이항분포-베타사전분포

- 확률모형

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

- 사전분포

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

- 사후분포

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

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

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
```

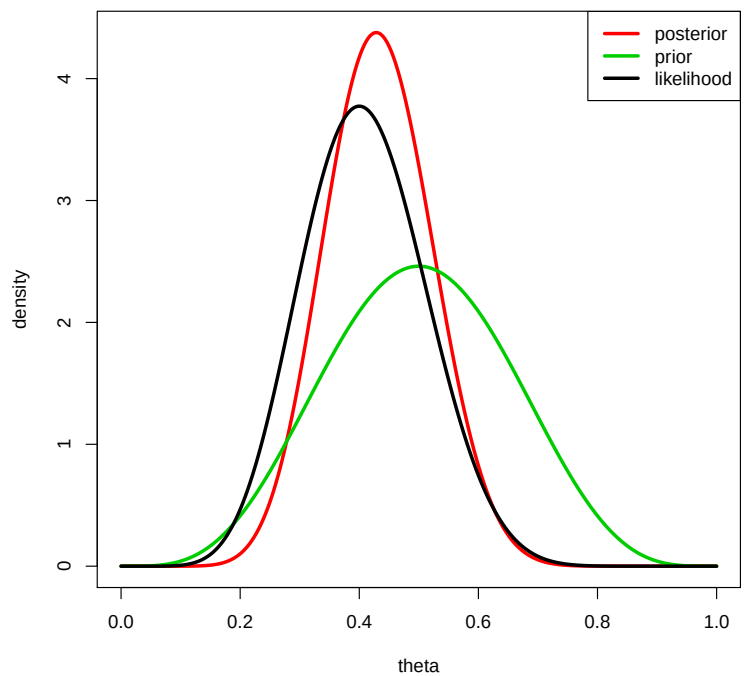
```
## [1] 0.007921147
```

```
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))
```



몬테카를로 방법

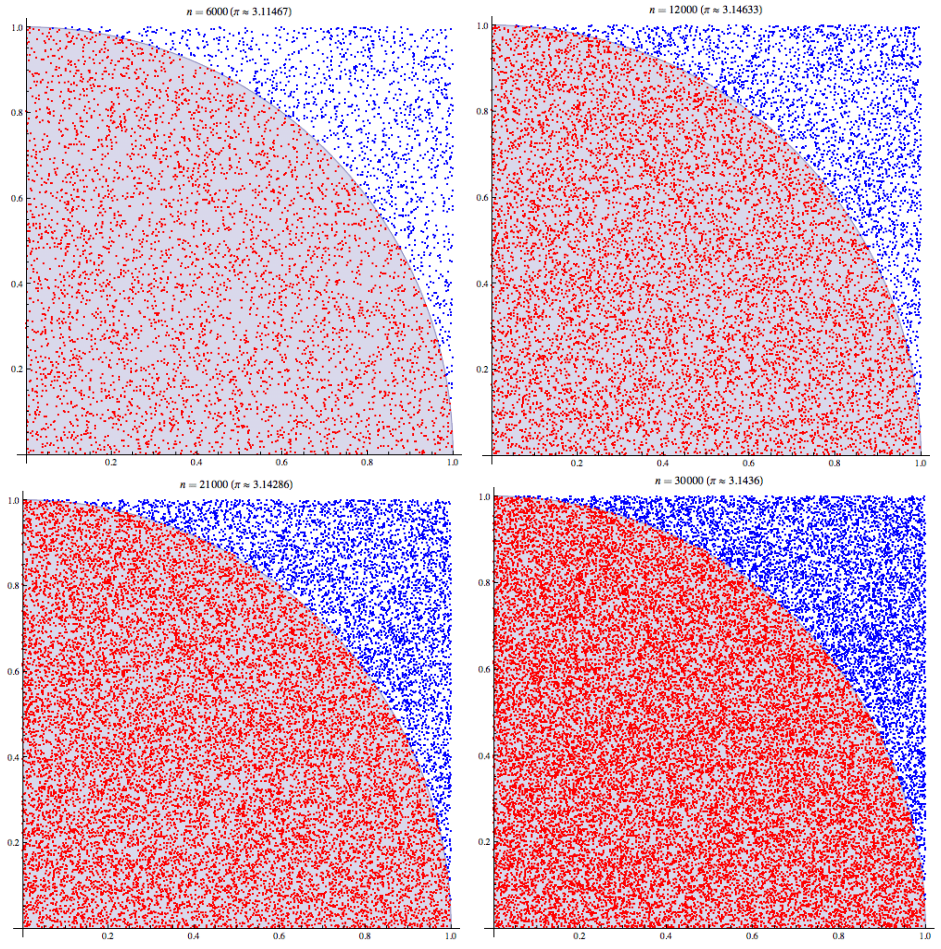
예. 원주율 문제

- 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

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

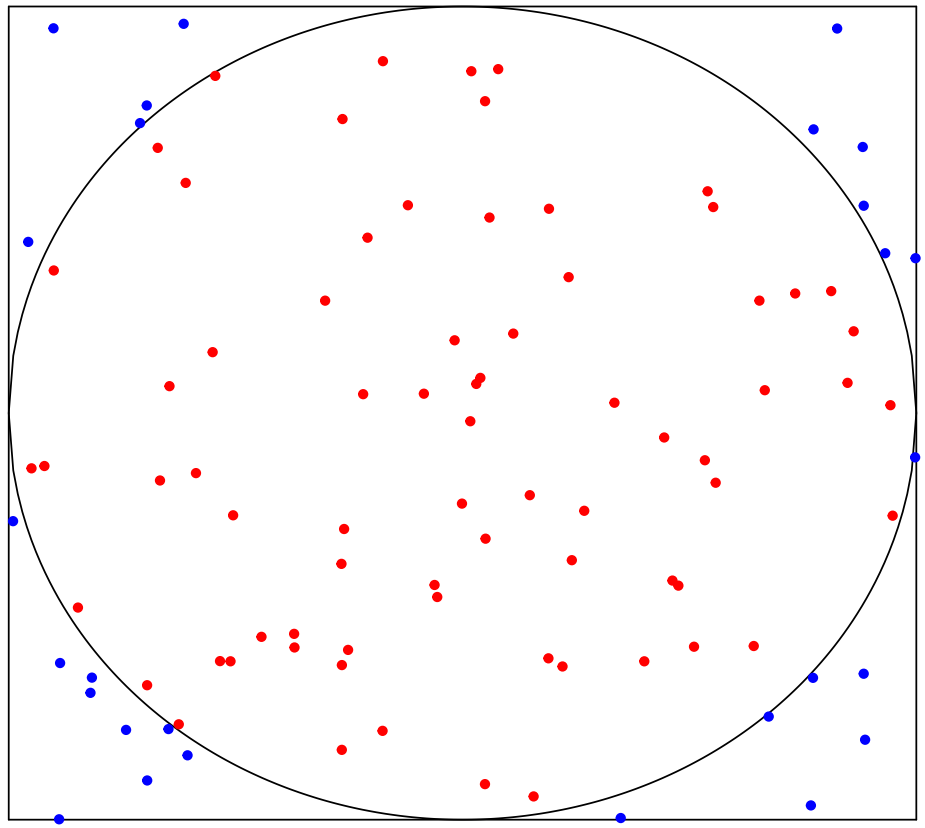


몬테카를로 방법

