5. 데이터 기반 추론 2

임채영
서울대학교 통계학과

이번 강의에서 다룰 내용

- 베이즈 추론의 개념
- MCMC, Variational Inference

베이즈 추론의 개념

베이즈 추론

- 베이즈 추론(Bayes Inference)는 통계적 추론의 한 방법으로, 추론해야 하는 대상의 사전 확률에서 데이터 관측을 통해 해당 대상의 사후 확률을 업데이트하여 추론하는 방법
- 베이즈 확률론을 기반으로 하며, 이는 추론하는 대상을 확률변수로 보아 그 변수의 확률분포를 추정하는 것을 의미

빈도주의 **v.s.** 베이즈 주의

- 빈도주의 (Frequentist approach)
- 모수를 이용해 확률분포를 정의할 수 있을 때의 경우로 한정해서 설명
- 확률변수가 특정한 분포를 따른다고 가정하고 그 분포의 모수를 추정한다. 이 때, 모수는 고정된 상수이다.
- 확률은 무한히 많은 시행에서의 상대적인 빈도로 정의
- 통계적 추론은 모수를 추정하는 것에 목적을 둔다.

빈도주의 **v.s.** 베이즈 주의

- 베이즈 주의 (Bayesian approach)

- 자료가 특정한 분포에서 나왔다고 할때, 그 분포의 모수를 고정된 상수가 아니라 확률변수로 가정한다. 즉, 모수도 분포를 가지는 것으로 간주한다.
- 확률을 빈도나 어떤 시스템의 물리적 속성으로 여기는 빈도주의와는 달리, 베이지안들은 주관주의 확률이론에 따라 확률을 어떤 사람이 특정한 순간에 주어진 명제나 사건에 대해 갖는 믿음의 정도(degree of belief)로 정의한다.
- 따라서 모수의 분포를 추정할 때 현재 관찰된 자료뿐만 아니라 이전의 자료나 연구자의 믿음 등도 고려된다.
- 새로운 자료가 수집되면 모수에 대한 추정이 업데이트 된다.

“ 동전 하나를 던졌을 때 앞면이 나올 확률이 50퍼센트이다.”에 대한 해석

빈도주의

- 동전 하나 던지기를 수천, 수만번 하면 그중에 50퍼센트는 앞면이 나오고, 50퍼센트는 뒷면이 나온다
- 객관적 확률로 해석

베이지안

- 동전 하나 던지기의 결과가 앞면이 나올 것이라는 확신은 50퍼센트이다.
- 주관적 확률로 해석

베이즈 추론에 필요한 요소

베이즈 추론을 구성하는 세 가지 요소

- 사전 분포(Prior distribution) : 모수 θ의 분포로 자료를 보기전 분석자의 θ에 관한 정보(불확실성의 정도)를 나타낸다. $\pi(\theta)$로 나타낸다.
- 데이터 모형(Data distribution) : 데이터의 분포에 관한 모형으로 $x \mid \theta \sim f(x \mid \theta)$ 또는 $\pi(x \mid \theta)$로 나타낸다.
- 사후 분포(Posterior distribution) : 데이터가 주어졌을 때, θ의 확률분포로 데이터를 본 후의 분석자의 θ 에 관한 불확실성을 나타낸다. $\pi(\theta|x)$로 나타낸다.

- 데이터 모형 $f(x \mid \theta)$, 사후분포 $\pi(\theta \mid x)$는 조건부 확률분포로 해석한다.
- 즉, $f(x \mid \theta) = f(x, \theta)/\pi(\theta)$, $\pi(\theta \mid x) = f(x, \theta)/f(x)$.
- 여기서는 $f(\cdot)$ 또는 $\pi(\cdot)$를 확률밀도함수(또는 확률질량함수)로 혼용해서 사용한다.
- 베이즈 정리를 이용하면, 사후확률은 다음과 같다.
- $\pi(\theta \mid x) = f(x \mid \theta)\pi(\theta)/f(x) \ \propto \ f(x \mid \theta)\pi(\theta)$
- 즉, 사후분포 $\propto$ 가능도 $\times$ 사전분포

모수에 대한 베이즈 추론

- 데이터가 주어졌을때의 모수에 대한 사후분포를 이용하여 추론을 한다.
- 사후분포에서 평균, 중앙값, 최빈값등을 모수 θ의 베이지안 추정값으로 사용할 수 있다. 이를 사후분포의 평균 (Posterior mean), 사후분포의 중앙값 (Posterior median), 사후분포의 최빈값 (Maximum a Posteriori, MAP)이라 한다.

정규분포 예제

- 분산 σ^2이 알려져 있을 때, 평균 θ, 분산 σ^2인 정규분포에서 관측한 데이터를 가지고 평균 θ에 대한 베이즈 추론을 해 보자.
- 모수 θ의 사전분포를 정규분포로 가정한다. $\theta \sim N(m, s^2)$
- 즉, 자료 관측 이전에 θ가 대략 m이라 믿으며, 그 불확실성이 대략 s만큼인 정규분포를 따른다고 믿는 것을 뜻한다.
- n개의 데이터 $\boldsymbol{X} = (X_1, X_2, ...X_n)$가 관측되었다고 하자. $X_i \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$.
- 이때 데이터의 확률모형은 다음과 같다.

$$
\begin{aligned}
f(x_1, \cdots, x_n|\theta, \sigma^2) &= \prod_{i=1}^{n} f(x_i|\theta, \sigma^2) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}\right)
\end{aligned}
$$

- 따라서 베이즈 정리에 의한 θ의 사후분포는 다음과 같다.

$$\pi(\theta|\boldsymbol{X}) \propto \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(\theta - m)^2}{2s^2}\right) \times \frac{1}{(2\pi\sigma^2)\backslash/2}^{n} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \theta)^2}{2\sigma^2}\right)$$

- 이 식을 정리하면

$$\theta \,|\, \boldsymbol{X} \;\sim\; N\left(\frac{\frac{\bar{X}}{\sigma^2/n} + \frac{m}{s^2}}{\frac{1}{\sigma^2/n} + \frac{1}{s^2}}, \left(\frac{1}{\sigma^2/n} + \frac{1}{s^2}\right)^{-1}\right)$$

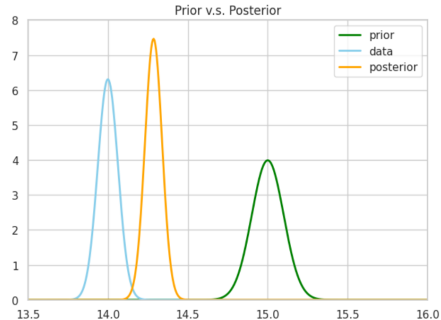
- 사후분포의 평균 $\frac{\frac{\bar{X}}{\sigma^2} + \frac{m}{s^2}}{\frac{1}{\sigma^2} + \frac{1}{s^2}}$은 데이터의 평균 ($\bar{X}$) 과 사전분포의 평균($m$)의 가중평균으로 볼 수 있다.

