S&DS 242/542: Theory of Statistics

Lecture 18: Bayesian Inference II

Bayesian inference

Given data $\mathbf{X} \sim f(\mathbf{x} \mid \theta)$ from a parametric model, Bayesian inference models θ as random, with a *prior distribution*

$$\Theta \sim f_\Theta(\theta)$$

The main object of interest is then the *posterior distribution*

$$f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$$

representing our updated belief about Θ, after having observed that the data is $\mathbf{X} = \mathbf{x}$.

This may be computed from

$$\underbrace{f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})}_{\text{posterior}} \propto \underbrace{f_{\mathbf{X}|\Theta}(\mathbf{x} \mid \theta)}_{\text{likelihood}} \underbrace{f_\Theta(\theta)}_{\text{prior}}$$

Bayesian point estimates and credible intervals

Either the mean or mode of the posterior distribution is commonly used as a Bayesian estimate $\hat{\theta}$ for θ.

The interval $I(\mathbf{X})$ from the lower-$\alpha/2$ to upper-$\alpha/2$ point of the posterior distribution forms a *Bayesian credible interval* with coverage $1-\alpha$. This ensures:

$$\mathbb{P}[\,\Theta \in I(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\,] = 1 - \alpha$$

where the probability is over the posterior distribution of Θ.

Example: Beta prior for a Bernoulli proportion

Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(p), with prior $P \sim \text{Beta}(\alpha, \beta)$. We computed the posterior distribution to be

$$P \sim \text{Beta}(S + \alpha,\, n - S + \beta), \qquad S = X_1 + \ldots + X_n$$

The posterior mean is

$$\hat{p} = \frac{S + \alpha}{n + \alpha + \beta},$$

a sample average as if we had observed — a priori — α heads and β tails before seeing our data. It is also a weighted average

$$\hat{p} = \underbrace{\frac{n}{n + \alpha + \beta}}_{\text{sample weight}} \cdot \underbrace{\frac{S}{n}}_{\text{sample mean}} + \underbrace{\frac{\alpha + \beta}{n + \alpha + \beta}}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}}$$

The lower-0.05 point to upper-0.05 point of $\text{Beta}(S + \alpha, n - S + \beta)$ forms a 90% Bayesian credible interval for p.

Example: Gamma prior for the Poisson model

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, with prior $\Lambda \sim \text{Gamma}(\alpha, \beta)$. We computed the posterior distribution to be

$$\Lambda \sim \text{Gamma}(S + \alpha,\ n + \beta), \qquad S = X_1 + \ldots + X_n$$

The posterior mean is

$$\hat{\lambda} = \frac{S + \alpha}{n + \beta},$$

again a sample average as if we had observed — a priori — β additional samples with sum α. It is also a weighted average

$$\hat{\lambda} = \underbrace{\frac{n}{\beta + n}}_{\text{sample weight}} \cdot \underbrace{\frac{S}{n}}_{\text{sample mean}} + \underbrace{\frac{\beta}{\beta + n}}_{\text{prior weight}} \cdot \underbrace{\frac{\alpha}{\beta}}_{\text{prior mean}}$$

The lower-0.05 to upper-0.05 points of $\text{Gamma}(S + \alpha, n + \beta)$ forms a 90% Bayesian credible interval for λ.

Example: Normal prior for a normal mean

Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1/\xi)$, where ξ is known and θ has prior $\Theta \sim \mathcal{N}(\mu_{\text{prior}}, 1/\xi_{\text{prior}})$. We computed the posterior distribution as

$$\Theta \sim \mathcal{N}(\mu_{\text{post}}, 1/\xi_{\text{post}})$$

where $\xi_{\text{post}} = n\xi + \xi_{\text{prior}}$, and the posterior mean takes a form

$$\mu_{\text{post}} = \frac{S + (\xi_{\text{prior}}/\xi)\mu_{\text{prior}}}{n + (\xi_{\text{prior}}/\xi)}, \qquad S = X_1 + \ldots + X_n$$

