\beta - 2\beta^T\Sigma_0\beta_0\right)\right\}\tau^{u-1}\exp(-v\tau) \\
&\times\tau^{\frac{n}{2}}\exp\left\{-\frac{\tau}{2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\right\} \\
&\text{where } \mathbf{X} = (X_1, \ldots, X_n)^T,\ \mathbf{Y} = (Y_1, \ldots, Y_n)^T
\end{aligned}$$

베이지안 회귀분석

- 깁스 샘플러
  1. β

$$
\begin{aligned}
& P(\beta|\tau, \mathbf{X}, \mathbf{Y}) \\
\propto\ & \exp\left\{-\frac{1}{2}\left(\beta^T\Sigma_0^{-1}\beta - 2\beta^T\Sigma_0\beta_0\right) - \frac{\tau}{2}\left(\beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{Y}\right)\right\} \\
=\ & \exp\left\{-\frac{1}{2}\beta^T\left(\Sigma_0^{-1} + \tau\mathbf{X}^T\mathbf{X}\right)\beta + \beta^T\left(\Sigma_0^{-1}\beta_0 + \tau\mathbf{X}^T\mathbf{Y}\right)\right\} \\
& \beta|\tau, \mathbf{X}, \mathbf{Y} \sim MVN\left(\hat{\beta}, \hat{\Sigma}\right) \\
& \text{where } \ \hat{\Sigma} = \left(\Sigma_0^{-1} + \tau\mathbf{X}^T\mathbf{X}\right)^{-1},\ \hat{\beta} = \hat{\Sigma}\left(\Sigma_0^{-1}\beta_0 + \tau\mathbf{X}^T\mathbf{Y}\right)
\end{aligned}
$$

  2. τ

$$
P(\tau|\beta, \mathbf{X}, \mathbf{Y}) \propto \tau^{u+\frac{n}{2}-1}\exp\left[-\left\{v + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right\}\tau\right]
$$

$$
\tau|\beta, \mathbf{X}, \mathbf{Y} \sim Gamma\left(u + \frac{n}{2}, v + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right)
$$

베이지안 회귀분석

천식자료(asthma.csv)는 424개 지역의 천식환자 수, 미세먼지 농도, 지가, 아파트 비율, 인구밀도로 이루어진 자료이다. 각 도시에 대하여 천식환자 수를 종속변수, 나머지를 설명변수로 하는 베이지안 회귀모형을 만들어보자.

- asthma : 10,000명 당 천식환자 수 (명/10,000)
- pm10 : 미세먼지 농도 ($\mu g/m^3$)
- hoval : 평균지가 (천원/m^2)
- apt_rate : 아파트 비율 (퍼센트)
- pop_den : 인구밀도 (명/km^2)

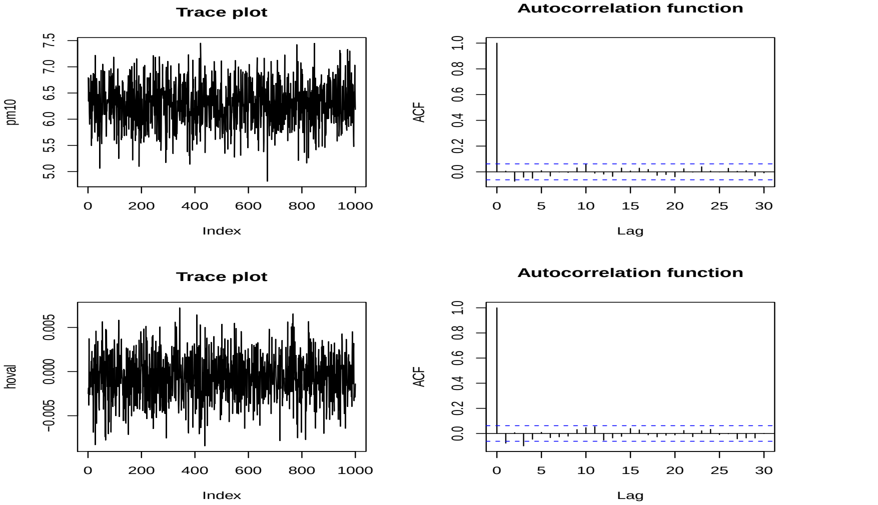
(SungChul Seo. et al., "GIS-based Association Between PM10 and Allergic Diseases in Seoul: Implications for Health and Environmental Policy")

베이지안 회귀분석

```r
data <- read.csv("C:/Users/JKH/Downloads/asthma.csv")
X <- as.matrix(data[,3:6])
Y <- as.matrix(data[,2])
beta0=rep(0,ncol(X)); Sigma0inv=diag(rep(1,ncol(X))); u=1; v=1  # prior

Gibbs_REG=function(beta0, Sigma0inv, u, v, X, Y, iter, burnin, thinning){
  k=ncol(X)
  n=nrow(X)
  XtX=t(X)%*%X

  betas=matrix(0,nrow=iter,ncol=k)
  taus=numeric(iter)
  beta=coefficients(lm(Y~X-1)) # initial values

  for(i in 1:iter){
    taus[i]=tau=rgamma(1,u+n/2,v+sum((Y-X%*%beta)*(Y-X%*%beta))/2)
    Sigma=solve(Sigma0inv+tau*XtX)
    betas[i,]=beta=Sigma%*%(Sigma0inv%*%beta0+tau*t(X)%*%Y)+
                                        matrix(rnorm(k)%*%chol(Sigma),ncol=1)
  }
  betas=betas[-(1:burnin),] # burn-in
  betas=betas[1:((iter-burnin)/thinning)*thinning,] # thinning
  taus=taus[-(1:burnin)]
  taus=taus[1:((iter-burnin)/thinning)*thinning]
  list(beta=betas,tau=taus)
}
res=Gibbs_REG(beta0,Sigma0inv,u,v,X,Y,20000,10000,10)
beta=res$beta
tau=res$tau
```