```
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)
}
```

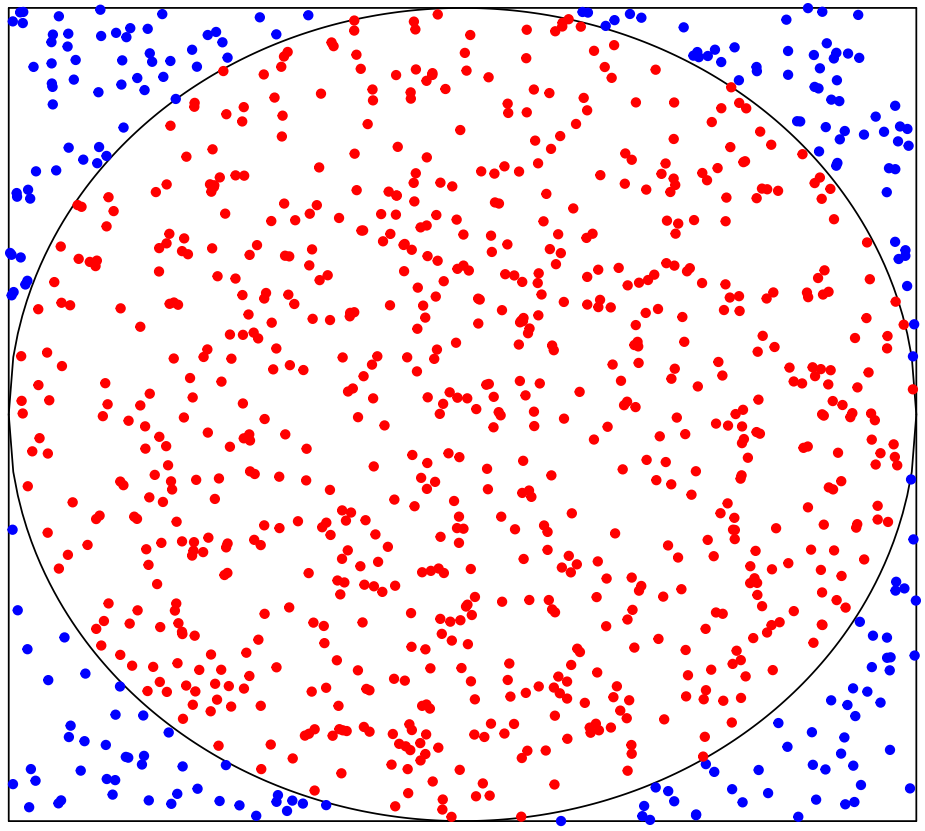
몬테카를로 방법

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



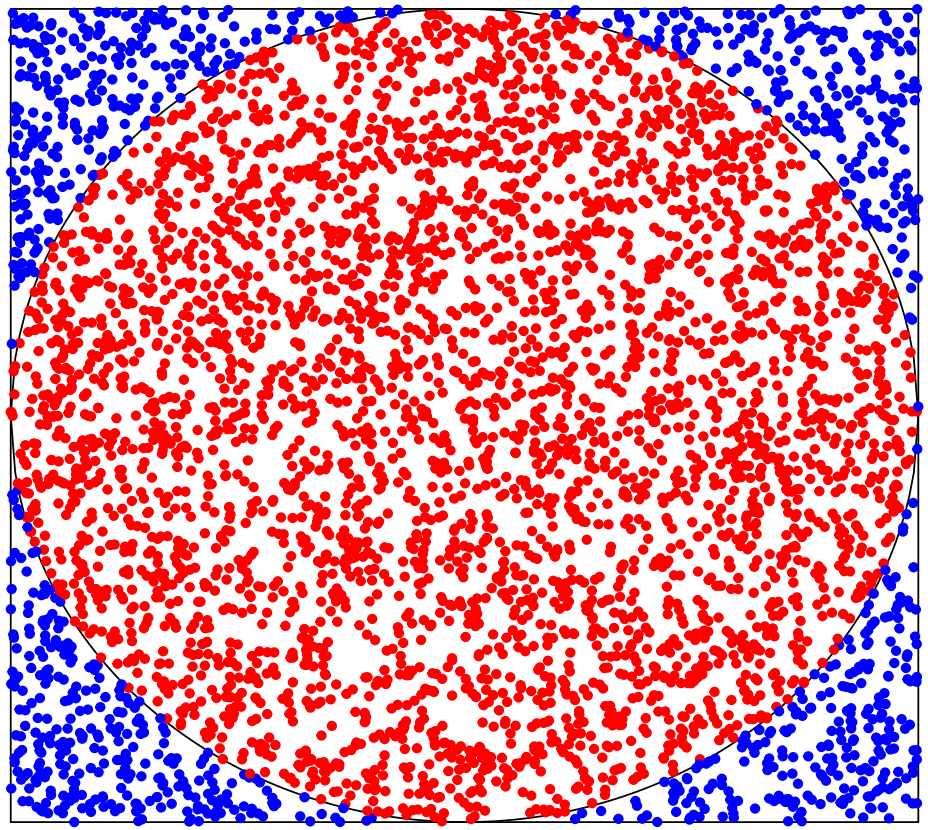
몬테카를로 방법

```
CalculatePi(1000)
```



몬테카를로 방법

```
CalculatePi(5000)
```



몬테카를로 방법

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

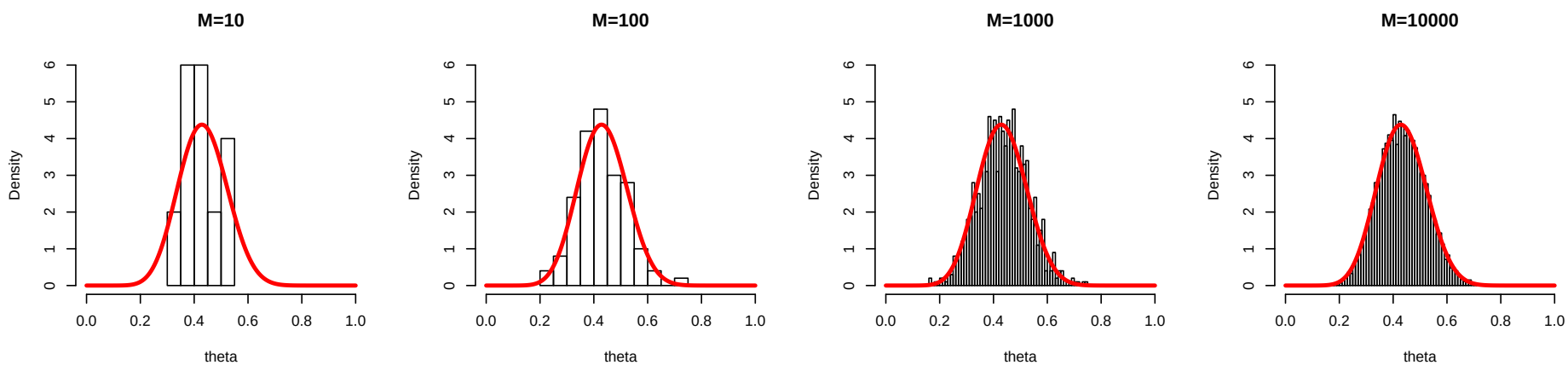
$$\theta^{(1)}, \ldots, \theta^{(M)} \overset{iid}{\sim} P(\theta|X_1, \ldots, X_n)$$

- $\theta^{(1)}, \ldots, \theta^{(M)}$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