정규분포 예제- 데이터 이용

배터리를 만드는 어떤 공장이 있다. 이 공장에서 하루동안 생산하는 배터리의 불량률(%)은 정규분포를 근사적으로 따르는데, 분산은 0.04이고 평균은 알려져있지 않다. 그런데 몇 년간의 자료를 토대로 보았을 때 불량률의 평균은 평균이 15%, 분산이 0.01인 정규분포를 따르는 것으로 확인되었다. 만약 10일 동안 공장을 가동하여 불량률의 10일동안의 평균이 14%로 관측되었다면, 이를 통하여 사후분포를 구하고 불량률의 베이즈 추정을 구해보자.

- μ에 대한 사전분포는 $N(m = 15, s^2 = 0.01)$이다.
- 데이터는 분산이 $\sigma^2 = 0.04$이고 평균이 θ인 정규분포 $N(\theta, \sigma^2 = 0.04)$를 따른다.
- 데이터의 평균 $\bar{X} = 14$이므로 앞에서 구한 θ의 사후분포식에 대입하면 $N(14.29, 0.053^2)$이다.

- θ에 대한 사전믿음(정규분포를 따른다는)에 데이터 $\bar{X} = 14$의 정보가 더해져 θ에 대한 사후분포가 만들어짐을 확인할 수 있다.
- 사전분포의 평균과 데이터의 평균 사이의 값을 평균으로 갖는 사후분포이다.
- 또한 사후분포의 분산이 사전분포의 분산보다 작아졌음을 알 수 있는데, 이는 θ가 사후분포의 평균 근방의 값을 가질 것에 대한 믿음이 더 커진 것이라 생각할 수 있다.



사전분포의 종류

- 사전분포는 데이터를 보기전에 모수 θ의 불확실성을 나타내는 확률분포이다.
- 켤레사전분포 (Conjugate prior): 사전분포와 사후분포가 같은 분포족 (distribution family)에 속하게 되는 사전분포를 말한다.

  예) 데이터가 정규분포를 따를때, 평균의 사전분포도 정규분포를 따른다고 할 경우, 평균의 사후 분포도 정규분포를 따름
- Informative prior (subjective prior) v.s. non-informative prior (diffuse prior, objective prior)

- 모수에 대해 구체적 정보를 주는 경우와 일반적 정보 또는 아예 정보가 없는 경우로 나눠서 생각할 수 있다.

  예) $\mu \sim N(1, 0.1^2)$ v.s. $\mu \propto 1$.

- Proper prior v.s. Improper prior

  사전분포가 잘 정의된 확률분포일때 (proper prior)와, 잘 정의되지 않은 확률분포 (improper prior), 즉, 적분 또는 합이 유한하지 않을때로 나누어 생각할 수 있다.

- 예를들어, $X \sim N(\mu, \sigma^2)$ 인 자료가 있다고 하자. 이때 분산은 알려져 있다고 가정한다. 평균에 대한 특별한 정보(믿음)가 없는경우 평균이 어떤 값을 가지던지 동일한 정도의 정보를 주도록 하고 싶으면 μ 가 균일분포를 따른다고 가정할 수 있다. 다만 평균의 범위가 $(-\infty, \infty)$이므로 이러한 분포는 적분가능하지 않다.

- 사전분포가 improper prior라도 사후분포가 proper prior가 될수 있다. 다만 항상 되는것은 아니기 때문에 확인이 필요하다.

Markov Chain Monte Carlo

사후분포 샘플링 접근법

- 베이즈 추론을 요약하면
  - 데이터의 분포에 관한 모형 (Likelihood, $L(\theta|X)$)과 모형을 정하는 모수 (parameter)의 사전분포 (prior distribution, $\pi(\theta)$)를 이용하여 모수의 사후분포(Posterior probability, $\pi(\theta|X)$)을 구하고 이를 이용하여 모수에 관한 추론을 하는 방법이다.
- 이때, Likelihood가 복잡하거나, 데이터가 아주 큰 경우, 일반적으로 사후확률을 이론적으로 구하기 어렵다.
- 여기에서는 사후분포로부터의 샘플들을 구하여 사후분포를 근사적으로 구하는 Monte Carlo 방법 , Markov Chain Monte Carlo 방법, Variational Inference 등을 소개한다.

몬테카를로 방법(Monte Carlo method)

- 어떤 값 (또는 함수값)을 근사적으로 계산하는데 있어 확률분포로부터 생성한 무작위 샘플(표본)들을 이용하는 방법을 몬테카를로 방법이라고 한다.
- 만약 사후분포로부터의 난수생성이 가능하다면 사후분포로부터 생성한 무작위 샘플(표본)을 이용하여 사후 평균, 사후 분산, 사후 중앙값 등 사후분포의 특성을 추정 할 수 있다.
- 몬테카를로 방법은 사후분포로부터의 난수생성이 쉽지 않으면 효용도가 떨어진다는 문제가 있다.

Markov Chain Monte Carlo(MCMC)방법

- Markov Chain Monte Carlo (MCMC)는 점근분포(limiting distribution)가 우리가 원하는 사후분포가 되는 마코프 체인을 만들어 체인이 충분히 진행 되었을 때 체인들의 값들을 사후분포의 샘플이라고 생각하고 이를 이용하여 추론하는 방법이다
- 깁스 샘플링 (Gibbs sampling) 또는 메트로폴리스-헤이스팅스 샘플링 (Metropolis-Hastings sampling), 해밀턴 몬테 카를로(Hamiltonian Monte Carlo) 등이 있다.
- MCMC방법은 복잡한 모형인 경우 또는 데이터가 큰 경우 계산속도가 느린 단점이 있다.