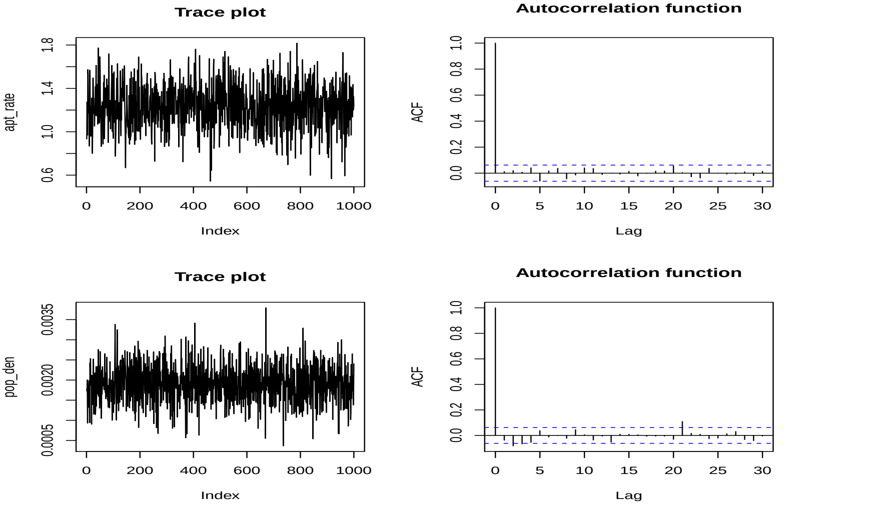
베이지안 회귀분석

- β_{pm10}, β_{hoval}



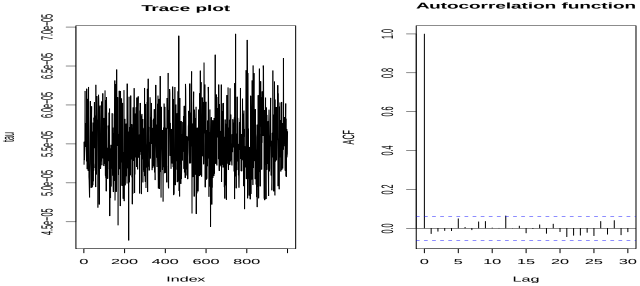
베이지안 회귀분석

- $\beta_{\text{apt\_rate}}$, $\beta_{\text{pop\_den}}$



베이지안 회귀분석

- τ

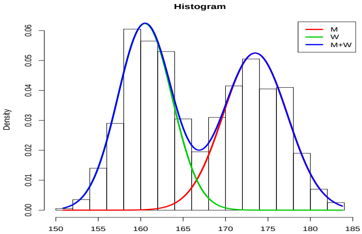


목차

혼합모형

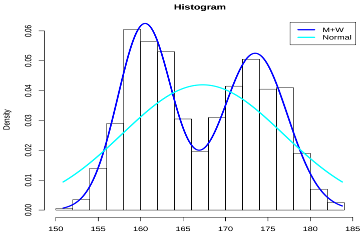
- 고등학교 3학년 학생들의 키의 확률분포를 모형화하고자 한다.
- 남학생들의 키를 정규분포로 모형화하여 모수 $(\mu_M,\ \sigma_M^2)$을 얻었고, 여학생들의 키를 정규분포로 모형화하여 모수 $(\mu_W,\ \sigma_W^2)$를 얻었다고 가정하자.
- 남학생과 여학생의 성비를 $\pi_M : \pi_W$ $(\pi_M + \pi_W = 1)$라 하면, 고등학교 3학년 학생들의 키의 확률분포를 아래와 같이 표현할 수 있다.

$$\pi_M Normal(\mu_M, \sigma_M^2) + \pi_W Normal(\mu_W, \sigma_W^2)$$



혼합모형

- 성별 정보를 무시한 채, 전체 자료를 정규분포로 모형화하면 자료를 잘 설명하지 못한다.
- 따라서 자료에 성별 정보가 없는 경우에 큰 문제가 발생한다.



혼합모형

- 이에 대한 해결책으로 성별을 나타내는 잠재변수 $(Z = (Z_1, \ldots, Z_n))$를 모형에 포함시키는 방법이 있다.

- 확률모형

$$Z_i = \begin{cases} M & \text{w.p. } \pi_M \\ W & \text{o.w.} \end{cases}$$

$$X_i | Z_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim Normal(\mu_{Z_i}, \sigma^2_{Z_i})$$

$$\text{where} \quad \boldsymbol{\mu} = (\mu_M, \mu_W), \ \boldsymbol{\sigma}^2 = (\sigma^2_M, \sigma^2_W)$$

- $\boldsymbol{\pi} \, (= (\pi_M, \pi_W))$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$의 추정을 통해 전체 키의 확률분포

$$\pi_M Normal(\mu_M, \sigma^2_M) + \pi_W Normal(\mu_W, \sigma^2_W)$$

의 추정이 가능하다.

혼합모형

- 확률분포가 $K(\in \mathbb{N})$개의 기본분포(*e.g.* 정규분포)들의 합으로 나타나는 것을 혼합모형이라 한다.

$$f(x;\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{\theta_k}(x)$$

- 혼합모형을 통해 모형기반 군집분석(model-based clustering)이 가능하다.

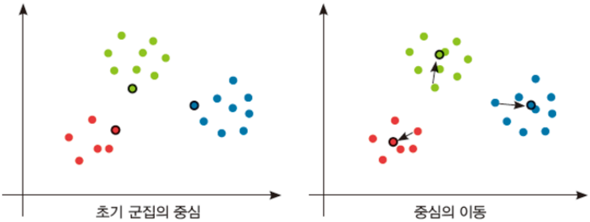
- 본 강의에서는 그룹의 수 K가 알려졌다고 가정한다.

K-평균 군집분석

- K-평균 군집분석
  - 사전에 결정된 군집수 K가 주어지면 유클리드 거리를 이용하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 나누는 방법

K-평균 군집분석

- K-평균 군집분석의 알고리즘
  1. 군집수 *K*가 주어지면 랜덤하게 초기 *K*개 군집의 중심 선택
  2. 각 관측값을 그 중심과 가장 가까운 거리에 있는 군집에 할당
  3. 군집 중심을 새로 계산
  4. 기존의 중심과 새로 계산된 중심 간에 차이 없을 때까지 2-3번 반복

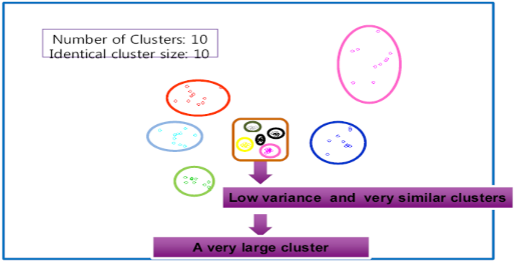


K-평균 군집분석

- K의 선택
  - 실제 데이터분석에서 적절한 군집 수를 미리 아는 경우는 드묾
  - 군집수의 변경해 가면서 결과를 보고 최적의 군집 수를 결정
    - 일반적으로 K값이 증가함에 따라 거리 제곱 합이 감소
    - K값을 증가시켜 가면서 거리 제곱 합을 계산하고 $K-1$때의 값보다 감소량이 줄어드는 K값을 선택

K-평균 군집분석

- K-평균 군집분석의 문제점
  - 군집간의 분산이 다른 경우 안좋음
  - 자료들간의 상관관계를 반영하지 못함
  - 이종의 변수가 섞여 있는 경우 거리를 정의하기 어려움



베이지안 정규분포 혼합모형

- 기본분포가 정규분포이고, 모수 $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$를 베이지안 방법으로 추정한다.