```
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
```

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.

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

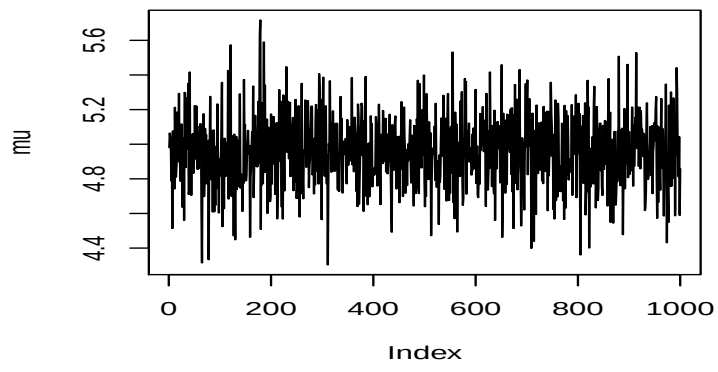
$$\mu|\lambda, X_1, \ldots, X_n \sim \mathit{Normal}\left(\frac{\kappa_0\mu_0 + \sum_{i=1}^n X_i}{n+\kappa_0}, \frac{1}{\lambda(n+\kappa_0)}\right)$$

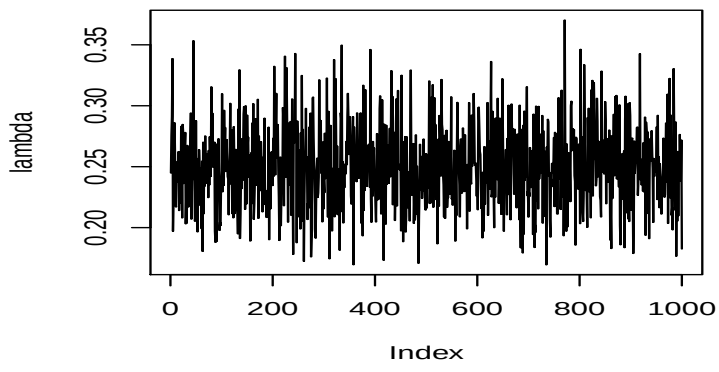
정규분포-정규감마사전분포

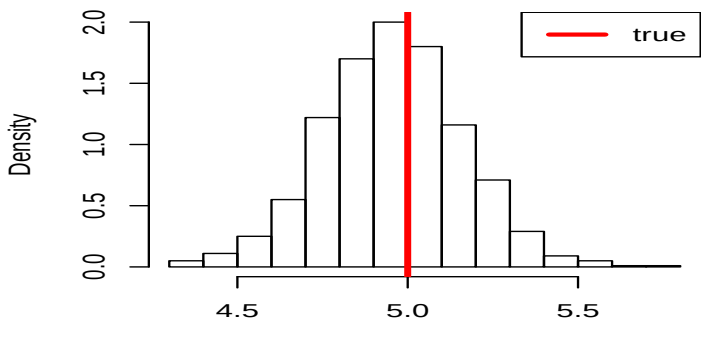
```
set.seed(100)
X=rnorm(100,5,2) # Normal(5,4)
Gibbs=function(X, alpha, beta, m0, k0, iter, burnin, thinning){
  mus=numeric(iter)
  lambdas=numeric(iter)
  Xsum=sum(X)
  n=length(X)
  mu=1
  for(i in 1:iter){
    lambdas[i]=lambda=rgamma(1,alpha+(n+1)/2,
                beta+(k0*(mu-m0)^2+sum((X-mu)^2))/2)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

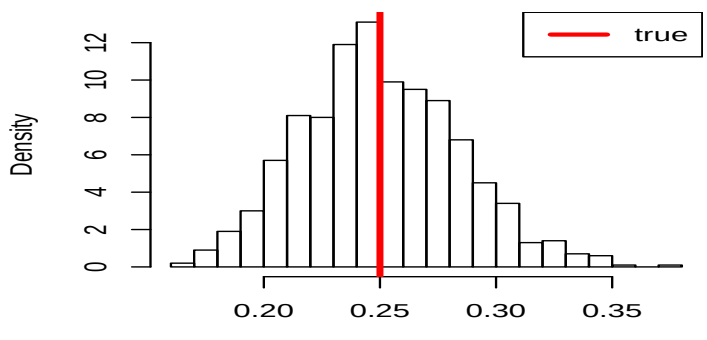
정규분포-정규감마사전분포

```
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)
```









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

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

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

$$P(\theta|X_1, \ldots, X_n) \propto P(\theta) \prod_{i=1}^{n} P(X_i|\theta)$$

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

$$\propto \frac{1}{1+\theta^2} \exp\left\{-\frac{n(\theta-\bar{X})^2}{2}\right\} \quad \text{where } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- 알려진 형태의 분포가 아니기 때문에 직접 표본을 추출하기가 어렵다.

정규분포-코시사전분포

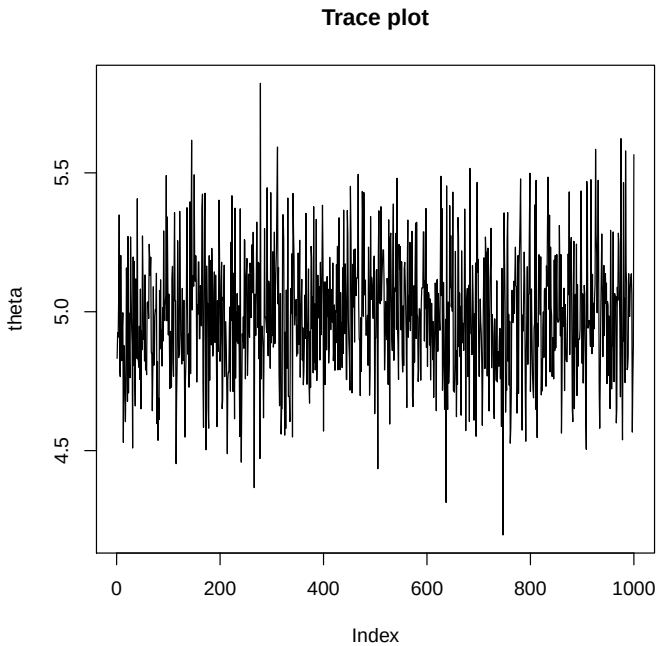
- 사후분포 : $P(\theta|X_1,\ldots,X_n) \propto \frac{1}{1+\theta^2}\exp\left\{-\frac{n(\theta-\bar{X})^2}{2}\right\} =: \pi(\theta)$
- 제안분포 : $q(\cdot|u) = Normal(u,1)$
- 수락확률 : $\alpha(\theta^{(t)},\tilde{\theta}) = \min\left\{\frac{\pi(\tilde{\theta})q(\theta^{(t)}|\tilde{\theta})}{\pi(\theta^{(t)})q(\tilde{\theta}|\theta^{(t)})},1\right\}$

