서울대학교 통계학과 **이재용 교수**

09강

베이지안 통계학

깁스 추출법

한국방송통신대학교
Korea National Open University

목차

01 깁스추출법

고차원 확률변수 생성의 문제점

◘ 고차원 확률변수 생성의 문제점

삼각분포에서 합격불합격 방법으로 확률변수의 생성

* 삼각분포의 밀도함수

$$f(x)=\begin{cases} x, & 0\le x\le 1, \\ 2-x, & 1\le x\le 2 \end{cases}$$

* 제안분포

$$g(x)=1, 0\le x\le 2$$

* 합격율

$$\frac{1}{2}$$

◘ p차원 삼각분포에서 합격불합격 방법으로 확률변수의 생성

* 삼각분포의 밀도함수

$$f(x)=\prod_{i=1}^{p} f(x_i), x=(x_1,\dots,x_p)$$

* 제안분포

$$g(x)=\prod_{i=1}^{p} g(x_i)$$

* 합격율

0.5^p

예. $0.5^{10}=0.000976,$

$0.5^{100}=7.888\times 10^{31}.$

차원이 커질수록 확률을 생성하기가 현실적으로 어려워진다.

깁스추출법 알고리듬

- 깁스추출법은 다변량 분포함수에서 확률변수를 추출하는 알고리듬이다.
- p차원 확률변수를 한번에 생성하지 않고, 한번에 하나의 원소를 조건부 분포에서 추출한다.
- 생성된 확률변수의 열은 목표분포 $\pi(x_1,\ldots,x_p)$를 정상분포로하는 마르코프체인이 된다.
- 적당한 조건하에서, 깁스알고리듬 결과인 $(X^{(0)},X^{(1)},X^{(2)}\ldots)$의 표본적률과 표본분위수는 $\pi(x_1,\ldots,x_p)$의 적률과 분위수로 수렴한다.
- 깁스추출법의 장점
  - 다변량 확률변수의 생성을 1차원 확률변수의 생성 문제로 바꾸었다.

$(x_1,\ldots,x_n)$의 밀도함수가 $\pi(x_1,\ldots,x_p)$이라하고,
$x_{-i}:=(x_1,\ldots,x_{i-1},x_{i+1},\ldots x_p)$라 표시하자.

- 깁스 알고리듬

초기값$(X_1^{(0)},\ldots,X_p^{(0)})$을 선정한다.
t=1,2,3,...,에 대하여

단계1. $X_1^{(t)}\sim \pi_{X_1|X_{-1}}\left(\cdot\,\middle|\,X_2^{(t-1)},\ldots,X_p^{(p-1)}\right)$ 을 추출한다.

단계2. $X_2^{(t)}\sim \pi_{X_2|X_{-2}}\left(\cdot\,\middle|\,X_1^{(t)},X_3^{(t-1)},\ldots,X_p^{(t-1)}\right)$ 을 추출한다.

$\vdots$

단계3. $X_j^{(t)}\sim \pi_{X_j|X_{-j}}\left(\cdot\,\middle|\,X_1^{(t)},\ldots,X_{j-1}^{(t)},X_{j+1}^{(t-1)},\ldots,X_p^{(t-1)}\right)$을 추출한다.

$\vdots$

단계4. $X_p^{(t)}\sim \pi_{X_p|X_{-p}}\left(\cdot\,\middle|\,X_1^{(t)},\ldots,X_{p-1}^{(t)}\right)$을 추출한다.

예. 이변량 정규분포

이변량정규분포

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right)$$

를 정상분포를 갖는 깁스추출표본을 구하자.

참고: 이변량 정규분포의 조건부 분포

$$(X_1, X_2) \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right) \text{일 때,}$$

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

$$X_2|X_1 = x_1 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

가 된다.

초기값 $(X_1^{(0)}, X_2^{(0)})$을 정한다. t=1,2,...,에 대해서,

(i) $x_1^{(t)} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(x_2^{(t-1)} - \mu_2\right), \sigma_1^2(1-\rho^2)\right)$를 발생한다.

(ii) $x_2^{(t)} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x_1^{(t)} - \mu_1\right), \sigma_2^2(1-\rho^2)\right)$를 발생한다.

