

STAT 431 — Applied Bayesian Analysis — Course Notes

Bayesian Computation: Gibbs Sampling and MCMC

Fall 2022

Sampling: Concepts, Notation, Facts

from posterior dist

mean, var - can't

so sample

to estimate

Suppose we want a (joint) sample of compute

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) \sim \text{some joint distribution } \mathcal{D}$$

where the joint distribution has a density

$$p(\theta_1, \dots, \theta_p)$$

We will say

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(S)}$$

is a **sample** from \mathcal{D} if

$$\boldsymbol{\theta}^{(s)} \sim \mathcal{D} \quad \text{for each } s$$

Notice: No need for independence — it could be a
dependent sample.

Notation:

$$\boldsymbol{\theta}^{(s)} = \left[\theta_1^{(s)}, \dots, \theta_p^{(s)} \right]$$

Fact:

$$\theta_j^{(1)}, \dots, \theta_j^{(S)}$$

is a sample from the marginal distribution of θ_j (under \mathcal{D})

Notation:

$$\boldsymbol{\theta}_{(-j)} = \boldsymbol{\theta} \text{ without } \theta_j$$

Then

$$p(\theta_j \mid \boldsymbol{\theta}_{(-j)})$$

is called the **full conditional density** for θ_j (corresponding to its **full conditional distribution**). \Rightarrow concept of Gibbs Sampling

Fact: If $\boldsymbol{\theta} \sim \mathcal{D}$ and we sample

$$\tilde{\theta}_1 \quad \text{from} \quad p(\cdot | \boldsymbol{\theta}_{(-1)})$$

(i.e. $\tilde{\theta}_1$ is sampled from the full conditional of θ_1), then

$$(\tilde{\theta}_1, \boldsymbol{\theta}_{(-1)}) \sim \mathcal{D}$$

Fact: If $\boldsymbol{\theta} \sim \mathcal{D}$ and we sample

$$\tilde{\theta}_1 \quad \text{from} \quad p(\cdot | \boldsymbol{\theta}_{(-1)})$$

(i.e. $\tilde{\theta}_1$ is sampled from the full conditional of θ_1), then

$$(\tilde{\theta}_1, \boldsymbol{\theta}_{(-1)}) \sim \mathcal{D}$$

Similarly for any element of $\boldsymbol{\theta}$: If we sample

$$\tilde{\theta}_j \quad \text{from} \quad p(\cdot | \boldsymbol{\theta}_{(-j)})$$

then

$$(\theta_1, \dots, \theta_{j-1}, \tilde{\theta}_j, \theta_{j+1}, \dots, \theta_p) \sim \mathcal{D}$$

Basic Gibbs Sampling

Based on the full conditionals ...

For simplicity, suppose the model has two parameters:

$$\boldsymbol{\theta} = (\theta_1, \theta_2)$$

Given data \mathbf{y} , the (joint) posterior density is

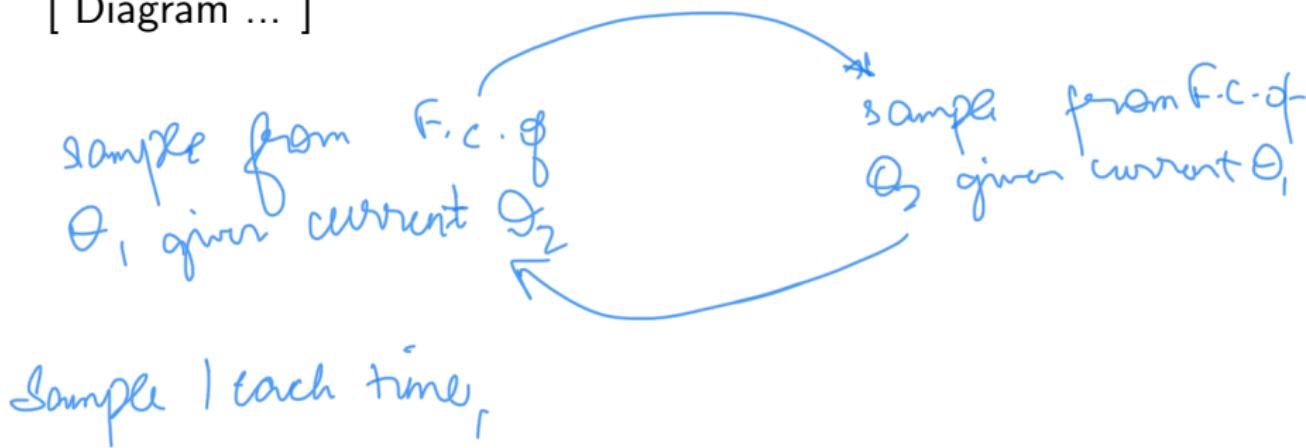
$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\theta_1, \theta_2 \mid \mathbf{y})$$