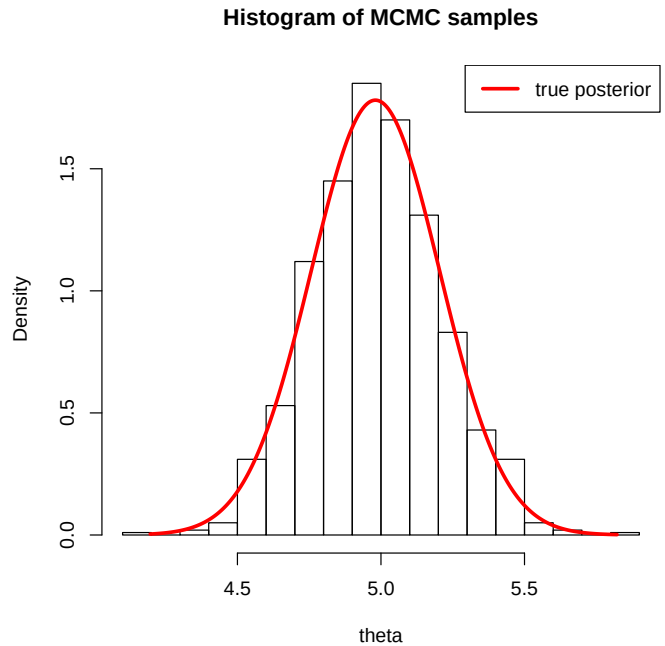
정규분포-코시사전분포

```
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

```
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

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

- 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

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

$$
\beta|\tau, \mathbf{X}, \mathbf{Y} \sim MVN\left(\hat{\beta}, \hat{\Sigma}\right)
$$

$$
\text{where } \hat{\Sigma} = \left(\Sigma_0^{-1} + \tau\mathbf{X}^T\mathbf{X}\right)^{-1},\ \hat{\beta} = \hat{\Sigma}\left(\Sigma_0^{-1}\beta_0 + \tau\mathbf{X}^T\mathbf{Y}\right)
$$

  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)

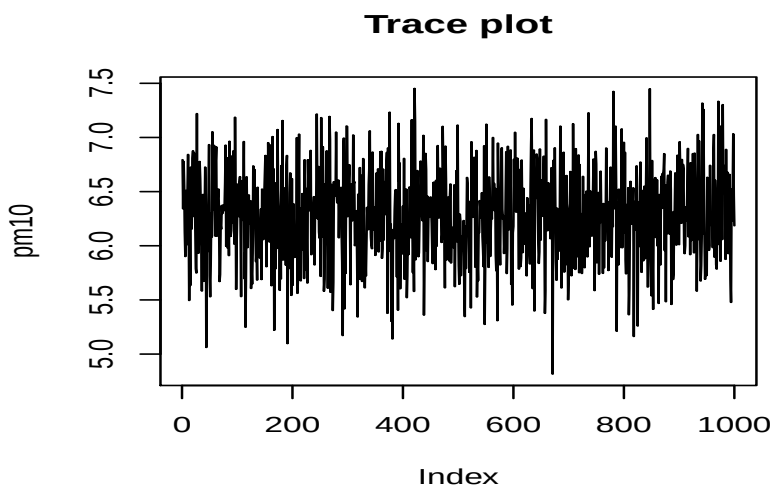
(SungChul Seo. et al., “GIS-based Association Between PM10 and Allergic Diseases in Seoul: Implications for Health and Environmental Policy”)

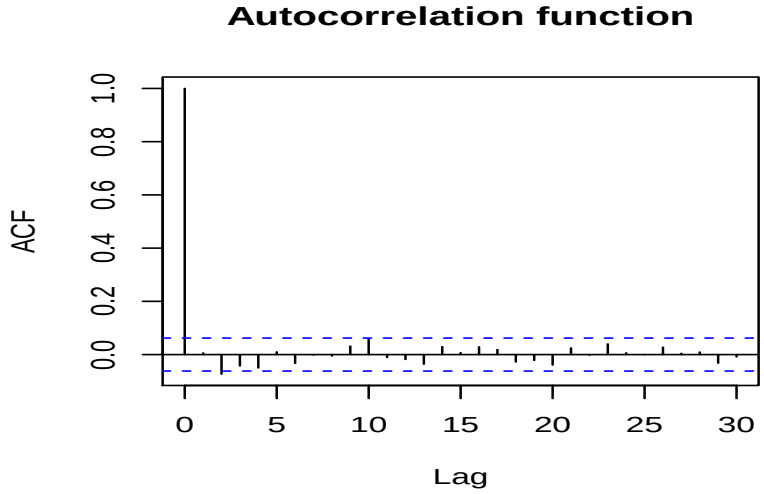
베이지안 회귀분석

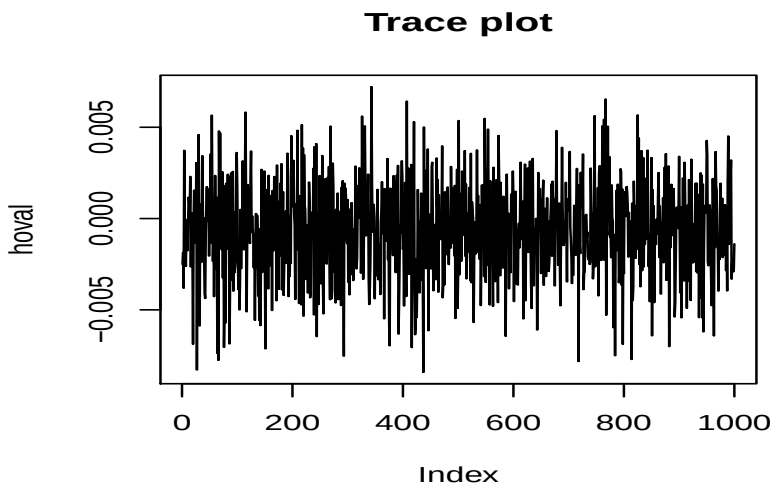
```
data <- read.csv("C:/Users/JKH/Downloads/asthma.csv")
X <- as.matrix(data[,3:6])
Y <- as.matrix(data[,2])
beta0=rep(0,ncol(X)); Sigma0inv=diag(rep(1,ncol(X))); u=1; v=1   # prior

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
```