$(x_1^{(t)}, x_2^{(t)}), t \geq 0$를 이용해 정규분포의 적률과 분위수를 구한다.

$$\int h\ (x_1, x_2)\pi(dx_1, dx_2) = \frac{1}{n}\sum_{t=1}^{n} h(x_1^{(t)}, x_2^{(t)})$$

$$\text{i.e., } P(x_1 \geq \widehat{0, x_2} \geq 0) = \frac{1}{n}\sum_{t=1}^{n} I\left(x_1^{(t)} \geq 0, x_2^{(t)} \geq 0\right).$$

예. 절단된 정규분포

$$\pi(x) \propto e^{-\frac{1}{2\sigma^2}(X-\mu)^2} I(x \geq A), A \in \mathbb{R}$$

라 하자.
π를 정상분포로 갖는 마르코프 체인 $x^{(t)}$을 생성하고자 한다.

아이디어

$$\pi(x) \propto \int I(0 \leq z \leq e^{-\frac{1}{2\sigma^2}(X-\mu)^2}) I(x \geq A) dz$$

이라는 것에 착안해서,

$$\pi(x, z) \propto I(0 \leq z \leq e^{-\frac{1}{2\sigma^2}(X-\mu)^2}) I(x \geq A)$$

이라 놓는다.
$\pi(x, z)$를 정상분포로 갖는 마르코프 체인$(x^{(t)}, z^{(t)})$ 를 생성하면,
x^t는 $\pi(x)$를 정상분포로 갖는 마르코프체인이 된다.

예. 절단된 정규분포. 알고리듬

◘ 조건부 분포

$$\pi(x|z) \propto I\left(AV(\mu - \sqrt{-2\sigma^2 \log z}) \leq x \leq \mu + \sqrt{-2\sigma^2 \log z}\right)$$

$$\pi(z|x) \propto I\left(0 \leq z \leq e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right)$$

$$\therefore x|z \sim Unif\left(AV(\mu - \sqrt{-2\sigma^2 \log z}) \leq x \leq \mu + \sqrt{-2\sigma^2 \log z}\right)$$

$$z|x \sim Unif\left(0, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right)$$

* 참고

1. 이것은 분할추출법(slice sampler)의 일종이다.
2. 변수들을 추가해서 추출하기 쉬운 분포를 얻는 방법을 자료확대(data augmentation, 자료덧붙임)이라 한다.

02 수렴진단

번인과 가늘게하기

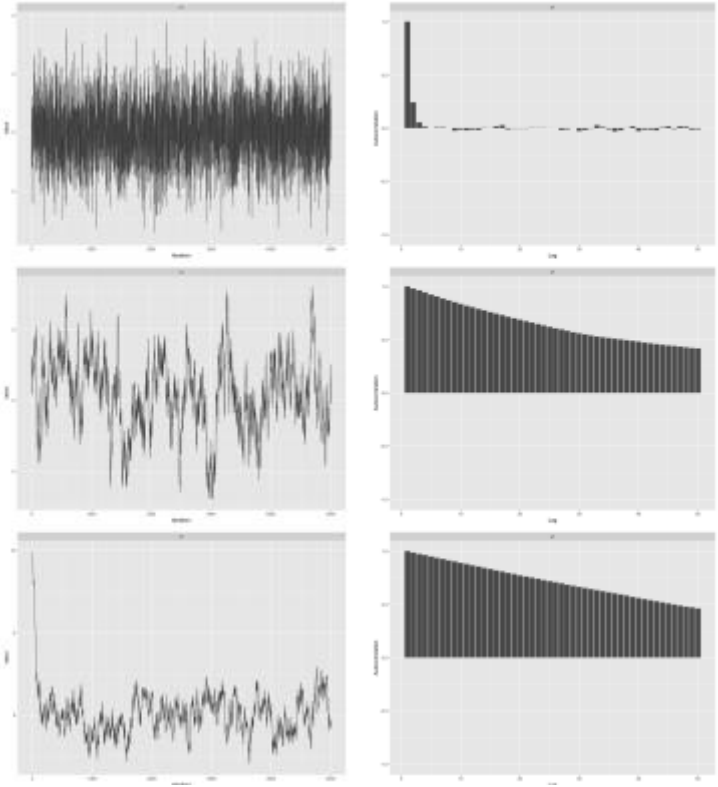
1) 번인(Burn-in)

* 엠씨엠씨 표본의 앞부분을 버리는 것을 말한다.

2) 가늘게하기(thinning)

* 엠씨엠씨 표본의 자기상관계수가 높을 때, 모든 표본을 다 사용하지 않고 r번째 표본만 따로 추출해서 사용하는 것을 말한다. 가늘게하기로 컴퓨터의 메모리를 절약할 수 있다.

시계열그림과 자기상관계수 그림을 이용한 수렴진단



맨 위 행의 그림은 마르코프체인에 문제가 없어 보인다.