- 먼저 사전분포로 사용될 두 분포를 소개한다.
  - 디리클레분포

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

$$\text{where } x_k \in (0, 1), \sum_{k=1}^{K} = 1$$

  - 역감마분포

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}} \quad \text{for } x \in \mathbb{R}^+$$

$\cdot$ $X \sim Gamma(\alpha, \beta) \;\Leftrightarrow\; \frac{1}{X} \sim Inv - Gamma(\alpha, \beta)$

베이지안 정규분포 혼합모형

- 확률모형

$$\begin{aligned} P(Z_i = k|\boldsymbol{\pi}) &= \pi_k \quad \text{for } k = 1, \ldots, K \\ X_i|Z_i = k, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\sim Normal_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{for } i = 1, \ldots, n \end{aligned}$$

- 편의상 $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}_d$로 가정

- 사전분포

$$\begin{aligned} \boldsymbol{\pi} &\sim Dirichlet(\alpha, \ldots, \alpha) \\ \sigma_k^2 &\sim Inv - Gamma(u/2, v/2) \\ \boldsymbol{\mu}_k|\sigma_k^2 &\sim Normal_d(\boldsymbol{\mu}_0, \lambda\sigma_k^2 \mathbf{I}_d) \end{aligned}$$

- 사후분포

$$\begin{aligned} P(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{Z}|\mathbf{X}) &\propto P(\boldsymbol{\pi}; \alpha) \prod_{k=1}^{K} \left\{ P(\boldsymbol{\mu}_k|\sigma_k^2; \boldsymbol{\mu}_0, \lambda) P(\sigma_k^2; u, v) \right\} \\ &\quad \times \prod_{i=1}^{n} P(Z_i|\boldsymbol{\pi}) P(X_i|Z_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\ \text{where} \quad & \mathbf{Z} = (Z_1, \ldots, Z_n) \text{ and } \mathbf{X} = (X_1, \ldots, X_n) \end{aligned}$$

베이지안 정규분포 혼합모형

- 깁스 샘플러

1. $\boldsymbol{\pi}$

$$
\begin{aligned}
P(\boldsymbol{\pi}|\mathbf{Z};\alpha) &\propto P(\boldsymbol{\pi};\alpha)\prod_{i=1}^{n} P(Z_i|\boldsymbol{\pi}) \\
&\propto \prod_{k=1}^{K}\pi_k^{\alpha-1}\prod_{i=1}^{n}\pi_k^{I(Z_i=k)} = \prod_{k=1}^{K}\pi_k^{\alpha+n_k-1} \\
\boldsymbol{\pi}|\mathbf{Z} &\sim \mathit{Dirichlet}(\alpha+n_1,\ldots,\alpha+n_K)
\end{aligned}
$$

여기서 $I(\cdot)$은 표시함수이고, $n_k = \sum_{i=1}^{n} I(Z_i = k)$이다.

베이지안 정규분포 혼합모형

2. $\boldsymbol{\mu}$

$$
\begin{aligned}
& P(\boldsymbol{\mu}_k|\sigma_k^2, \mathbf{Z}, \mathbf{X}; \boldsymbol{\mu}_0, \lambda) \\
\propto\ & P(\boldsymbol{\mu}_k|\sigma_k^2; \boldsymbol{\mu}_0, \lambda) \prod_{Z_i=k} P(X_i|\boldsymbol{\mu}_k, \sigma_k^2) \\
\propto\ & \exp\left(-\frac{1}{2\lambda\sigma_k^2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) - \frac{1}{2\sigma_k^2}\sum_{Z_i=k}(X_i - \boldsymbol{\mu}_k)^T(X_i - \boldsymbol{\mu}_k)\right) \\
& \boldsymbol{\mu}_k|\sigma_k^2, \mathbf{Z}, \mathbf{X} \sim Normal_d(\hat{\boldsymbol{\mu}}_k, \hat{\sigma}_k^2 \mathbf{I}_d) \\
& \quad \text{where } \hat{\sigma}_k^2 = \frac{\lambda\sigma_k^2}{n_k\lambda + 1} \text{ and } \hat{\boldsymbol{\mu}}_k = \frac{1}{n_k\lambda + 1}\boldsymbol{\mu}_0 + \frac{n_k\lambda}{n_k\lambda + 1}\left(\frac{1}{n_k}\sum_{Z_i=k} X_i\right)
\end{aligned}
$$

베이지안 정규분포 혼합모형

3. $\boldsymbol{\sigma}^2$

$$
\begin{aligned}
& P(\sigma_k^2|\boldsymbol{\mu}_k, \mathbf{Z}, \mathbf{X}; \boldsymbol{\mu}_0, \lambda, u, v) \\
\propto\ & P(\boldsymbol{\mu}_k|\sigma_k^2; \boldsymbol{\mu}_0, \lambda) P(\sigma_k^2; u, v) \prod_{Z_i=k} P(X_i|\boldsymbol{\mu}_k, \sigma_k^2) \\
\propto\ & \left(\sigma_k^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\lambda\sigma_k^2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)\right\} \left(\sigma_k^2\right)^{-\frac{u}{2}-1} \exp\left(-\frac{v}{2\sigma_k^2}\right) \\
& \times \prod_{Z_i=k}\left[\left(\sigma_k^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_k^2}(X_i - \boldsymbol{\mu}_k)^T(X_i - \boldsymbol{\mu}_k)\right\}\right] \\
=\ & (\sigma_k^2)^{-\hat{u}-1} \exp\left(-\frac{\hat{v}}{\sigma_k^2}\right)
\end{aligned}
$$

$$
\sigma_k^2|\boldsymbol{\mu}_k, \mathbf{Z}, \mathbf{X} \sim Inv - Gamma(\hat{u}, \hat{v}) \quad \text{where } \hat{u} = \frac{u + n_k + 1}{2} \text{ and}
$$

$$
\hat{v} = \frac{1}{2\lambda}\left\{\lambda v + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0) + \lambda \sum_{Z_i=k}(X_i - \boldsymbol{\mu}_k)^T(X_i - \boldsymbol{\mu}_k)\right\}
$$

베이지안 정규분포 혼합모형

4. **Z**

$$
\begin{aligned}
P(Z_i = k|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, X_i) &\propto P(Z_i = k|\boldsymbol{\pi})P(X_i|\boldsymbol{\mu}_k, \sigma_k^2) \\
&\propto \frac{\pi_k}{(\sigma_k^2)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i - \boldsymbol{\mu}_k)^T(X_i - \boldsymbol{\mu}_k)}{2\sigma_k^2}\right\}
\end{aligned}
$$

베이지안 정규분포 혼합모형

- 고등학교 3학년 학생들의 키 자료를 깁스 샘플러 방법으로 분석한다.