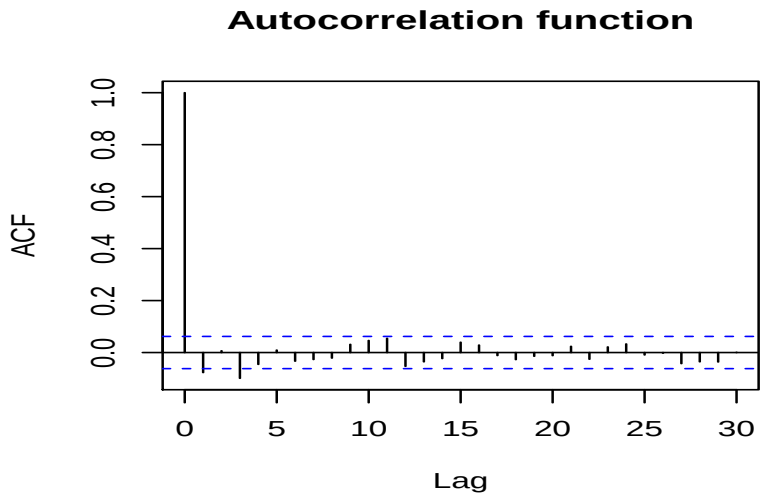
베이지안 회귀분석

- β_{pm10}, β_{hoval}



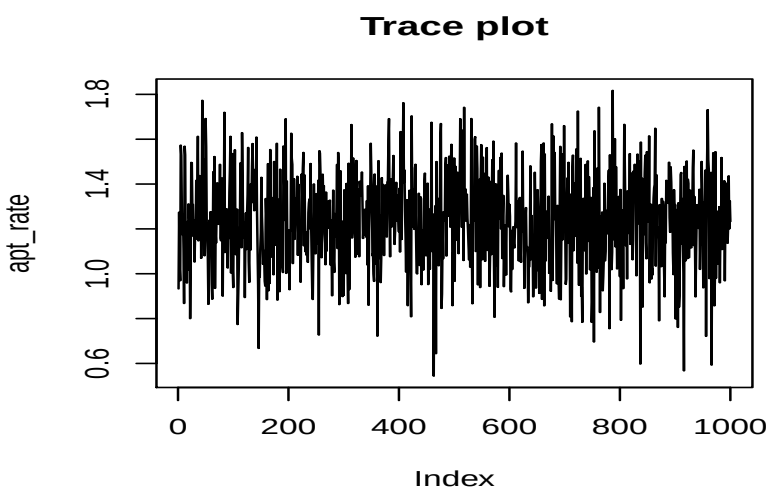


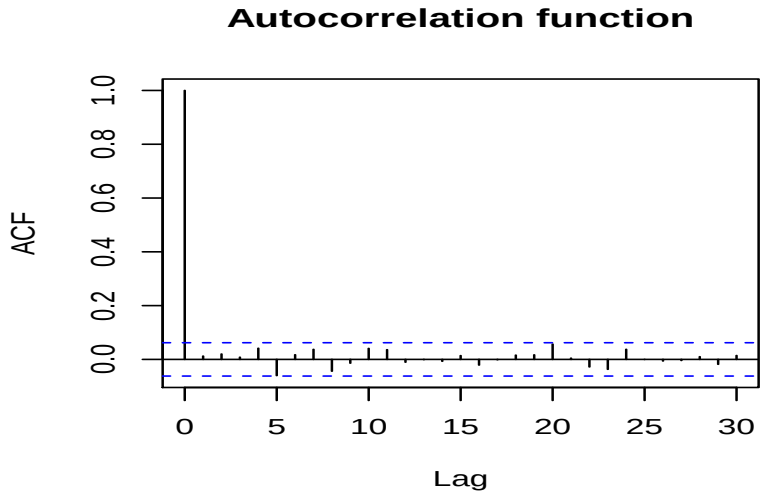


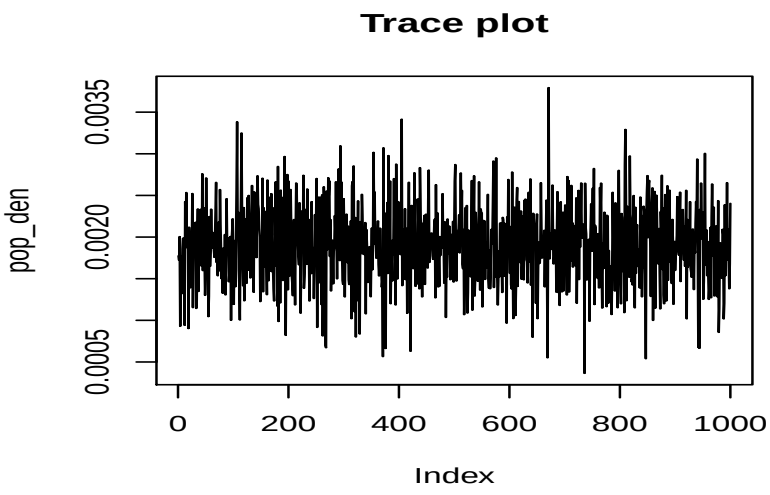


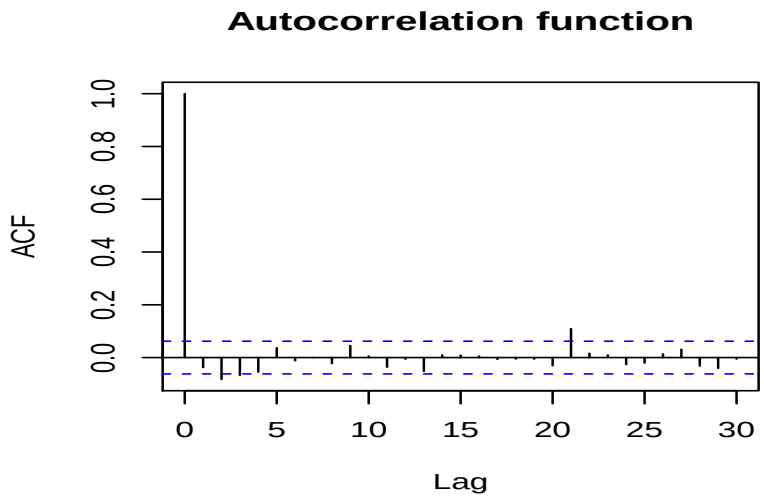
베이지안 회귀분석

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



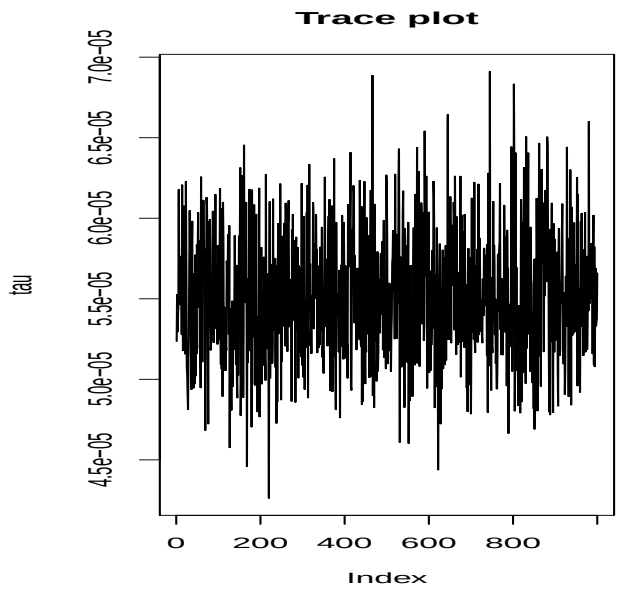


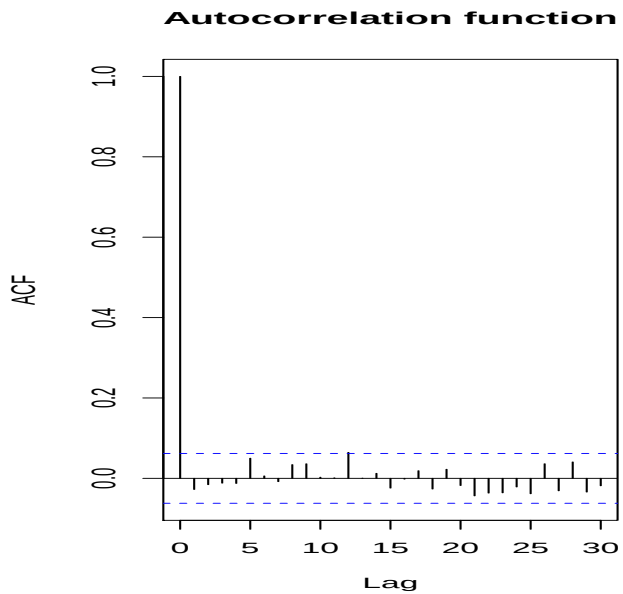




베이지안 회귀분석

- τ





혼합모형

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

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

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

- 본 강의에서는 그룹의 수 K가 알려졌다고 가정한다.