마코프 체인

- 마코프 체인**(Markov Chain)**이란 여러 가능한 상태(state) 사이에서, 어느 한 상태로부터 다른 상태로의 전이(transition)를 겪는 수학적 시스템을 뜻한다.
- X_t를 $\pi_t(\cdot)$를 분포로 갖는 시각 t에서의 상태 벡터라고 하자.
- 무기억성(memorylessness):
  다음의 상태(X_{t+1})는 오로지 현재의 상태(X_t)에만 의존하며 그 이전에 일어난 일련의 상태와는 무관하다.
- 이런 성질을 마코프 성질(Markov Property)이라고 한다.
- 마코프 체인(Markov Chain)은 마코프 성질을 가지는 확률변수 $X_1, X_2, \cdots$들의 수열을 뜻한다.

- 즉, 주어진 현재의 상태에 대하여 미래와 과거의 상태는 독립이다.

$$P(X_{n+1} = x | X_n = x_n, \cdots, X_1 = x_1) = P(X_{n+1} = x | X_n = x_n)$$

- 전이행렬(transition matrix)또는 전이함수를 이용하여 현재의 상태로부터 그 다음, 다다음 상태의 확률분포를 계속하여 계산하는 것이 가능하다.
- 이러한 과정을 반복하다 보면 특정 조건 하에서 현재 상태(state)의 확률분포가 그 전 상태의 확률분포와 같아지는 때가 온다.
- 이렇게 평형상태에 도달한 확률분포를 정상분포(stationary distribution) 또는 점근분포(limiting distribution) 라고 하며, 이 분포는 초기값에 의존하지 않는다.

MCMC의 원리

- STEP 1 : 점근분포가 $\pi(x)$인 마코프 체인 $X_0, \cdots, X_t$를 만들어 X_t들을 저장한다.
- STEP 2 : t를 충분히 크게 하여 X_t가 $\pi(x)$를 따른다고 가정하는데 무리가 없으면, 이후에 생성되는 $X_{t+1}, \cdots X_{t+m}$을 저장한다.
- 이렇게 생성된 $X_{t+1}, \cdots X_{t+m}$를 $\pi(x)$를 따르는 난수로 본다.
- 수렴한 분포로부터의 난수 생성이라고 할 수 있기 위해서는 STEP 1에서의 t를 충분히 크게 해주어야 하는데, 이 과정은 결과가 초기값에 영향을 받지 않게 하는 과정으로 burn-in period라고 한다.
- 그렇다면 어떻게 이러한 마코프 체인을 만들 수 있을까?

깁스 샘플링

- p차원의 모수의 사후분포에 대한 문제에서 p개의 1차원 모수의 조건부 사후분포로 나눠서 마코프 체인을 만들어서 난수를 생성하는 방법이다.
- 이 때, 1차원 모수의 조건부 사후분포는 난수를 생성하기 쉬운 분포(아는 분포)이다.
- 예를 들어 데이터 $Y_1, \cdots, Y_n$이 $N(\mu, \sigma^2)$을 따른다고 했을때 관심모수는 $\Theta = (\mu, \sigma^2)^T$이므로 2차원 모수로 볼 수 있고, 사후확률분포는 $\pi(\Theta|Y_1, \cdots, Y_n) = \pi(\mu, \sigma^2|Y_1, \cdots, Y_n)$이다.
- 이 경우 1차원 모수의 조건부 사후 분포는 $\pi(\mu|\sigma^2, Y_1, \cdots, Y_n)$, $\pi(\sigma^2|\mu, Y_1, \cdots, Y_n)$ 이다.

알고리즘

1. p차원 모수 벡터 $\Theta = (\theta_1, \cdots, \theta_p)^T$에 대하여 초기값 $\Theta^{(0)} = (\theta_1^{(0)}, \cdots, \theta_p^{(0)})^T$를 설정한다.

   $t = 1$ 부터 T까지 다음을 반복한다.

2. $\theta_1^{(t)} \sim \pi(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, Data)$

   $\theta_2^{(t)} \sim \pi(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, Data)$

   $\cdots$

   $\theta_p^{(t)} \sim \pi(\theta_p | \theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{p-1}^{(t)}, Data)$

- 이 알고리즘은 다음과 같은 확률벡터열을 만들어낸다.

$$\Theta^{(1)} = (\theta_1^{(1)}, \cdots, \theta_p^{(1)})'$$
$$\Theta^{(2)} = (\theta_1^{(2)}, \cdots, \theta_p^{(2)})'$$
$$\cdots$$
$$\Theta^{(T)} = (\theta_1^{(T)}, \cdots, \theta_p^{(T)})'$$

- 이 확률벡터열에서, $\Theta^{(t)}$는 오로지 $\Theta^{(t-1)}$을 통해서만 $\Theta^{(0)}, \cdots, \Theta^{(t-1)}$에 의존한다. 즉, $\Theta^{(t)}$는 주어진 $\Theta^{(t-1)}$에 대하여 $\Theta^{(0)}, \cdots, \Theta^{(t-2)}$에 조건부 독립이 된다. 즉, 마코프 성질을 만족시킨다.
- 이때 마코프 체인의 시작점 $\Theta^{(0)}$는 적당히 값을 정한다.