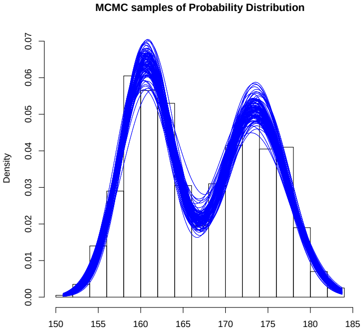
```
set.seed(10)
M=rnorm(500,173.5,3.8); W=rnorm(500,160.5,3.2);  # true
X=c(M,W)
alpha=1; mu0=mean(X); lambda=1; u=5; v=5;

Gibbs_MIX=function(alpha, mu0, lambda, u, v, X, iter, burnin, thinning){
  n=length(X); idx=list()
  pis=mus=sigmas=matrix(0,nrow=iter,ncol=2)
  pi=rep(0.5,2)
  mu=c(mean(X[X<median(X)]),mean(X[X>=median(X)]))
  sigma=c(var(X[X<median(X)]),var(X[X>=median(X)]))

  for(i in 1:iter){
    # Z
    u=runif(n)
    Z=rep(2,n)
    for(j in 1:n){
      p=log(pi)-log(sigma)/2-(X[j]-mu)^2/(2*sigma)
      p=exp(p-max(p))
      if(u[j]<p[1]/sum(p)) Z[j]=1
    }
    idx[[1]]=which(Z==1); idx[[2]]=which(Z==2)
    n_k=c(length(idx[[1]]),length(idx[[2]]))

    # pi
    pis[i,1]=pi[1]=rbeta(1,alpha+n_k[1],alpha+n_k[2])
    pis[i,2]=pi[2]=1-pi[1]
```

베이지안 정규분포 혼합모형

```
    # mu, sigma
    q=1/(n_k*lambda+1)
    m0=c(mean(X[idx[[1]]]),mean(X[idx[[2]]]))
    m=q*mu0+(1-q)*m0
    s=lambda*sigma/(n_k*lambda+1)
    mus[i,1]=mu[1]=rnorm(1,m[1],sqrt(s[1]))
    mus[i,2]=mu[2]=rnorm(1,m[2],sqrt(s[2]))
    sigmas[i,1]=sigma[1]=1/rgamma(1,(u+n_k[1]+1)/2,
            (lambda*v+(mu[1]-mu0)^2+lambda*sum((X[idx[[1]]]-mu[1])^2))/2)
    sigmas[i,2]=sigma[2]=1/rgamma(1,(u+n_k[2]+1)/2,
            (lambda*v+(mu[2]-mu0)^2+lambda*sum((X[idx[[2]]]-mu[2])^2))/2)
  }
  pis=pis[-(1:burnin),] # burn-in
  pis=pis[1:((iter-burnin)/thinning)*thinning,] # thinning
  mus=mus[-(1:burnin),]
  mus=mus[1:((iter-burnin)/thinning)*thinning,]
  sigmas=sigmas[-(1:burnin),]
  sigmas=sigmas[1:((iter-burnin)/thinning)*thinning,]
  list(pi=pis,mu=mus,sigma=sigmas)
}
res=Gibbs_MIX(alpha,mu0,lambda,u,v,X,10000,5000,10)
pi=res$pi
mu=res$mu
sigma=res$sigma
```

베이지안 정규분포 혼합모형

```
xx=seq(from=min(X),to=max(X),length.out=1000)
l=length(mu[,1])/5
y=matrix(0,nrow=l,ncol=1000)
for(i in 1:l)
  y[i,]=pi[i*5,1]*dnorm(xx,mu[i*5,1],sqrt(sigma[i*5,1]))+
    pi[i*5,2]*dnorm(xx,mu[i*5,2],sqrt(sigma[i*5,2]))
hist(X,20,probability=T,main="MCMC samples of Probability Distribution"
     ,xlab="",ylim=c(0,max(y)))
for(i in 1:l)
  lines(xx,y[i,],lty=3,lwd=0.1,col=4)
```

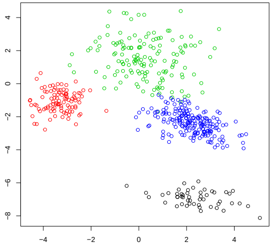


k-평균 vs 혼합모형

- 4개의 군집을 갖는 2차원 상의 자료를 임의로 생성

```
library(MASS);library(mclust);set.seed(50)
```

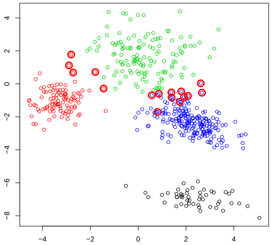
```
mu=matrix(rnorm(8,sd=4),ncol=2)
Sigma=array(0,dim=c(4,2,2))
x=c()
for(i in 1:4){
  diag(Sigma[i,,])=rgamma(2,3,4)
  Sigma[i,1,2]=Sigma[i,2,1]= runif(1,-0.5,0.5)*sqrt(Sigma[i,1,1]*Sigma[i,2,2])
  x=rbind(x,mvrnorm(50*i,mu[i,],Sigma[i,,]))
}
clust0=c(rep(1,50),rep(2,100),rep(3,150),rep(4,200))
plot(x,col=clust0,xlab="",ylab="")
```



k-평균 vs 혼합모형

- k-평균

```
k_means=kmeans(x,4)
clust1=c(1,3,4,2)[k_means$cluster]
plot(x,col=clust1,xlab="",ylab="")
points(x[which(clust0!=clust1),],col="red",cex=2,lwd=3)
```

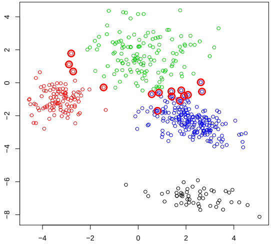


- 빨간색 큰 원은 오분류를 나타냄

k-평균 vs 혼합모형

- 혼합모형 : $\Sigma_k = \sigma^2 \mathbf{I}$ (Mclust의 "EII" 옵션)

```
mix_EII=Mclust(x,4,"EII")
clust2=apply(mix_EII$z,1,which.max)
plot(x,col=clust2,xlab="",ylab="")
points(x[which(clust0!=clust2),],col="red",cex=2,lwd=3)
```

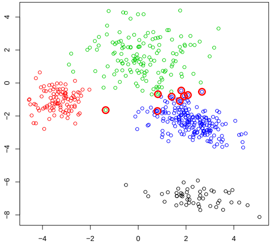


- k-평균과 유사

k-평균 vs 혼합모형

- 혼합모형 : $\Sigma_k = \sigma_k^2 \mathbf{I}$ (Mclust의 "VII" 옵션)