두번째 행은 자기 상관계수가 너무 크다.

이 때는 가늘게하기를 할 필요가 있다.

세번째 행은 마르코프체인이 초기값에 의존한다.

번인을 할 필요가 있다.

03 실습

실습. 이변량정규분포. 문제

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right)$$

를 정상분포를 갖는 깁스추출표본을 구하시오,

$\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1, \rho = 0.3, 0.9$일 때
두 가지의 경우에 대해
깁스추출표본을 구하시오.

1. x_1, x_2의 요약통계량을 구하시오
2. ρ와 $\mathbb{P}(x_1 > 0, x_2 > 0)$을 추정하시오.

실습. 이변량정규분포.

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right)$$

를 정상분포를 갖는 깁스추출표본을 구하시오, $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1, \rho = 0.3, 0.9$일 때 두 가지의 경우에 대해 깁스추출표본을 구하시오.

1. x_1, x_2의 요약통계량을 구하시오
2. ρ와 $\mathbb{P}(x_1 > 0, x_2 > 0)$을 추정하시오.

초기값 $(x_1^{(0)}, x_2^{(0)})$을 정한다.

$t=1,2,\ldots,m$에 대해서,

1. $x_1^{(t)} \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(x_2^{(t-1)} - \mu_2\right), \sigma_1^2(1-\rho^2))$를 발생한다.
2. $x_2^{(t)} \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}\left(x_1^{(t)} - \mu_1\right), \sigma_2^2(1-\rho^2))$를 발생한다.

실습. 이변량정규분포. 코드

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right)$$

를 정상분포를 갖는 깁스추출표본을 구하시오, $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1, \rho = 0.3, 0.9$일 때 두 가지의 경우에 대해 깁스추출표본을 구하시오.

1. x_1, x_2의 요약통계량을 구하시오
2. ρ와 $\mathbb{P}(x_1 > 0, x_2 > 0)$을 추정하시오.

- 파라미터 값들을 정한다.

```
m = 5000
mu1 = 0
mu2 = 0
sig1 = 1
sig2 = 1
rho = 0.5
```

- 깁스 샘플러의 초기화

```
po.x1 = NULL
po.x2 = NULL
x1 = mu1
x2 = mu2
```

- 깁스추출

```
for(j in 1:m) {
    cmean = mu1 + rho*sig1/sig2*(x2 - mu2)
    csd = sig1*sqrt(1-rho^2)
    x1 = rnorm(1, cmean, csd)
    cmean = mu2 + rho*sig2/sig1*(x1 - mu1)
    csd = sig2*sqrt(1-rho^2)
    x2 = rnorm(1, cmean, csd)
    po.x1 = c(po.x1, x1)
    po.x2 = c(po.x2, x2)
}
```

실습. 이변량정규분포. 코드

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2, & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2, & \sigma_2^2\end{pmatrix}\right)$$

를 정상분포를 갖는 깁스추출표본을 구하시오, $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1, \rho = 0.3, 0.9$일 때 두 가지의 경우에 대해 깁스추출표본을 구하시오.

1. x_1, x_2의 요약통계량을 구하시오
2. ρ와 $\mathbb{P}(x_1 > 0, x_2 > 0)$을 추정하시오.

coda를 이용한 요약통계량과 그림들

```
library(coda)
library(dplyr)
post.mcmc = as.mcmc(post.df)
summary(post.mcmc)
#summary(post.mcmc,
# quantiles = c(0.05, 0.25, 0.5, 0.75, 0.95))
densplot(post.mcmc)
traceplot(post.mcmc)
autocorr.plot(post.mcmc)
```

변수를 선택할 때

```
post.df %>% select(x1)
%>% as.mcmc %>% traceplot
```

rho와 확률의 추정

```
post.df %>% mutate(x12=x1*x2,
pr = as.numeric(x1 > 0 & x2 > 0))
%>% as.mcmc %>% summary
```

ggmcmc를 이용한 요약통계량과 그림들

```
library(ggmcmc)
ggs.post = ggs(post) # mcmc.list를 ggmcmc로 쓸 수 있
는 tbl_df 객체로 만든다.
ggs_density(ggs.post)
ggs_traceplot(ggs.post)
ggs_traceplot(ggs.post)
ggs_autocorrelation(ggs.post)
```

변수를 선택할 때

```
post.df %>% select(x1) %>% as.mcmc %>% ggs
    %>% ggs_traceplot
```

참고문헌

1. Hoff, P. D. (2009). A first course in Bayesian statistical methods의 6장. Springer Science & Business Media.

수고하셨습니다.

감사합니다.