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

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

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

베이지안 정규분포 혼합모형

- 확률모형

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

  - 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

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

where $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$

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. $\boldsymbol{\mu}$

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

$$
\boldsymbol{\mu}_k | \sigma_k^2, \mathbf{Z}, \mathbf{X} \sim \mathit{Normal}_d(\hat{\boldsymbol{\mu}}_k, \hat{\sigma}_k^2 \mathbf{I}_d)
$$

$$
\text{where } \hat{\sigma}_k^2 = \frac{\lambda\sigma_k^2}{n_k\lambda + 1} \text{ and } \hat{\boldsymbol{\mu}}_k = \frac{1}{n_k\lambda + 1} \boldsymbol{\mu}_0 + \frac{n_k\lambda}{n_k\lambda + 1} \left(\frac{1}{n_k} \sum_{Z_i = k} X_i \right)
$$

베이지안 정규분포 혼합모형

3. σ^2

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

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

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

베이지안 정규분포 혼합모형

4. $\mathbf{Z}$

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

베이지안 정규분포 혼합모형

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

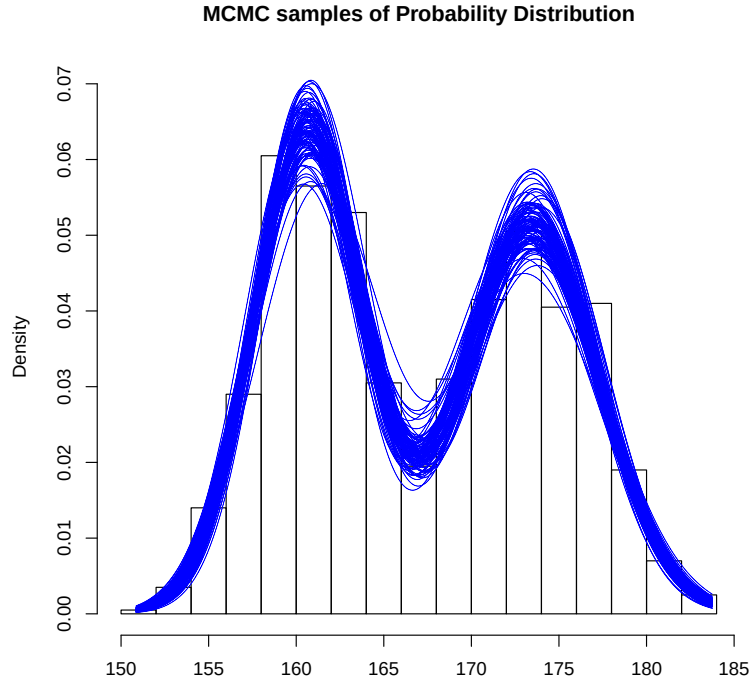
```
set.seed(10)
M=rnorm(500,173.5,3.8); W=rnorm(500,160.5,3.2);  # true
X=c(M,W)
alpha=1; mu0=mean(X); lambda=1; u=5; v=5;

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)
```

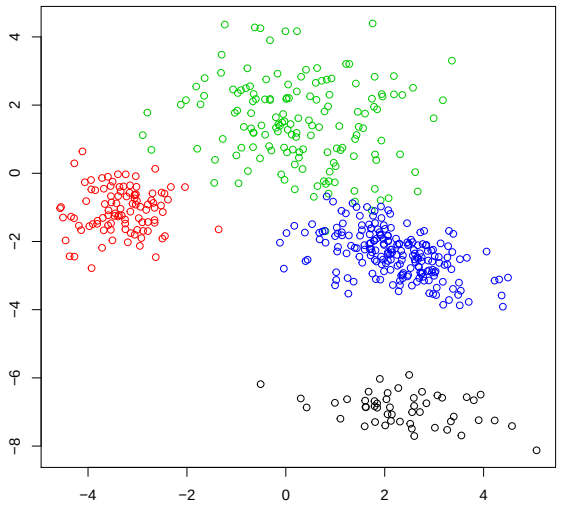


k-평균 vs 혼합모형

- 4개의 군집을 갖는 2차원 상의 자료를 임의로 생성

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

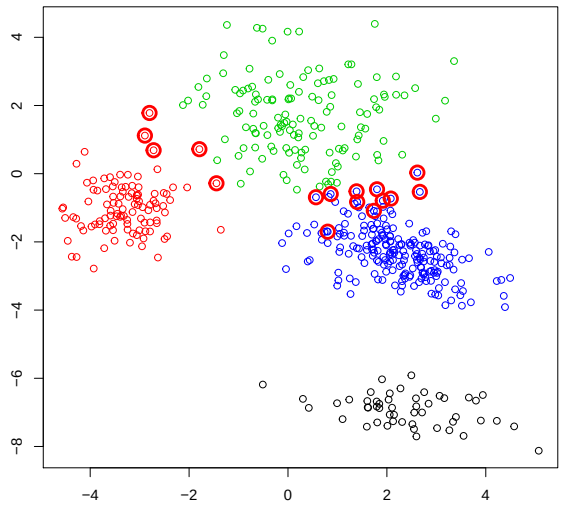
```
mu=matrix(rnorm(8,sd=4),ncol=2)
Sigma=array(0,dim=c(4,2,2))
x=c()
for(i in 1:4){
  diag(Sigma[i,,])=rgamma(2,3,4)
  Sigma[i,1,2]=Sigma[i,2,1]= runif(1,-0.5,0.5)*sqrt(Sigma[i,1,1]*Sigma[i,2,2])
  x=rbind(x,mvrnorm(50*i,mu[i,],Sigma[i,,]))
}
clust0=c(rep(1,50),rep(2,100),rep(3,150),rep(4,200))
plot(x,col=clust0,xlab="",ylab="")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



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)
```