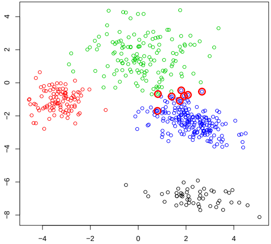
```
mix_VII=Mclust(x,4,"VII")
clust3=apply(mix_VII$z,1,which.max)
plot(x,col=clust3,xlab="",ylab="")
points(x[which(clust0!=clust3),],col="red",cex=2,lwd=3)
```



k-평균 vs 혼합모형

- 혼합모형 : $\Sigma_k = diag(\sigma^2_{k1}, \ldots, \sigma^2_{kd})$ (Mclust의 "VEI" 옵션)

```
mix_VEI=Mclust(x,4,"VEI")
clust4=apply(mix_VEI$z,1,which.max)
plot(x,col=clust4,xlab="",ylab="")
points(x[which(clust0!=clust4),],col="red",cex=2,lwd=3)
```



혼합모형을 이용한 표본 추출

- 두 단계를 통해 혼합모형에서 표본을 추출할 수 있다.

$$x \sim \sum_{k=1}^{K} \pi_k P_{\theta_k} \iff \begin{cases} z \sim (\pi_1, \ldots, \pi_K) \\ x|z = k \sim P_{\theta_k} \end{cases}$$

혼합모형을 이용한 표본 추출

- t-분포

$$
\begin{aligned}
f(x;\nu) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } x \in \mathbb{R}^+ \\
&\propto \Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{x^2+\nu}{2}\right)^{-\frac{\nu+1}{2}} \\
&= \int_0^\infty z^{-\frac{\nu+1}{2}-1}\exp\left(-\frac{x^2+\nu}{2z}\right)dz \\
&\propto \int_0^\infty \frac{1}{\sqrt{z}}\exp\left(-\frac{x^2}{2z}\right) z^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2z}\right)dz
\end{aligned}
$$

$$
x \sim t(\nu) \iff \begin{cases} z \sim Inv - Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\ x|z \sim Normal(0, z) \end{cases}
$$

혼합모형을 이용한 표본 추출

- 음이항분포

$$
\begin{aligned}
f(k; r, p) &= \frac{(k+r-1)!(r-1)!}{k!}(1-p)^r p^k \quad \text{for } k = 0, 1, \ldots \\
&\propto \frac{1}{k!} p^{k+r} \Gamma(k+r) \\
&= \frac{1}{k!} \int_0^\infty z^{k+r-1} e^{-\frac{z}{p}} dz \\
&= \int_0^\infty \frac{z^k}{k!} e^{-z} \cdot z^{r-1} e^{-\frac{1-p}{p} z} dz
\end{aligned}
$$

$$
k \sim Neg - Binomial(r, p) \Longleftrightarrow \begin{cases} z \sim Gamma\left(r, \frac{1-p}{p}\right) \\ k|z \sim Poisson(z) \end{cases}
$$

직접해보기

시리얼 자료(cereals.csv)

- 미국에서 판매되는 77가지 시리얼에 관한 자료
- 자료의 세부 정보는 아래와 같음

1. Name: 시리얼의이름
2. Mfr: 시리얼제조사
   A: American Home Food Products, G: General Mills, K: Kelloggs,
   N: Nabisco, P: Post , Q: Quaker Oats, R: Ralston Purina
3. type: 차갑게 먹는가 따듯하게 먹는가(cold / hot)
4. Calories: 1회 제공량 당 칼로리
5. Protein: 단백질 함량(그램)
6. Fat: 지방 함량(그램)
7. Sodium: 소금 함량(밀리그램)
8. fiber: 식이섬유 함량(그램)
9. carbo: 복합탄수화물 함량(그램)
10. sugars: 설 탕함량(그램)
11. potass: 칼륨 함량(밀리그램)
12. vitamins: FDA 기준치 대비 비타민, 미네랄 함량 %
13. shelf: 진열대 위치(바닥부터 1,2,3층)
14. weight: 1회 제공량 당 무게(온스)
15. cups: 1회 제공량 당 컵 단위(ex. 1.5 컵, 0.9컵 등)

16. rating: 소비자 조사에 의한 시리얼 평점

직접해보기

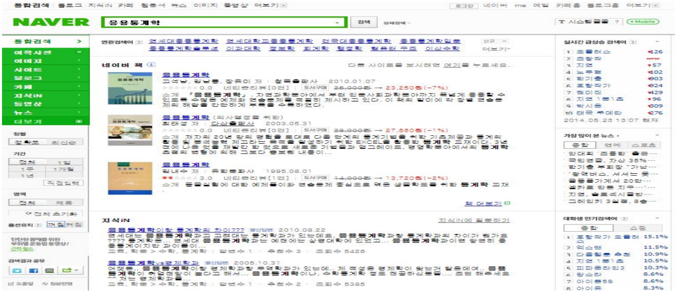
1. 9개의 변수(칼로리, 단백질, 지방, 소금, 식이섬유, 복합 탄수화물, 설탕, 칼륨, 비타민 함량)를 선택
2. 결측치가 있는 자료는 제거
3. k-평균과 혼합모형으로 군집분석 시행

목차

Information Retrieval

Information Retrieval

- Query 와 문서 리스트



TF-IDF

- 주어진 문서를 벡터로 만드는 방법
- 사전에 W개의 단어가 있을 경우, 각 문서를 W 차원의 벡터로 만든다.
- W차원의 벡터의 각 원소는 관련 단어가 문서에서 나온 빈도이다.
- 하지만 단순 빈도를 사용할 경우 자주 나오는 단어 (예: ‘the’, ’a’)의 빈도가 너무 높아진다.
- 이러한 문제를 해결하기 위해서 TF-IDF를 사용한다.
- k번째 문서에서 i번째 단어의 TF-IDF값은

$$\mathrm{TF}-\mathrm{IDF}(k,i)=\mathrm{Freq}(k,i)\times log(N/n_i)$$

  이고 여기서 N은 전체 문서 수, n_i는 i번째 단어를 포함하는 문서의 수이다.
- Query와 문서들의 유사성을 TF-IDF의 유사성 (cos유사성, 상관계수)를 이용하여 측정한다.

Information Retrieval

<u>키워드 빈도를 이용한 IR 예제</u>

- Query:“Applied multivariate analysis"
- Term-frequency Data Base

	regression	histogram	Factor analysis	Multivariate	Asymptotic	clustering	Dimension reduction	Analysis	REL	MATCH
Doc1	x	x	x			X	x		R	
Doc2				X*	x			x*		M
Doc3			x	x*				X*	R	M

숨겨진 의미: Latent Semantic

- 비슷한 단어 문제를 해결하기 위한 방법
- IR 은 단어들의 동시사용빈도(co-occurrent frequency)를 이용한 latent semantic을 추출
- "Applied" 와 "Regression" 또는 "Histogram"의 동시사용빈도는 "Asymptotic"보다 훨씬 크다.