for which the full conditionals are

$$p(\theta_1 \mid \theta_2, \mathbf{y}) \quad \text{and} \quad p(\theta_2 \mid \theta_1, \mathbf{y})$$

Idea: Alternate between sampling from the full conditional for θ_1 and the full conditional for θ_2 (once each time), updating each value after sampling.

[Diagram ...]



Result: A sequence of **iterates**

$$\underbrace{\theta_1^{(1)}, \theta_2^{(1)}}_{\theta^{(1)}}, \underbrace{\theta_1^{(2)}, \theta_2^{(2)}}_{\theta^{(2)}}, \underbrace{\theta_1^{(3)}, \theta_2^{(3)}}_{\theta^{(3)}}, \dots$$

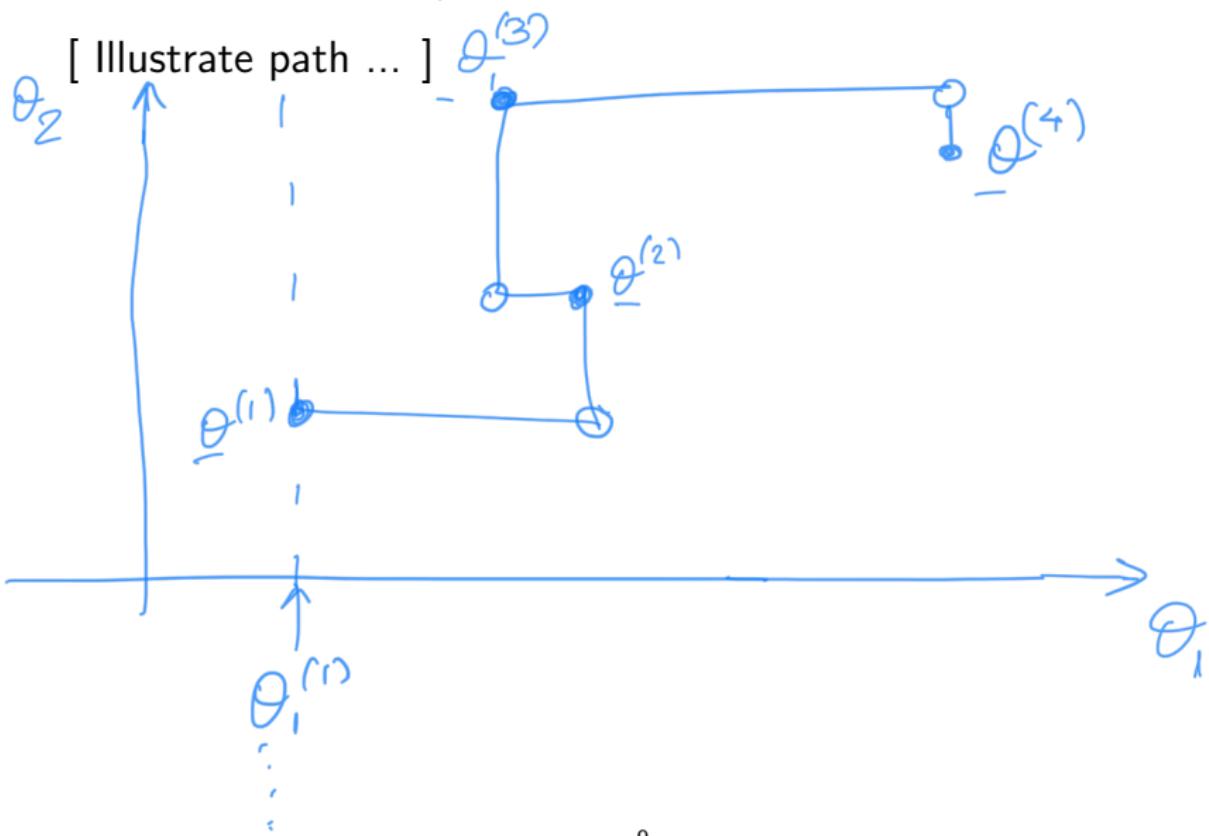
(The initial value $\theta_1^{(1)}$ may be chosen deterministically or at random — ideally it shouldn't matter.)