예시- 정규분포

- 데이터가 정규분포 $N(\mu, \sigma^2)$를 따를 때의 μ와 σ^2에 대해 적절한 사전분포를 이용하여 깁스샘플링을 통한 베이즈 추정을 해보자
- 예를 들어, 생산된 제품들의 일별 정상품 비율은 정규분포를 따른다고 한다. 30일을 무작위로 골라 정상품 비율에 대한 평균과 분산을 베이즈 방법으로 추정하기위해 깁스샘플링을 이용한다.
- 데이터는 $X_1, \cdots, X_{n=30} \sim$ *i.i.d* $N(\mu, \sigma^2)$이고
- μ의 사전분포로는 실수 위에서의 균일분포(improper prior), σ^2의 사전분포로는 Jeffreys' prior인 $1/\sigma^2$을 가정한다.
- 즉 $\pi(\mu) \propto 1$, $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$

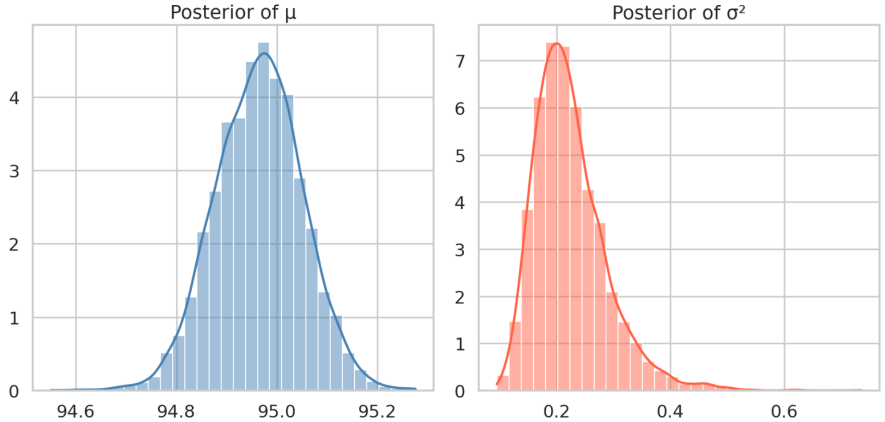
- 베이즈 정리에 의하여 다음과 같은 조건부 확률분포를 구 할 수 있다.

$$
\begin{aligned}
\mu \,|\, \boldsymbol{X}, \sigma^2 &\sim N\left(\bar{X}, \frac{\sigma^2}{n}\right) \\
\sigma^2 \,|\, \boldsymbol{X}, \mu &\sim \text{IGamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right)
\end{aligned}
$$

  - IGamma는 역감마분포로 확률밀도함수는 $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\beta/x\right)$ 이다. 여기서 α는 shape parameter, β는 scale parameter라고 한다.
- 위에서 구한 조건부 확률분포를 이용하여 깁스샘플링을 진행한다.

정규분포 깁스 샘플링 모의실험

- 평균이 95% , 표준편차가 0.5%인 정규분포로부터 30개의 데이터를 임의로 만들어서 데이터 인것 처럼 이용해서 10000개의 체인을 생성하고 5000개를 버린후 나머지로 μ, σ^2의 사후분포를 추정해보자.



깁스샘플링의 변형

- 1차원 모수의 조건부 사후분포중에 난수생성이 쉽지 않은 경우도 있다.
- 이 경우 해당 조건부 사후분포에서의 난수 생성은 proposal density를 이용하는 Metropolis-Hastings (M-H) 알고리즘을 적용할 수 있다.
- 이러한 MCMC 알고리즘은 M-H algorithm within Gibbs sampling이라고 한다.
- M-H 알고리즘은 다음 슬라이드에서 소개한다.

메트로폴리스-헤이스팅스 샘플링

- M-H 알고리즘에서는 각 시각 t에 대하여, 마코프 체인의 다음 상태인 X_{t+1}를 다음과 같은 과정을 통하여 결정한다.

(1) 제안분포(proposal distribution) $q(\cdot|X_t)$ 로부터 후보(candidate point)가 되는 샘플 Y을 뽑는다.
  - 제안분포는 현재의 상태인 X_t에 의존할 수 있다.
  - 예를 들어, $q(\cdot|X_t)$은 평균이 X_t이고 고정된 공분산행렬을 가지는 다변량 정규분포일 수 있다.

(2) 각 시각 t에 대하여, 다음과 같은 Acceptance ratio을 계산하여 마코프 체인의 다음 상태인 X_{t+1}을 결정한다.

$$\alpha(X, Y) = \min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right)$$

- 즉, 후보 Y는 $\alpha(X_t, Y)$의 확률로 다음 상태(X_{t+1})로 받아들여질지 아닐지 결정된다.
- 위의 식은 마코프 체인의 수렴을 위한 조건중 하나인 정상분포의 존재성을 위한 Detailed Balance 조건($\pi(X|Y)\pi(Y) = \pi(Y|X)\pi(X)$)으로부터 얻어진 식이다.
- 제안분포와 관련된 확률식 $\pi(Y|X) = q(Y|X)\alpha(X, Y)$을 위의 Detailed Balance 조건식에 대입해 $\frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}$항을 얻을 수 있다.

알고리즘

1. 초기값 $\theta^{(0)}$을 설정한다. 예를들어 $\theta^{(0)} = 1$.

   $t = 1$부터 T까지 다음을 반복한다.
2. $q(\theta|\theta^{(t-1)})$로부터 체인의 다음상태에 대한 후보로 $\tilde{\theta}$를 생성한다.
3. acceptance ratio α를 계산한다.
4. Unif(0,1)을 따르는 r을 생성한다.
5. $r < \alpha$(또는 $log(r) < log(\alpha)$)이면 $\theta^{(t)} = \tilde{\theta}$, 아니면 $\theta^{(t)} = \theta^{(t-1)}$로 둔다.

M-H 알고리즘의 장단점

- 원래 분포인 $\pi(\cdot)$을 정확히 몰라도 쓸 수 있다 . 즉, 정확한 확률분포가 아니라 그에 비례하는 비정규화 분포(un-normalized distribution)만 알아도 알고리즘을 사용할 수 있다.
- 목표로 하는 $\pi(\cdot)$의 분포를 따르는 X_t로 수렴하는데까지 시간이 오래 걸릴 수 있다.