Information Retrieval

Latent semantics 예제

- psychology, history, English, philosophy => Human and social science
- mathematics, statistics, chemistry, physics, biology => Natural science

Information Retrieval

Singular value decomposition for IR (Deerwester *et al.*, 1990)

$$\underset{n \times W}{C} \approx \underset{n \times K}{T} \times \underset{K \times K}{S} \times \underset{K \times W}{D}$$

- K : semantic의 수
- T를 바탕으로 문서들의 유사성을 측정

토픽모형

- 문서 자료로부터 토픽(단어들이 공통으로 출현하는 패턴)들을 찾아내기 위해 개발된 분석 방법
- 단어들의 순서는 모형에 영향을 끼치지 않고, 단어들의 빈도만 모형에 영향을 끼침 (Bag-of-Words)

토픽모형

- 총 n개의 문서가 있고, j번째 문서에는 N_j개의 단어가 들어있다.
- 사전의 크기를 W로 나타내고, 사전을 $\{1, \ldots, W\}$와 매칭시킨다.
- j번째 문서의 i번째 단어를 x_{ji} $(\in \{1, \ldots, W\})$라 하자.
- 토픽은 W개의 단어들에 대한 확률분포이다.
- 총 K개의 토픽이 있고, k번째 토픽 ϕ_k를 아래와 같이 표현한다.

$$\phi_k = (\phi_{k1}, \ldots, \phi_{kW}) \quad \text{where } \phi_k \in (0,1)^W, \ \sum_{w=1}^{W} \phi_{kw} = 1$$

토픽모형

- 토픽모형에서는 j번째 문서에 있는 단어들$(x_{j1}, \ldots, x_{jN_j})$이 j번째 문서의 확률분포 P_j에서 독립적으로 추출되었다고 가정한다.

$$x_{ji} \overset{iid}{\sim} P_j \quad \text{for} \quad i = 1, \ldots, N_j$$

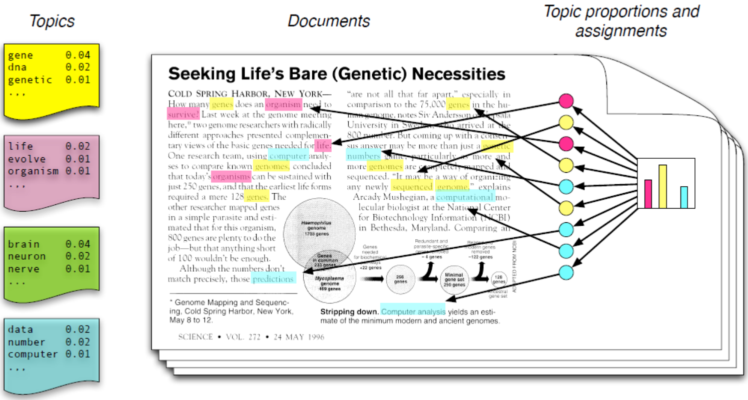
- 토픽모형에서는 문서분포 P_j를 토픽들의 혼합모형으로 가정한다.

$$P_j = \sum_{k=1}^{K} \theta_{jk} \phi_k \quad \text{for} \quad j = 1, \ldots, n$$

여기서 $\theta_j = (\theta_{j1}, \ldots, \theta_{jK})$는 j번째 문서에서 토픽들의 비율을 나타낸다.

- 토픽들은 모든 문서에서 기본분포로 사용되고, 토픽의 비율은 문서마다 다르다.

토픽모형



[Blei, MLSS 2012]

토픽모형

- 단어 x_{ji}를 문서분포 P_j에서 추출하는 것은, 먼저 이 단어에 해당하는 토픽을 z_{ji}를 결정하고, 해당 토픽의 분포에서 x_{ji}를 추출하는 것과 동일하다.

$$x_{ji} \sim \sum_{k=1}^{K} \theta_{jk}\phi_k \Longleftrightarrow \begin{cases} z_{ji} \sim (\theta_{j1}, \ldots, \theta_{jK}) \\ x_{ji}|z_{ji} = k \sim (\phi_{k1}, \ldots, \phi_{kW}) \end{cases}$$

- 표기의 편의를 위해 다음과 같이 정의한다.

$$\begin{aligned} \mathbf{x} &= (x_{ji} : \ j = 1, \ldots, n; \ i = 1, \ldots, N_j) \\ \mathbf{z} &= (z_{ji} : \ j = 1, \ldots, n; \ i = 1, \ldots, N_j) \\ \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_K) \\ \boldsymbol{\theta} &= (\theta_{jk} : \ j = 1, \ldots, n; \ k = 1, \ldots, K) \\ \theta_j &= (\theta_{j1} \ldots, \theta_{jK}) \\ \boldsymbol{\phi} &= (\phi_{kw} : \ k = 1, \ldots, K; \ w = 1, \ldots, W) \\ \phi_k &= (\phi_{k1}, \ldots, \phi_{kW}) \end{aligned}$$

Latent Dirichlet Allocation

- 확률모형

$$
\begin{aligned}
z_{ji}|\theta_j &\sim \theta_j \\
x_{ji}|z_{ji} = k, \phi &\sim \phi_k \quad \text{for } \; j = 1, \ldots, n; \; i = 1, \ldots, N_j
\end{aligned}
$$

- 사전분포

$$
\begin{aligned}
\theta_j &\sim Dirichlet(\alpha, \ldots, \alpha) \quad \text{for } \; j = 1, \ldots, n \\
\phi_k &\sim Dirichlet(\beta, \ldots, \beta) \quad \text{for } \; k = 1, \ldots, K
\end{aligned}
$$

Latent Dirichlet Allocation

- 사후분포

$$
\begin{aligned}
& P(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z} | \mathbf{x}) \\
\propto & \left[\prod_{j=1}^{n} P(\theta_j; \alpha) \right] \left[\prod_{k=1}^{K} P(\phi_k; \beta) \right] \left[\prod_{j=1}^{n} \prod_{i=1}^{N_j} P(x_{ji} | z_{ji}, \phi) P(z_{ji} | \theta_j) \right] \\
\propto & \left[\prod_{j=1}^{n} \prod_{k=1}^{K} \theta_{jk}^{\alpha - 1} \right] \left[\prod_{k=1}^{K} \prod_{w=1}^{W} \phi_{kw}^{\beta - 1} \right] \left[\prod_{j=1}^{n} \prod_{i=1}^{N_j} \prod_{k=1}^{K} \left\{ \theta_{jk}^{I(z_{ji}=k)} \prod_{w=1}^{W} \phi_{kw}^{I(z_{ji}=k, x_{ji}=w)} \right\} \right] \\
\propto & \left[\prod_{j=1}^{n} \prod_{k=1}^{K} \theta_{jk}^{\alpha + N_{jk} - 1} \right] \left[\prod_{k=1}^{K} \prod_{w=1}^{W} \phi_{kw}^{\beta + N_{kw} - 1} \right]
\end{aligned}
$$

$$
\text{where } \; N_{jk} = \sum_{i=1}^{N_j} I(z_{ji} = k), \; N_{kw} = \sum_{j=1}^{n} \sum_{i=1}^{N_j} I(z_{ji} = k, x_{ji} = w)
$$