This is as if we had observed ξ_{prior}/ξ additional samples with mean μ_{prior}, and is again a weighted average

$$\mu_{\text{post}} = \underbrace{\frac{n}{n + (\xi_{\text{prior}}/\xi)}}_{\text{sample weight}} \cdot \underbrace{\frac{S}{n}}_{\text{sample mean}} + \underbrace{\frac{\xi_{\text{prior}}/\xi}{n + (\xi_{\text{prior}}/\xi)}}_{\text{prior weight}} \cdot \underbrace{\mu_{\text{prior}}}_{\text{prior mean}}$$

The interval $\mu_{\text{post}} \pm z^{(0.05)}\sqrt{\frac{1}{\xi_{\text{post}}}}$ is a 90% credible interval for θ.

Bayesian vs. frequentist coverage

Bayesian credible intervals vs. confidence intervals

A Bayesian credible interval $I(\mathbf{X})$ with coverage $1-\alpha$ is an interval which guarantees

$$\mathbb{P}[\,\Theta \in I(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\,] = 1 - \alpha$$

Here Θ is random, and $1-\alpha$ is the probability of $\Theta \in I(\mathbf{x})$ over the randomness of Θ, given the observed data $\mathbf{X} = \mathbf{x}$.

A frequentist confidence interval $I(\mathbf{X})$ with coverage $1-\alpha$ is an interval which guarantees

$$\mathbb{P}_\theta[\,\theta \in I(\mathbf{X})\,] = 1 - \alpha$$

The parameter θ is fixed, and $1-\alpha$ is the probability of $\theta \in I(\mathbf{X})$ over the randomness of $\mathbf{X} \sim f(\mathbf{x} \mid \theta)$.

Example: Intervals for a normal model

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$, where the variance is known and equal to 1.

In a frequentist analysis, we might estimate θ by its MLE $\hat{\theta} = \bar{X}$. Then $\text{Var}_\theta[\bar{X}] = \frac{1}{n}$, and a frequentist confidence interval for θ would be

$$\bar{X} \pm \frac{1}{\sqrt{n}} \cdot z^{(\alpha/2)}$$

This guarantees, for *any fixed value* of the true parameter θ,

$$\mathbb{P}_\theta \left[\theta \in \bar{X} \pm \frac{1}{\sqrt{n}} \cdot z^{(\alpha/2)} \right] = 1 - \alpha$$

where $\mathbb{P}_\theta$ is over the randomness of $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$.

Example: Intervals for a normal model

In a Bayesian analysis, suppose we choose the prior distribution $\Theta \sim \mathcal{N}(0, 1/\xi_{\mathrm{prior}})$. The posterior distribution is then

$$\Theta \sim \mathcal{N}(\mu_{\mathrm{post}}, 1/\xi_{\mathrm{post}})$$

where

$$\mu_{\mathrm{post}} = \frac{n}{n+\xi_{\mathrm{prior}}}\bar{X}, \qquad \xi_{\mathrm{post}} = n + \xi_{\mathrm{prior}}.$$

A Bayesian credible interval would be

$$\frac{n}{n+\xi_{\mathrm{prior}}}\bar{X} \pm \frac{1}{\sqrt{n+\xi_{\mathrm{prior}}}} \cdot z^{(\alpha/2)}$$

- This is centered at the posterior mean, which shrinks $\bar{X}$ towards the prior mean of 0.
- Its width is narrower than the frequentist interval.

Example: Intervals for a normal model

This posterior credible interval has the guarantee, for any realization of the data $\mathbf{x} = (x_1, \ldots, x_n)$ with mean $\bar{x}$,

$$\mathbb{P}\left[\Theta \in \frac{n}{n+\xi_{\text{prior}}}\bar{x} \pm \frac{1}{\sqrt{n+\xi_{\text{prior}}}} \cdot z^{(\alpha/2)} \;\middle|\; \mathbf{X} = \mathbf{x}\right] = 1-\alpha$$

where $\mathbb{P}$ is over the posterior distribution of Θ given $\mathbf{X} = \mathbf{x}$.