Note: These samples will generally be dependent because each is sampled based on the previous one.

Algorithm:

1. Choose initial value $\theta_1^{(1)}$ and sample $\theta_2^{(1)}$ from $p(\theta_2 \mid \theta_1^{(1)}, \mathbf{y})$
2. For $s = 2$ to S ,
 - 2.1 Sample $\theta_1^{(s)}$ from $p(\theta_1 \mid \theta_2^{(s-1)}, \mathbf{y})$
 - 2.2 Sample $\theta_2^{(s)}$ from $p(\theta_2 \mid \theta_1^{(s)}, \mathbf{y})$
3. Use iterates $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}$ for inference.

The iterates form a "path of samples":



Fact: If $\theta_1^{(1)}$ is drawn from its posterior marginal $p(\theta_1 | \mathbf{y})$, then the sequence of iterates

$$\theta^{(1)}, \theta^{(2)}, \dots \theta^{(S)}$$

is a (dependent) sample from the posterior.

Principle: Under certain conditions, regardless of the value of $\theta_1^{(1)}$, the iterates will converge in distribution to the posterior.

like CLT

So, “eventually” (for s large enough)

$$\theta^{(s)}, \dots \theta^{(S)}$$

will be *approximately* a (dependent) sample from the posterior.

Issues (addressed later):

- ▶ Where to start? (choosing $\theta_1^{(1)}$)
- ▶ How long until “close enough” to posterior?
- ▶ How many samples needed?
- ▶ How to detect problems?

Concern: Why would sampling from the full conditionals necessarily be any easier than direct sampling from the posterior?

Answer: You can often make the full conditionals easy to sample by using a special form of prior ...

Semi-Conjugacy

As an example, suppose

$$\underbrace{Y_1, \dots, Y_n}_{\mathbf{Y}} \mid \mu, \sigma^2 \sim iid \text{ Normal}(\mu, \sigma^2)$$

with both μ and σ^2 unknown.

Recall:

- ▶ For fixed σ^2 , the normal distribution is conjugate for μ .
- ▶ For fixed μ , the inverse gamma distribution is conjugate for σ^2 .

We call each of these distributions **semi-conjugate** (also called **conditionally conjugate** in BSM).

A possible **semi**-conjugate prior specification:

$$\begin{aligned}\mu &\sim \text{Normal}(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim \text{InvGamma}(\alpha, \beta)\end{aligned}\quad \left.\right\} \text{independent}$$

You can show that this prior is **NOT** (fully) conjugate.

(Indeed, μ and σ^2 turn out to be *dependent* under the corresponding posterior.)

However, semi-conjugacy makes each full conditional easy to sample, since each parameter has a conjugate prior when the other parameter is fixed.

Formally, for a given likelihood family, we will call a prior semi-conjugate for a parameter θ_j

if it would be conjugate when $\boldsymbol{\theta}_{(-j)}$ is held fixed:

For any \mathbf{y} , the prior full conditional

$$\pi(\theta_j \mid \boldsymbol{\theta}_{(-j)})$$

and the posterior full conditional

$$p(\theta_j \mid \boldsymbol{\theta}_{(-j)}, \mathbf{y})$$

are of the same distributional family.

for semi-conjugacy

Also, notice:

Assuming $p = 2$ parameters (for simplicity), we have

$$p(\theta_1 \mid \theta_2, \mathbf{y}) = \frac{p(\theta_1, \theta_2 \mid \mathbf{y})}{p(\theta_2 \mid \mathbf{y})} \underset{\text{in } \theta_1}{\propto} p(\theta_1, \theta_2 \mid \mathbf{y})$$

posterior

So the joint posterior density is actually a kernel of the full conditional for θ_1 . (Similarly for θ_2 .)

Thus, choosing a prior that makes this the kernel of an easily-sampled distribution will allow easy Gibbs sampling for θ_1 .