Latent Dirichlet Allocation

- 깁스 샘플러
  1. $\boldsymbol{\theta}$

$$P(\theta_j|\mathbf{z};\alpha) \propto \prod_{k=1}^{K} \theta_{jk}^{\alpha+N_{jk}-1}$$
$$\theta_j|\mathbf{z} \sim Dirichlet(\alpha+N_{j1},\ldots,\alpha+N_{jK})$$

  2. ϕ

$$P(\phi_k|\mathbf{z};\beta) \propto \prod_{w=1}^{W} \phi_{kw}^{\beta+N_{kw}-1}$$
$$\phi_k|\mathbf{z} \sim Dirichlet(\beta+N_{k1},\ldots,\beta+N_{kW})$$

  3. $\mathbf{z}$

$$P(\mathbf{z}|\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{x}) \propto \prod_{j=1}^{n}\prod_{i=1}^{N_j} P(z_{ji}|\theta_j)P(x_{ji}|\theta_{ji},\boldsymbol{\phi})$$
$$P(z_{ji}=k|\theta_j,\boldsymbol{\phi},x_{ji}) \propto \theta_{jk}\phi_{kx_{ji}}$$

Latent Dirichlet Allocation

- 깁스 샘플러는 추출하는 변수간의 연관성이 클 때 수렴이 느리다.
- 사후분포에서 몇몇 변수들을 적분하여 없앰으로써 변수간의 연관성을 줄여서 수렴 속도를 향상시키는 방법을 붕괴 깁스 샘플러(collapsed Gibbs sampler)라 한다.
- LDA 모형에서는 $\boldsymbol{\theta}$와 $\boldsymbol{\phi}$를 적분하여 없애고 $\mathbf{z}$만 추출하는 깁스 샘플러가 가능하다.
- $\boldsymbol{\theta}$와 $\boldsymbol{\phi}$는 추출된 $\mathbf{z}$와 사후분포의 조건부분포 $P(\boldsymbol{\theta}|\mathbf{z})$, $P(\boldsymbol{\phi}|\mathbf{z})$를 통해 얻을 수 있다.

Latent Dirichlet Allocation

- 붕괴 깁스 샘플러

$$P(\boldsymbol{\theta}, \phi, \mathbf{z}|\mathbf{x}) \propto \left[\prod_{j=1}^{n}\prod_{k=1}^{K}\theta_{jk}^{\alpha+N_{jk}-1}\right]\left[\prod_{k=1}^{K}\prod_{w=1}^{W}\phi_{kw}^{\beta+N_{kw}-1}\right]$$

$$P(\mathbf{z}|\mathbf{x}) \propto \prod_{j=1}^{n}\frac{\prod_{k=1}^{K}\Gamma(\alpha+N_{jk})}{\Gamma\left(K\alpha+\sum_{k=1}^{K}N_{jk}\right)}\prod_{k=1}^{K}\frac{\prod_{w=1}^{W}\Gamma(\beta+N_{kw})}{\Gamma\left(W\beta+\sum_{w=1}^{W}N_{kw}\right)}$$

$$P(z_{ji}=k|\mathbf{z}^{-ji}, x_{ji}) \propto \frac{\beta+N_{kx_{ji}}^{-ji}}{W\beta+\sum_{w=1}^{W}N_{kw}^{-ji}}\left(\alpha+N_{jk}^{-ji}\right)$$

$$\text{where } \mathbf{z}^{-ji} = \left(z_{j'i'} : (j', i') \neq (j, i)\right),\ N_{jk}^{-ji} = \sum_{i' \neq i} I(z_{ji'}=k),$$

$$N_{kw}^{-ji} = \sum_{(j',i')}\sum_{\neq(j,i)} I(z_{j'i'}=k, x_{j'i'}=w)$$

Latent Dirichlet Allocation

- LDA 모형을 쇼핑자료에 적용 가능
  - 고객:문서, 상품:단어
- 인터넷 쇼핑몰의 쇼핑자료를 R 패키지를 이용하여 LDA 모형으로 분석
- 쇼핑몰에서 판매되는 상품들을 70가지 상품으로 중분류(예: 의류, 화장품, 전자제품, 건강제품 등)
- 인터넷 쇼핑몰 회원 중 임의 추출된 10,000명을 대상으로 2015년 한 해 동안 70가지 상품에 대한 구매 횟수를 저장

(a)

A0
A1
A2
A3
A4
A5
A6
A7
A8

(b)

고객	상품	횟수
1	6	12
1	10	2
1	11	6
1	13	4
1	29	2
1	37	5
2	2	1
2	6	6

(a) “품목.txt”의 자료 형태

(b) “쇼핑.csv”의 자료 형태

Latent Dirichlet Allocation

- R의 "lda" 패키지에서 붕괴 깁스 샘플러 알고리즘을 구현
- 이 패키지를 사용하기 위해서는 자료를 “쇼핑.dat”의 형태로 변형해야 함
- “쇼핑.dat”는 각각의 줄이 한 고객에 해당하며 첫 번째 숫자는 해당 고객이 몇 종류의 상품을 구매하였는지를 나타냄
- 이어지는 $a:b$는 상품 a를 b회 구매했음을 의미

```
6 5:12 9:2 10:6 12:4 28:2 36:5
11 1:1 5:6 8:1 10:5 13:5 14:2 16:2 17:9 26:1 36:4 45:1
10 5:10 10:8 12:2 17:1 20:1 31:2 36:8 39:1 45:1 60:1
11 5:5 10:10 12:1 13:18 14:1 17:7 26:1 31:1 34:3 36:5 39:2
```

“쇼핑.dat”의 자료 형태

Latent Dirichlet Allocation

```
library(lda); library(ggplot2); library(reshape2); library(cowplot)
```

```
setwd("D:/Dropbox/R/LDA")
shop=read.csv("쇼핑.csv")

preproc = function(data,name){
  tab=table(data[,1])
  line=paste(data[,2]-1,data[,3],sep=":")
  line=tapply(line,data[,1],FUN=identity)
  line=Map(c, tab, line)
  line=lapply(line,FUN=function(x) paste(x,collapse=" "))
  line=paste(line,collapse="\n")

  loc=paste0(name,".dat")
  outfile=file(loc)
  writeLines(line,outfile)
  close(outfile)
}
preproc(shop,"쇼핑")

shop2=read.documents("쇼핑.dat"); item_name=read.vocab("품목.txt")
head(shop2,1)


## [[1]]
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    5    9   10   12   28   36
## [2,]   12    2    6    4    2    5


set.seed(100)
n=length(shop2); K=11; W=70; alpha=1.0; beta=1.0
lda=lda.collapsed.gibbs.sampler(shop2, K, item_name, 500, alpha, beta)
```