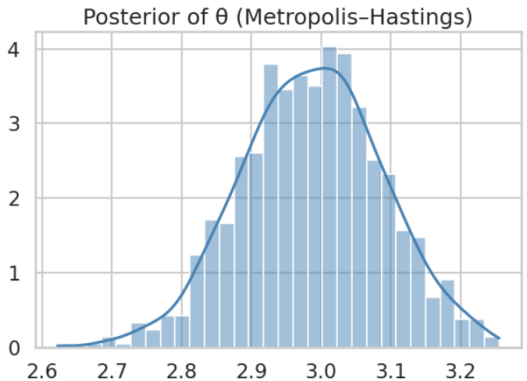
예시 - **Normal-Cauchy** 모델

- $Y_1, \cdots, Y_n \sim$ *i.i.d* $N(\theta, 1)$인 데이터를 생각하자.
- θ의 사전확률로 $\pi_0(\theta) = \frac{1}{\pi(1+\theta^2)}$를 가정하자. 첫번째 $\pi_0(\cdot)$는 확률밀도함수, 두번째 π는 원주율.
- 베이즈 정리에 의해 θ의 사후확률은 다음과 같다.

$$\begin{aligned} \pi(\theta|Y_1, \cdots, Y_n) &\propto \exp\left(-\frac{\sum_{i=1}^n (Y_i - \theta)^2}{2}\right) \times \frac{1}{1+\theta^2} \\ &\propto \exp\left(-\frac{n(\theta - \bar{Y})^2}{2}\right) \times \frac{1}{1+\theta^2} \end{aligned}$$

- 이 사후분포의 형태 (θ의 함수로서)는 일반적인 꼴 (아는 분포)이 아니므로, M-H 알고리즘을 적용한다. 제안분포로 예를들어 $q(\theta|\theta^*) = \frac{1}{\sqrt{2\pi n^{-1}}} e^{-\frac{(\theta-\bar{Y})^2}{2/n}}$, i.e. $N(\bar{Y}, \frac{1}{n})$를 사용할 수 있다.

- $Y_1, \cdots, Y_{100}$의 데이터가 있다고 하자. (실제로는 $Y \sim N(3,1)$에서 생성)
- M-H 알고리즘을 이용하여 $\theta^{(t)}$를 10,000개 생성한 후, 앞의 9000개를 burn-in period 로 보고 버리고 나머지 1000개를 가지고 사후분포를 추정해보자.



해밀턴 몬테 카를로**(HMC)**

- 사후분포의 형태에 따라 깁스 샘플링이나 M-H 샘플링으로 생성된 마코프 체인 X_n이 타겟 분포(점근분포) $\pi(x)$를 충분히 대표하지 못할 수 있다 (mixing is slow). 예- 상관계수가 1에 가까운 이변량 정규분포의 평균 추정
- HMC는 해밀턴 동역학 (Hamiltonian Mechanics) 아이디어를 사용해서 생성된 마코프 체인이 타겟 분포를 충분히 대표하도록 하는 방법이다.

해밀턴 동역학

- 해밀턴 동역학에서 입자들의 위치(position)를 θ, 운동량(momentum)을 η라고 했을 때 입자의 총 에너지(Hamiltonian, $H(\theta, \eta)$)를 위치에너지(Potential energy, $U(\theta)$)와 운동에너지(Kinetic energy, $K(\eta)$)의 합으로 표현한다. $H(\theta, \eta) = U(\theta) + K(\eta)$
- 입자의 에너지가 낮은 곳은 (H가 낮은 곳)에서는 입자들이 거의 정지하거나 자주 머물면서(U가 작고) 운동량의 에너지도 작은편(K가 작은)이고, 에너지가 높은 곳 (H가 높은곳)에서는 자주 머물지 않거나 (U가 크고) 또는 운동량의 에너지가 높은 (K가 큰)경우라고 한다면, 입자가 특정 위치와 운동량이 가지는 확률 (Joint probability)을 다음과 같이 해석할 수 있다. $\pi(\theta, \eta) \propto e^{-H(\theta,\eta)}$.

HMC 아이디어

- $\pi(\theta, \eta) \propto e^{-H(\theta,\eta)} = e^{-U(\theta)-K(\eta)}$에서 입자의 위치 θ를 우리가 추정하고자 하는 모수 θ, $\pi(\theta)$를 사후분포라고 생각하면 결합확률분포 ($\pi(\theta, \eta)$)를 점근 분포(타겟 분포)로 가지는 마코프 체인 $(\theta^{(r)}, \eta^{(r)})$을 생성하면 $\theta^{(r)}$들은 $\pi(\theta)$들로부터 나온 샘플이라고 할 수 있다.
- 입자들의 위치(θ)와 운동량η의 시간에 따른 변화를 표현하는 미분방정식 (Hamiltonian's differential equations)인 $\frac{d\theta}{dt} = \frac{\partial H}{\partial \eta}$, $\frac{d\eta}{dt} = -\frac{\partial H}{\partial \theta}$로부터 $(\theta^{(r)}, \eta^{(r)})$에서 $(\theta^{(r+1)}, \eta^{(r+1)})$로 이동하는 방법을 찾을 수 있다.
- 한편, 운동량(η) = 질량(m) $\times$ 속도(v), 운동에너지 $K = \frac{1}{2}mv^2$으로부터 $K(\eta) = \frac{1}{2m}\eta^2$이 되므로 η의 분포 $\pi(\eta) \propto e^{-K(\eta)}$는 분산을 m으로 갖는 정규분포 $N(0, m)$라고 볼 수 있어서 다루기 쉬운 분포이고, θ의 사후분포에 관심을 갖는 측면에서 η는 보조변수(auxiliary variable)로 본다.

- *m*의 조정을 통해 (θ, η)가 $\pi(\theta, \eta)$에 따라 (θ, η)가 있을 수 있는 공간(parameter space, phase space)에서 잘 움직이도록 (well mixing) 할 수 있다.
- 요약하자면, HMC는 관심 모수 θ를 입자의 위치로 보고 $\pi(\theta)$에 따라 θ가 움직인다고 할때 물리학적 구조를 차용하여 입자의 물리적 상태를 (θ, η)로 표현해서 $\pi(\theta, \eta)$에 따라 움직이게 함으로써 $\pi(\theta)$의 정보를 얻는것이다.
- 마코프 체인 생성은 해당 미분방정식을 수치적으로 풀어서 진행하게 되며, 알려진 방법으로 Leapfrog알고리즘 등이 있다.
- Leapfrog알고리즘을 향상시킨 NUTS(No-U-Turn sampler)방법이 개발되었고, 베이지안 분석을 수행할때 NUTS기반 사후분포 추출 소프트웨어인 STAN이 가장 널리 사용되며, STAN을 파이썬에서 사용할 수 있게하는 모듈로 pyMC, CmdStanPy 가 있다.