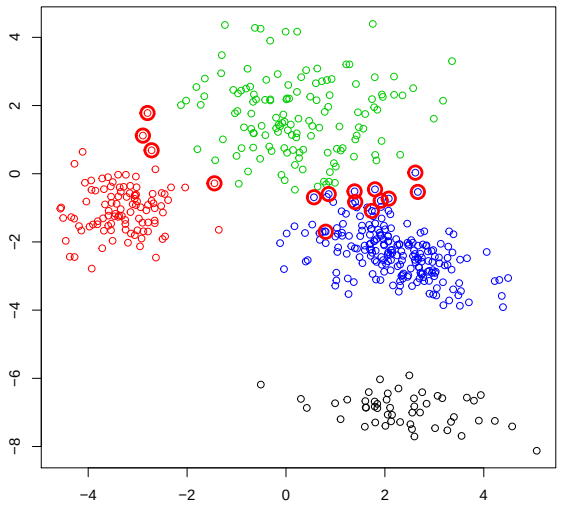


- 빨간색 큰 원은 오분류를 나타냄

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)
```

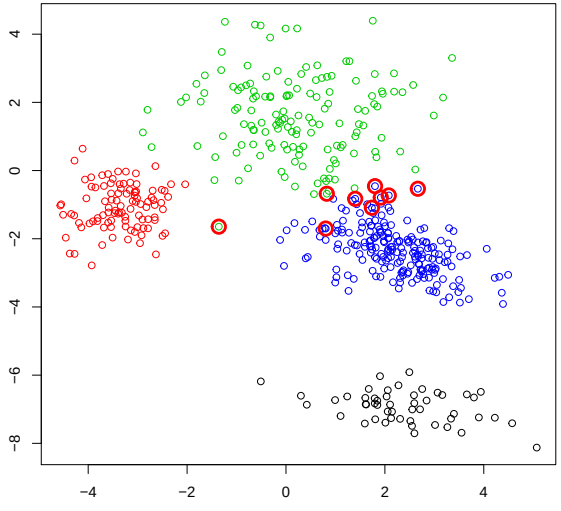


- k-평균과 유사

k-평균 vs 혼합모형

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

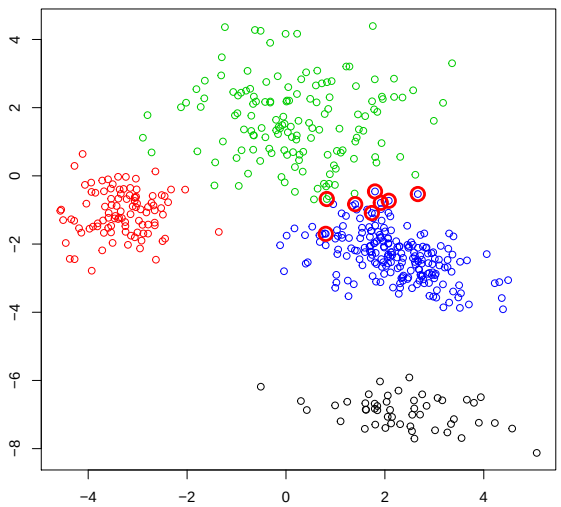
```
mix_VII=Mclust(x,4,"VII")
clust3=apply(mix_VII$z,1,which.max)
plot(x,col=clust3,xlab="",ylab="")
points(x[which(clust0!=clust3),],col="red",cex=2,lwd=3)
```



k-평균 vs 혼합모형

- 혼합모형 : $\Sigma_k = diag(\sigma_{k1}^2, \ldots, \sigma_{kd}^2)$ (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)
```



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

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.

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

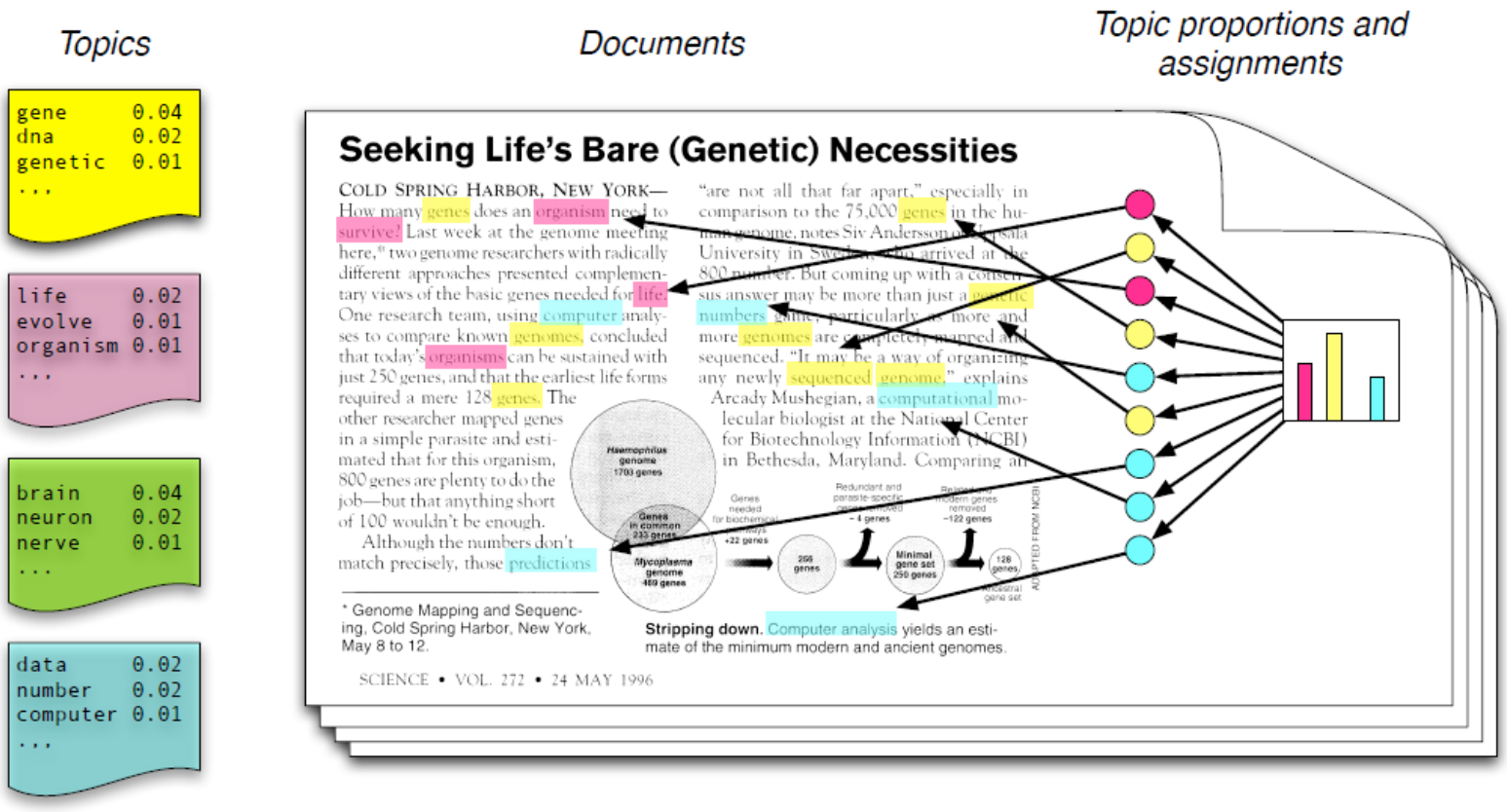
- 토픽모형에서는 문서분포 P_j를 토픽들의 혼합모형으로 가정한다.

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

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.

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

- 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

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

- 사전분포

$$\theta_j \sim \textit{Dirichlet}(\alpha, \ldots, \alpha) \quad \text{for } j = 1, \ldots, n$$
$$\phi_k \sim \textit{Dirichlet}(\beta, \ldots, \beta) \quad \text{for } k = 1, \ldots, K$$

Latent Dirichlet Allocation

- 사후분포

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

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. $\boldsymbol{\theta}$

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

$$\theta_j|\mathbf{z} \sim \mathit{Dirichlet}(\alpha + N_{j1}, \ldots, \alpha + N_{jK})$$

  2. $\boldsymbol{\phi}$

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

$$\phi_k|\mathbf{z} \sim \mathit{Dirichlet}(\beta + N_{k1}, \ldots, \beta + N_{kW})$$

  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 θ와 ϕ를 적분하여 없애고 $\mathbf{z}$만 추출하는 깁스 샘플러가 가능하다.
- θ와 ϕ는 추출된 $\mathbf{z}$와 사후분포의 조건부분포 $P(\theta|\mathbf{z})$, $P(\phi|\mathbf{z})$를 통해 얻을 수 있다.

Latent Dirichlet Allocation

- 붕괴 깁스 샘플러

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

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

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

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

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

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 *a*를 *b*회 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.