Example: Normal Sample, *Semi-Conjugate Prior*

$$\underbrace{Y_1, \dots, Y_n}_{\mathbf{Y}} \mid \mu, \sigma^2 \sim iid \text{ Normal}(\mu, \sigma^2)$$

$$\left. \begin{array}{l} \mu \sim \text{Normal}(\mu_0, \sigma_0^2) \\ \sigma^2 \sim \text{InvGamma}(\alpha, \beta) \end{array} \right\} \text{independent}$$

Recall: This prior is semi-conjugate ...

Getting the full conditional for μ is just like treating σ^2 as known — recall, under a $\text{Normal}(\mu_0, \sigma_0^2)$ prior,

$$\mu \mid \sigma^2, \mathbf{y} \sim \text{Normal}(\mu_1, 1/\tau_1^2)$$

where

$$\mu_1 = \frac{n\tau^2 \bar{y} + \tau_0^2 \mu_0}{n\tau^2 + \tau_0^2} \quad \tau_1^2 = n\tau^2 + \tau_0^2$$

with

$$\tau^2 = 1/\sigma^2 \quad \tau_0^2 = 1/\sigma_0^2$$

Getting the full conditional for σ^2 is just like treating μ as known — recall, under an $\text{InvGamma}(\alpha, \beta)$ prior,

$$\sigma^2 | \mu, \mathbf{y} \sim \text{InvGamma}(n/2 + \alpha, SSE/2 + \beta)$$

where

$$\begin{aligned} SSE &= \sum_i (y_i - \mu)^2 \\ &= (n - 1)s^2 + n(\bar{y} - \mu)^2 \end{aligned}$$

and s^2 is the usual sample variance.

The Gibbs sampler just alternates between sampling from these full conditionals.

We illustrate with Jevons's coin data ...

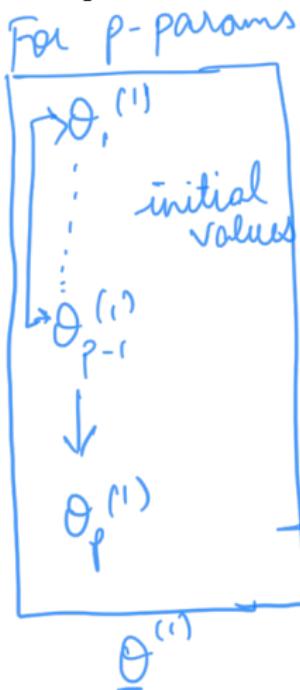
R Example 3.4:

Gibbs Sampler for Semi-Conjugate Prior
(Normal Sample)



Generalize: Gibbs Sampler for p Parameters

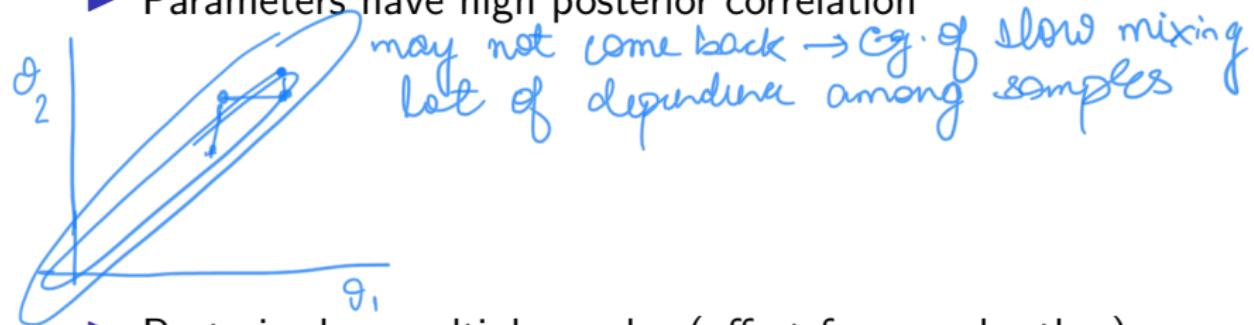
[Diagram of sampling ...]