Variational Inference

변분추론

- 변분추론(Variational Inference)은 또 다른 근사 베이지안 방법(Approximate Bayesian method)으로 사후분포에 가까우면서, 샘플링이 쉬운 분포를 찾아 추론을 하는것이며, 일반적으로 MCMC보다 속도가 빠르다.
- 사후분포가 복잡한 모형인 경우 고려할 수 있다.
- 베이지안 분석에서 시작되었지만, MCMC의 경우는 분포로부터 표본을 샘플링하는 기법으로, 변분추론의 경우는 분포를 근사시키는 기법으로 사용되고 있다.

모형 가정

- 변분추론 설명을 위해서 다음과 같은 가정을 한다.
- 데이터: $\boldsymbol{X} = (X_1, \cdots, X_n)$
- 은닉변수(잠재변수): $\boldsymbol{Z} = (Z_1, \cdots, Z_m)$
- 추가 모수: α
- 목적: 사후분포 $p(\boldsymbol{Z}|\boldsymbol{X}, \alpha)$와 가까우면서 다루기 쉬운 분포(근사분포)를 찾아서 $\boldsymbol{Z}$를 생성하거나 사후분포의 특성값들을 근사적으로 구한다.

근사분포 찾기

- $p(\boldsymbol{Z}|\boldsymbol{X}, \alpha)$와 가까운 $q(\boldsymbol{Z}|\nu)$를 찾는다고 하자.
- 이때, $q(\boldsymbol{Z}|\boldsymbol{\nu})$는 ν에 따라 움직이는 분포들의 모임 (클래스 $\mathcal{Q}$)에 속한다고 하면, 이러한 분포들 중 $p(\boldsymbol{Z}|\boldsymbol{X}, \alpha)$에 가장 가깝도록 하는 ν를 찾는 문제로 볼수도 있다.
- 여기서 ν는 변분 모수 (variational parameter), $\mathcal{Q}$를 변분분포족(variational family, 변분분포 집)라 부른다.
- 두 분포가 가깝다는 기준, 즉 분포사이의 "가까움"을 나타내는 기준이 필요하다. 이를 위해 KL divergence를 소개한다.

쿨백-라이블러 발산 **(Kullback-Leibler Divergence)**

- 정보이론 (Information theory)에서 온 개념
- 두 분포사이의 "가까움"을 나타내는 값

$$
\begin{aligned}
KL(q \,\|\, p) &= E_q \left(\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} \right) \\
&= E_q \left(\log q(\boldsymbol{Z}) - \log p(\boldsymbol{Z}|\boldsymbol{X}) \right) \\
&= \sum_{\boldsymbol{z}} \log \left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{X})} \right) q(\boldsymbol{z}) \quad \text{(discrete)} \\
&= \int log \left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{X})} \right) q(\boldsymbol{z}) d\boldsymbol{z} \quad \text{(continuous)}
\end{aligned}
$$

- $KL(q \,\|\, p) \geq 0$. 만약 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$라면 $KL(q \,\|\, p) = 0$.

근사분포 찾는 문제는 분포들 사이의 가까움을 나타내는 K-L Divergence가 가장 작은 q를 찾는 문제로 생각한다. 즉,

$$q^* = \text{argmin}_{q \in \mathcal{Q}} KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X}))$$

- $KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X}))$ 계산을 위해서는 $\log(p(\boldsymbol{Z}|\boldsymbol{X})) = \log(p(\boldsymbol{Z}, \boldsymbol{X})/p(\boldsymbol{X}))$를 계산해야 하는데, 일반적으로 $\log(p(\boldsymbol{X}))$의 계산이 복잡하다.
- 따라서 $KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X}))$대신 좀 더 계산이 쉬운 다른 값(ELBO)을 이용한다.

Evidence Lower Bound(ELBO)

- $KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X})) = -ELBO(q) + \log p(\boldsymbol{X})$,
- $ELBO(q) = E_q(\log p(\boldsymbol{Z}, \boldsymbol{X})) - E_q(\log q(\boldsymbol{Z}))$ 임을 보일수 있다.
- $\log p(\boldsymbol{X})$는 q에 대해서 상수이므로 KL을 최소화 시키는 q를 찾는데 필요가 없다.
- 따라서

$$q^* = \text{argmin}_{q \in \mathcal{Q}} KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X})) = \text{argmax}_{q \in \mathcal{Q}} ELBO(q(\boldsymbol{Z}))$$

Evidence Lower Bound 이름의 유래

- $KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X})) \geq 0$으로부터
- $\log p(\boldsymbol{X}) \geq ELBO(q(\boldsymbol{Z}))$
- $\log p(\boldsymbol{X})$는 관측값의 Likelihood로 evidence라고도 부름.
- 따라서, $ELBO(q(\boldsymbol{Z})) = E_q(\log p(\boldsymbol{Z}, \boldsymbol{X})) - E_q(\log q(\boldsymbol{Z}))$는 evidence의 lower bound가 된다.