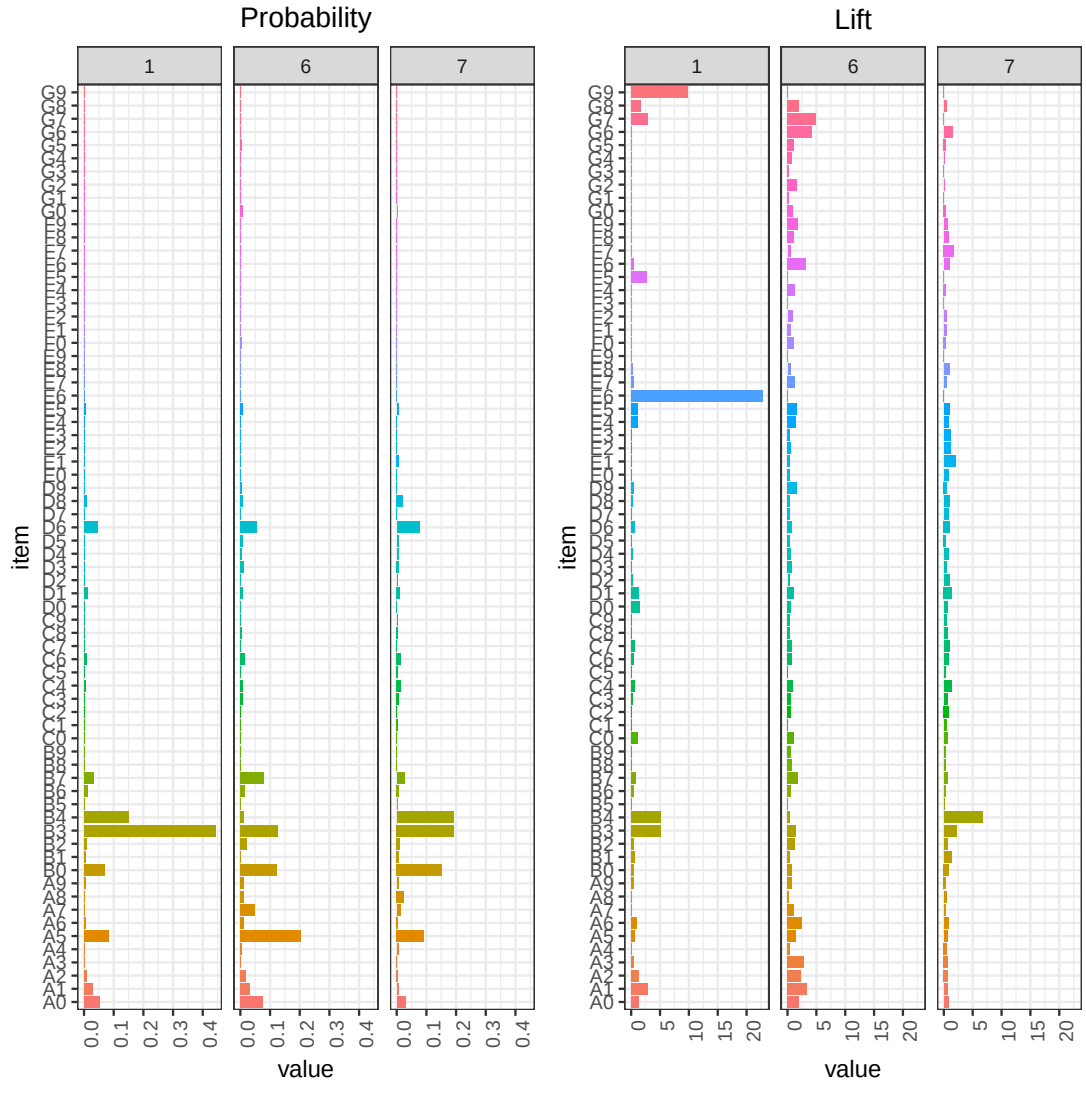
Latent Dirichlet Allocation

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

```
# 상품의 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"
```

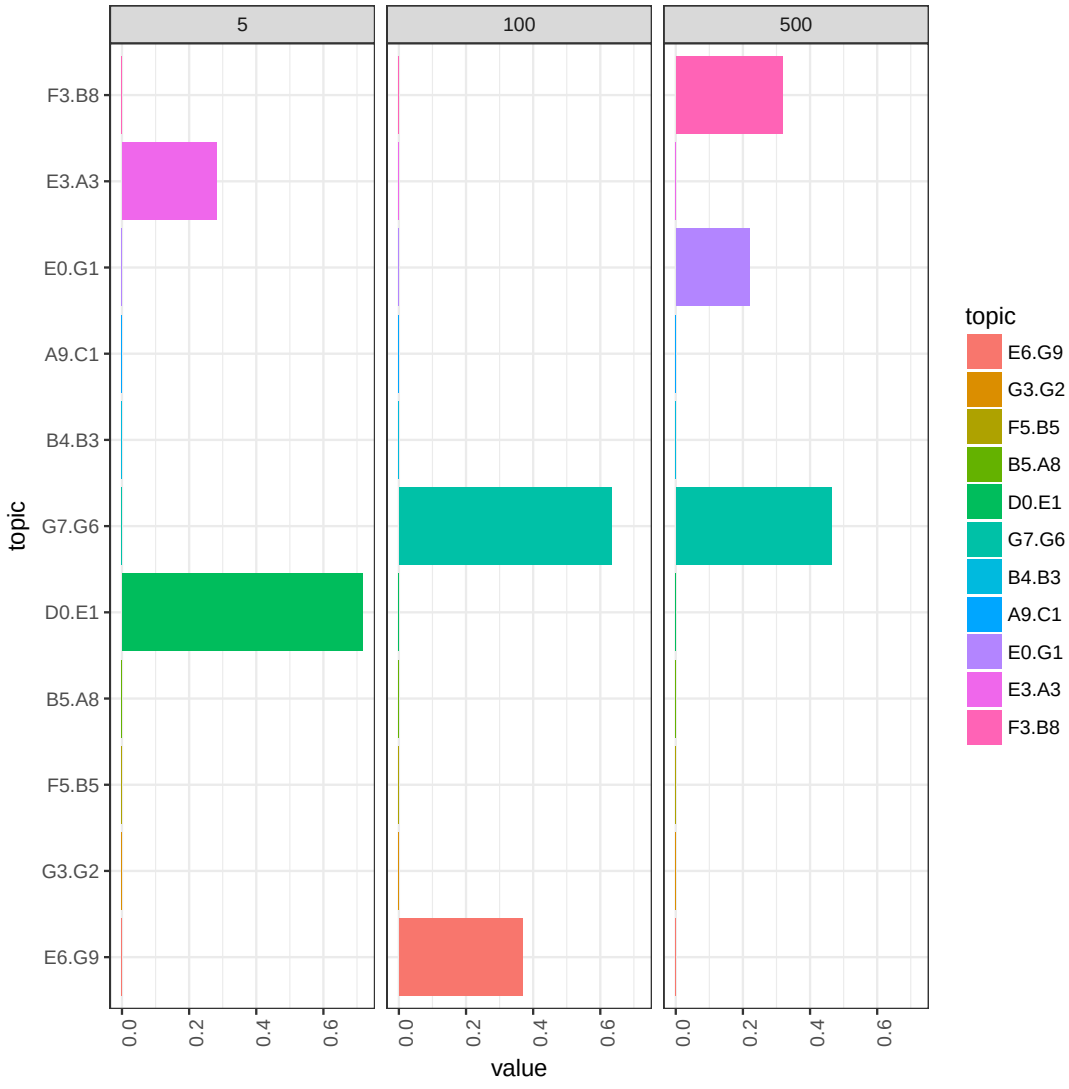
Latent Dirichlet Allocation

- 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