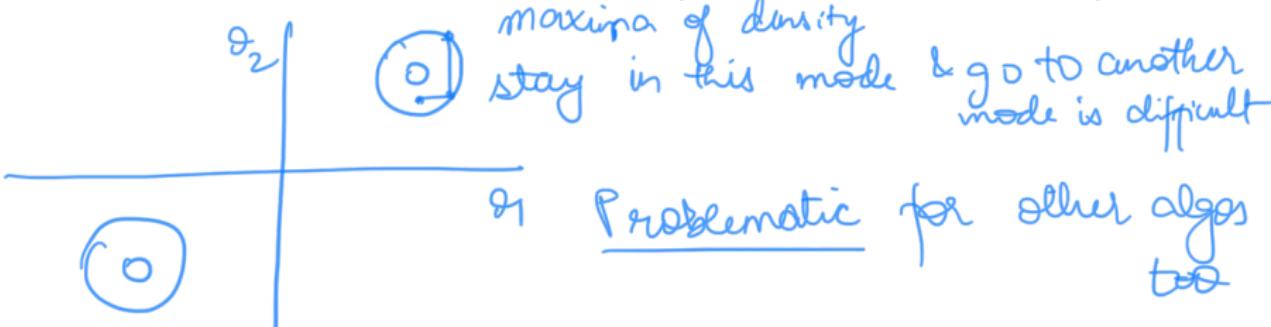
Difficult situations for Gibbs sampling:

(much higher than previous)

- ▶ Parameters have high posterior correlation



- ▶ Posterior has multiple modes (offset from each other)



Markov Chain Monte Carlo (MCMC)

A sequence of random variables

$$X_0, X_1, X_2, \dots$$

is a **Markov chain (MC)** if, for each $t \geq 2$, X_t is conditionally independent of

$$X_0, \dots X_{t-2}$$

given X_{t-1} .

That is,

$$f(x_t | x_{t-1}, \dots x_0) = f(x_t | x_{t-1}).$$

X_t is the **state** of the MC at time t .

The **transition kernel** is the conditional density

$$\textcolor{blue}{f} p(x_t \mid x_{t-1})$$

which determines how X_t can be generated based on X_{t-1} .

The kernel is **time-invariant** if it does not depend on t .
(Similarly for the MC.)

More generally, MCs may be sequences of random *vectors*.

A Gibbs sampler is a time-invariant Markov chain:

- ▶ $\theta^{(s)}$ is generated using only $\theta^{(s-1)}$
- ▶ the distributions used in the generation of $\theta^{(s)}$ do not depend on s (except through the value of $\theta^{(s-1)}$)

Under certain conditions, states of a time-invariant Markov chain converge in distribution to a unique distribution \mathcal{D} as $t \rightarrow \infty$:

$$X_t \xrightarrow[t \rightarrow \infty]{} \mathcal{D}$$

for (almost) any X_0 .

(The “certain conditions” are technical and often difficult to check.)

Under certain conditions, states of a time-invariant Markov chain converge in distribution to a unique distribution \mathcal{D} as $t \rightarrow \infty$:

$$X_t \xrightarrow[t \rightarrow \infty]{} \mathcal{D}$$

for (almost) any X_0 .

(The “certain conditions” are technical and often difficult to check.)

For a Gibbs sampler, the distribution \mathcal{D} is the posterior:

$$\boldsymbol{\theta}^{(s)} \xrightarrow[s \rightarrow \infty]{} \text{the posterior}$$

for (almost) any choice of $\boldsymbol{\theta}^{(1)}$.

Practical approach to running MCMC:

(multi mode \rightarrow identify problem)

(1) Choose several different **initial values** ($\theta^{(1)}$ s).

(Better if they are far apart.)

(2) For each initial value, run a separate **chain** for S **iterations**.

(3) **Monitor** the chains for:

- ▶ whether they seem to be converging to the same distribution
- ▶ how many iterations until convergence

Increase S if necessary.

(4) If converged, declare the first S_b iterates

$$\theta^{(1)}, \dots, \theta^{(S_b)}$$

warm-up or

of each chain to be a burn-in period.

Ignore the burn-in iterates, and use the rest for inference.

(5) Estimate the Monte Carlo error in your inferences.

Run more iterations until it is sufficiently small.

Note: Some samplers also need an initial period of
+ adaptation to find a good sampling scheme when
semi-conjugacy does not hold. Only the iterates after both
adaptation and burn-in should be used for inference.