ELBO의 이해

$$
\begin{aligned}
ELBO(q) &= E_q\left(\log p(\boldsymbol{Z}, \boldsymbol{X})\right) - E_q\left(\log q(\boldsymbol{Z})\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) + E_q(\log p(\boldsymbol{Z})) - E_q\left(\log q(\boldsymbol{Z})\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) - E_q\left(\log(q(\boldsymbol{Z})/p(\boldsymbol{Z}))\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) - KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}))
\end{aligned}
$$

- 마지막 식의 첫번째 항의 $\log p(\boldsymbol{X}|\boldsymbol{Z})$는 잠재변수 $\boldsymbol{Z}$가 주어졌을때의 관측값 $\boldsymbol{X}$의 확률(log scale)로 $\boldsymbol{Z}$의 log-likelihood로 볼수 있으므로 첫번째 항은 $\boldsymbol{Z}$의 log-likelihood의 기대값으로 볼수 있다.
- 따라서 ELBO를 최대화 하는것은 $\boldsymbol{Z}$값이 $p(\boldsymbol{X}|\boldsymbol{Z})$를 크게 하도록 하는 (Likelihood를 증가시키는 또는 데이터 $\boldsymbol{X}$를 더 잘 설명하는)는 $q(\boldsymbol{Z})$를 찾으려 하는것으로 볼 수 있다.

$$
\begin{aligned}
ELBO(q) &= E_q\left(\log p(\boldsymbol{Z}, \boldsymbol{X})\right) - E_q\left(\log q(\boldsymbol{Z})\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) + E_q(\log p(\boldsymbol{Z})) - E_q\left(\log q(\boldsymbol{Z})\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) - E_q\left(\log(q(\boldsymbol{Z})/p(\boldsymbol{Z}))\right) \\
&= E_q\left(\log p(\boldsymbol{X}|\boldsymbol{Z})\right) - KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}))
\end{aligned}
$$

- 마지막 식의 두번째 항은 $\boldsymbol{Z}$의 사전분포 $p(\boldsymbol{Z})$와 $q(\boldsymbol{Z})$사이의 KL 발산이다.
- 따라서, ELBO를 최대화 하는것은 사전분포 $p(\boldsymbol{Z})$와 가까운 $q(\boldsymbol{Z})$를 찾으려 하는것으로 볼 수 있다.
- $ELBO(q)$를 최대화 하는것은 likeihood와 prior사이에서 적절한 q를 찾게되는것이다.

ELBO 최대화를 위한 $\mathcal{Q}$

$$q^* = \text{argmin}_{q \in \mathcal{Q}} KL(q(\boldsymbol{Z}) \,\|\, p(\boldsymbol{Z}|\boldsymbol{X})) = \text{argmax}_{q \in \mathcal{Q}} ELBO(q(\boldsymbol{Z}))$$

- ELBO 최대화 문제에서 $q \in \mathcal{Q}$를 찾는것은 $\mathcal{Q}$를 어떤분포 집합 (probability distribution family)으로 놓느냐에 따라 계산 복잡도가 달라진다.
- 예를들어 mean-field variational family는 잠재변수가 서로 독립이면서 각기 다른 변분인자 (variational factor)에 의존하는것으로 가정한다. 즉, $q(\boldsymbol{z}) = \prod_{i=1}^{m} q_i(z_i)$.
- 이 경우 $ELBO(q)$를 최대화하는 $\{q_i^*\}$를 찾는 문제가 된다.

Variational family

- variational family는 데이터 $\boldsymbol{X}$에 의존하지 않는다.
- mean-field variational family보다 복잡한 family를 고려할수도 있으나 계산상의 복잡도가 커진다.
- 구체적으로 어떤 variational family (확률분포)를 고려할지는 문제에 따라 다르다.
- variational family가 정해지면 최대화 시키는 최적화 알고리즘을 상황에 맞게 적용한다.
- 주어진 데이터 $\boldsymbol{X}$와, variational family $\mathcal{Q}$가 정해져서 $\{q_i^*\}$를 찾으면 필요에 따라 $\{q_i^*\}$를 이용하여 Z_i를 생성할수 있다.
- 생성된 $\{Z_i\}$를 이용하여 데이터 $\boldsymbol{X}$와 유사한 $\boldsymbol{X}^*$를 생성할수도 있다 (generative model).

Mean-Field Variational Bayes (MFVB) case

- $\mathcal{Q}$가 mean-field variational family인 경우, 즉 $q(\boldsymbol{z}) = \prod_{i=1}^{m} q_i(z_i)$인 경우 ELBO를 최대화 하는 알고리즘으로 coordinate ascent 알고리즘이 있다. 여기서 coordinate은 각 q_i별로 따로 찾는것을 의미한다.
- 즉, 반복알고리즘으로 $q_j(z_j), j \neq i$가 주어졌을때 $q_i(z_i)$를 찾는것이다.

Coordinate Ascent for MFVB

- mean-field varational family가정하에서 $ELBO(q)$는 다음과 같이 표현되므로 Coordinate Ascent 알고리즘을 적용할 수 있다.

$$\begin{aligned} ELBO(q) &= E_q(\log p(\boldsymbol{Z}, \boldsymbol{X})) - E_q(\log q(\boldsymbol{Z})) \\ &= \int q_i(z_i) E_{-i} \log p(\boldsymbol{Z}, \boldsymbol{X}) dz_i - \sum_{j=1}^{m} E_{q_j} \log q_j(Z_j) \end{aligned}$$

- $E_{-i} \log p(\boldsymbol{Z}, \boldsymbol{X})$는 $q_i(z_i)$를 제외한 나머지 $q_j(z_j)$들로 계산한 기댓값이다.

- 이를 다시 정리하면 다음과 같다.

$$\begin{aligned} ELBO(q) &= \int q_i(z_i) \log \frac{\exp(E_{-i} \log p(\boldsymbol{Z}, \boldsymbol{X}))}{q_i(z_i)} - \sum_{j \neq i}^{m} E_{q_j} \log q_j(Z_j) \\ &= -KL(q_i \| \tilde{q}_i) + \text{ constant independent of } q_i \end{aligned}$$

- $\tilde{q}_i(z_i) \propto \exp(E_{-i} \log p(z_i, \boldsymbol{Z}_{-i}, \boldsymbol{X})) \propto \exp(E_{-i} \log p(z_i | \boldsymbol{Z}_{-i}, \boldsymbol{X}))$.