Latent Dirichlet Allocation

- document_sums와 topics는 각각 N_{jk}^{T}와 N_{kw}를 의미

```
# theta를 추정
theta=t(lda$document_sums)
for(i in 1:n) theta[i,]=theta[i,]/sum(theta[i,])
round(theta[1:3,1:10],2)


##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 0.00 0.35    0    0 0.06    0 0.58 0.00 0.00     0
## [2,] 0.19 0.00    0    0 0.00    0 0.00 0.00 0.81     0
## [3,] 0.00 0.31    0    0 0.00    0 0.00 0.03 0.23     0

# phi를 추정
phi=lda$topics
for(i in 1:K)  phi[i,]=phi[i,]/sum(phi[i,])
round(phi[1:3,1:10],2)


##        A0   A1   A2 A3   A4   A5   A6   A7   A8   A9
## [1,] 0.05 0.03 0.01  0 0.00 0.08 0.00 0.00 0.00 0.01
## [2,] 0.05 0.01 0.01  0 0.02 0.22 0.01 0.07 0.02 0.01
## [3,] 0.02 0.00 0.00  0 0.01 0.09 0.00 0.03 0.11 0.01
```

Latent Dirichlet Allocation

- 토픽을 대표하는 주요 상품들로 토픽의 이름을 붙이면 토픽을 효과적으로 나타낼 수 있다.
- 확률이 큰 상품들을 주요 상품으로 본다면 상품간의 구매 빈도 차이가 상당히 클 때 의미가 없어진다.
  - A 상품의 구매 빈도가 다른 상품들의 구매 빈도보다 훨씬 크면 여러 토픽들에서 A 의 확률이 클 것이고, 결국 여러 토픽들의 주요 상품이 같아질 것이다.
- 해당 토픽에서만 특별히 확률이 큰 상품들을 주요 상품으로 보기 위해 상품의 확률을 쇼핑자료 전체에서 해당 상품이 차지하는 비율로 나눈 값을 리프트(lift)로 정의하고, 이 값이 큰 상품들을 주요 상품으로 보았다.

Latent Dirichlet Allocation

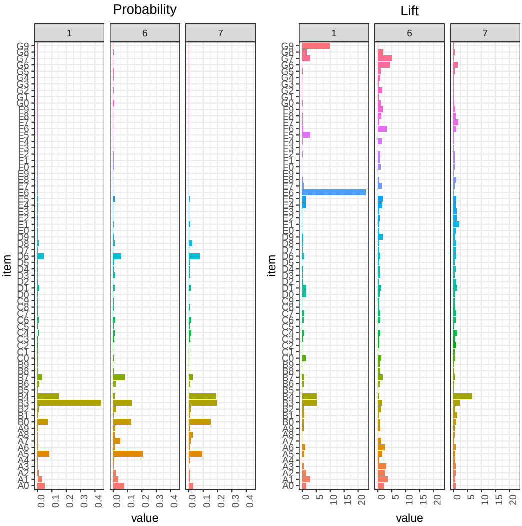
- 토픽마다 리프트가 큰 두 상품을 주요 상품으로 정의하고 이것으로 토픽의 이름을 붙였다.

```
# 삼품의 비율
p=colSums(lda$topics)/sum(lda$topics)

# lift 계산
lift=matrix(0,nrow=K,ncol=W)
colnames(lift)=item_name
for(i in 1:K){
  lift[i,p!=0]=phi[i,p!=0]/p[p!=0]
  lift[i,p==0]=0
}

# 토픽마다 lift 상위 2개 상품으로 이름을 붙임
topic_name=c()
for(i in 1:K){
  sorted=sort(lift[i,],decreasing=T)[1:2]
  topic_name=c(topic_name,paste(names(sorted),collapse=" "))
}
topic_name
```

```
##  [1] "E6 G9" "G3 G2" "F5 B5" "B5 A8" "D0 E1" "G7 G6" "B4 B3" "A9 C1"
##  [9] "E0 G1" "E3 A3" "F3 B8"
```

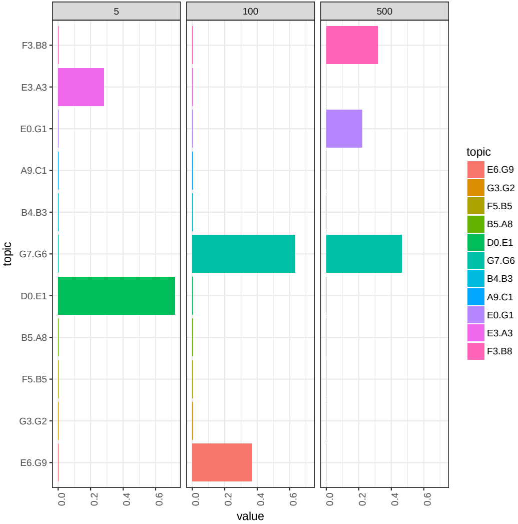
Latent Dirichlet Allocation

- 1,6,7번째 토픽의 확률과 리프트



Latent Dirichlet Allocation

- 5, 100, 500번째 고객의 토픽 비율



직접해보기

KBS(KBS.txt)

- 2015년 사회 분야의 KBS 기사들을 전처리한 자료
- 메르스 관련 기사들은 분석에서 제외
- 총 56,008개의 기사, 12,573 종류의 단어

직접해보기

1. 자료를 lda 패키지에서 사용하는 형태로 전처리
2. 토픽 수를 20, iteration 수를 500으로 하여 LDA 분석 실시
3. 토픽마다 확률이 높은 단어들 찾아보기

참고자료 및 출처

- 김도형, KOSSDA Summer 2016 Workshop: Bayesian Statistics Intro, 2016.
- Peter D. Hoff, A First Course in Bayesian Statistical Methods, Springer, 2010.
- SungChul Seo. et al., "GIS-based Association Between PM10 and Allergic Diseases in Seoul: Implications for Health and Environmental Policy", Allergy Asthma Immunol Res. 2016 Jan; 8(1): 32-40.
- W.R.Gilks. et al., Markov Chain Monte Carlo In Practice, Chapman & Hall, 1998.
- Murphy, Kevin P. "Conjugate Bayesian analysis of the Gaussian distribution." def 1.2σ2 (2007): 16.
- `https://www.johndcook.com/t_normal_mixture.pdf`
- `http://www.math.usu.edu/jrstevens/biostat/PoissonNB.pdf`

참고자료 및 출처

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
- David Blei, "Probabilistic Topic Models", MLSS, 2012.
- 김용대, 정구환, "토픽모형을 이용한 빅데이터 기반 마이크로 세그멘테이션 방법론 연구", 통계연구, 2016.
- Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.