- 앞 슬라이드의 $ELBO(q)$를 보면 $q_j(z_j), j \neq i$가 주어졌을때 $ELBO(q)$를 최대화 되도록 (또는 $KL(q_i \| \tilde{q}_i)$를 최소화 되도록) $i = 1, \cdots, m$에 대해서 차례대로 $q_i(z_i)$를 찾아가는 문제로 볼 수 있다.
- 그런데 $KL(q_i \| \tilde{q}_i)$는 $q_i = \tilde{q}_i$일 때 0이 되므로 결국, $q_j(z_j), j \neq i$가 주어졌을때 다음과 같이 주어진 $q_i^*(z_i)$를 수렴할때까지 반복적으로 찾는 문제가 된다.

$$\begin{aligned} q_i^*(z_i) &\propto \exp\left(E_{-i}\left(\log p(z_i | \boldsymbol{Z}_{-i}, \boldsymbol{X}\right)\right) \\ &\propto \exp\left(E_{-i}\left(\log p(z_i, \boldsymbol{Z}_{-i}, \boldsymbol{X}\right)\right) \end{aligned}$$

MFVB 예제 **-** 정규분포

- 데이터 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$일때 $\pi(\mu, \sigma^2) = \pi(\mu)\pi(\sigma^2) \propto \frac{1}{\sigma^2}$과 같은 사전분포를 가정하자. 앞서 깁스샘플링 예제와 동일
- 이 경우 $Z_1 = \mu$, $Z_2 = \sigma^2$으로 생각한다.
- 앞선 깁스샘플링 예제로부터 μ,σ^2의 조건부 사후분포가 다음과 같음을 알수 있다.

$$\begin{aligned} \mu \,|\, \boldsymbol{X}, \sigma^2 &\sim N\left(\bar{X}, \frac{\sigma^2}{n}\right) \\ \sigma^2 \,|\, \boldsymbol{X}, \mu &\sim \text{IGamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right) \end{aligned}$$

- 이정보로 부터 먼저 $q(\sigma^2)$가 주어졌을때 $q^*(\mu)$를 계산해보자.

$$
\begin{aligned}
q^*(\mu) &\propto \exp\left(E_{q(\sigma^2)}\left(\log p(\mu|\sigma^2, \boldsymbol{X})\right)\right) \\
&\propto \exp\left(E_{q(\sigma^2)}\left(-\frac{1}{2}\log\sigma^2 - \frac{n(\mu-\bar{X})^2}{2}\frac{1}{\sigma^2}\right)\right) \\
&\propto \exp\left(-\frac{nE_{q(\sigma^2)}[1/\sigma^2]}{2}(\mu-\bar{X})^2\right) \\
&\sim N\left(\bar{X}, \frac{1}{nE_{q(\sigma^2)}[1/\sigma^2]}\right)
\end{aligned}
$$

- 다음은 $q(\mu)$가 주어졌을때 $q^*(\sigma^2)$를 계산해보자.

$$
\begin{aligned}
q^*(\sigma^2) &\propto \exp\left(E_{q(\mu)}\left(\log p(\sigma^2|\mu, \boldsymbol{X})\right)\right) \\
&\propto \exp\left(E_{q(\mu)}\left(-\left(\frac{n}{2}+1\right)\log\sigma^2 - \frac{1}{\sigma^2}\frac{\sum(X_i-\mu)^2}{2}\right)\right) \\
&\propto (\sigma^2)^{-\frac{n}{2}-1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum E_{q(\mu)}(X_i-\mu)^2}{2}\right)\right) \\
&\sim \text{IGamma}\left(\frac{n}{2}, \frac{\sum E_{q(\mu)}(X_i-\mu)^2}{2}\right)
\end{aligned}
$$

- $q^*(\mu)$, $q^*(\sigma^2)$의 분포가 각각 정규분포, 역감마분포인것을 위의 유도식들로 부터 구하였으므로, $q(\mu) = q^*(\mu) = N(\mu_q, \sigma_q^2)$, $q(\sigma) = q^*(\sigma^2) = \text{IGamma}(\alpha_q, \beta_q)$로 놓으면 각 모수들은 다음과 같이 계산된다.

$$
\begin{aligned}
\mu_q &= \bar{X} \\
\sigma_q^2 &= \frac{1}{nE_{q(\sigma^2)}[1/\sigma^2]} = \frac{\beta_q}{n\alpha_q} \\
\alpha_q &= \frac{n}{2} \\
\beta_q &= \frac{\sum E_{q(\mu)}(X_i - \mu)^2}{2} = \frac{n\sigma_q^2 + \sum(X_i - \mu_q)^2}{2}
\end{aligned}
$$

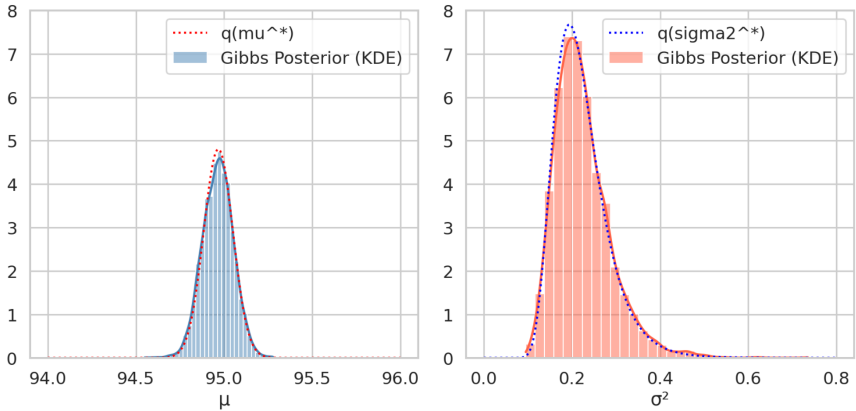
- 이 예제에서 μ_q, α_q는 바뀌지 않는다. 따라서 $q^*(\mu)$, $q^*(\sigma^2)$를 찾기 위해서는 β_q와 σ_q^2만 반복 알고리즘을 통해 찾으면 된다. 즉,
  (1) β_q의 초기값을 정한다.
  (2) 수렴할때까지 다음을 반복한다.

$$\sigma_q^2 \leftarrow \frac{\beta_q}{n\alpha_q}$$

$$\beta_q \leftarrow \frac{n\sigma_q^2 + \sum(X_i - \mu_q)^2}{2}$$

정규분포 깁스 샘플링 모의실험과 비교

- 깁스 샘플링 모의실험 예제와 동일한 데이터로 MFVB로 구한 $q^*(\mu)$, $q^*(\sigma^2)$와 깁스 샘플링으로 구한 μ와 σ^2의 사후분포를 비교해 보자.



The End