Q: Suppose there is a fixed true parameter θ. Does this Bayesian credible interval cover θ with probability $1-\alpha$, in the frequentist sense (over the randomness of the data $\mathbf{X}$)?

A: Not necessarily, and this coverage probability depends on the value of θ.

Example: Intervals for a normal model

The coverage probability is, in the frequentist sense,

$$
\begin{aligned}
&\mathbb{P}_\theta\left[\theta \in \frac{n}{n+\xi_{\text{prior}}}\bar{X} \pm \frac{1}{\sqrt{n+\xi_{\text{prior}}}}\cdot z^{(\alpha/2)}\right] \\
&= \mathbb{P}_\theta\left[\bar{X} \in \frac{n+\xi_{\text{prior}}}{n}\left(\theta \pm \frac{1}{\sqrt{n+\xi_{\text{prior}}}}\cdot z^{(\alpha/2)}\right)\right] \\
&= \mathbb{P}_\theta\left[\underbrace{\sqrt{n}(\bar{X}-\theta)}_{=Z} \in \frac{\xi_{\text{prior}}}{\sqrt{n}}\theta \pm \sqrt{\frac{n+\xi_{\text{prior}}}{n}}\cdot z^{(\alpha/2)}\right]
\end{aligned}
$$

Example: Intervals for a normal model

When $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$, $Z = \sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0, 1)$. So the coverage probability is

$$\mathbb{P}_{Z \sim \mathcal{N}(0,1)} \left[Z \in \frac{\xi_{\text{prior}}}{\sqrt{n}} \theta \pm \sqrt{\frac{n + \xi_{\text{prior}}}{n}} \cdot z^{(\alpha/2)} \right]$$

- If θ is far from 0 (the mean of our assumed prior distribution), then this coverage probability can be smaller than $1 - \alpha$, and it approaches 0 as $\theta \to \infty$.
- On the other hand, if $\theta = 0$, then this coverage probability is larger than $1 - \alpha$ because $\sqrt{(n + \xi_{\text{prior}})/n} > 1$.

So the interval has under-coverage when θ is far from the prior mean of 0, and over-coverage when θ is close to the prior mean.

The Bayesian coverage guarantee

To summarize, a Bayesian credible interval does not guarantee frequentist coverage for each fixed θ. However, it instead guarantees coverage in an average sense:

Since $\mathbb{P}[\,\Theta \in I(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\,] = 1 - \alpha$ for any possible realization of the data $\mathbf{X} = \mathbf{x}$, this guarantee holds also unconditionally, i.e.

$$\mathbb{P}[\,\Theta \in I(\mathbf{X})\,] = 1 - \alpha$$

This probability is over the joint distribution of Θ and $\mathbf{X}$.
The marginal distribution of Θ is the prior distribution $f_\Theta(\theta)$, so we may write this in turn as

$$\begin{aligned} 1 - \alpha = \mathbb{P}[\,\Theta \in I(\mathbf{X})\,] &= \int \mathbb{P}[\Theta \in I(\mathbf{X}) \mid \Theta = \theta]\, f_\Theta(\theta)\, d\theta \\ &= \int \mathbb{P}_\theta[\theta \in I(\mathbf{X})]\, f_\Theta(\theta)\, d\theta \end{aligned}$$

The Bayesian coverage guarantee

$$1 - \alpha = \int \mathbb{P}_\theta[\theta \in I(\mathbf{X})] \, f_\Theta(\theta) \, d\theta$$

Thus the frequentist coverage probability, when *averaged* over values of θ according to the prior, is guaranteed to be $1 - \alpha$.

- Suppose we conduct many experiments, where the parameter θ is the same in all experiments. If we compute a 90% frequentist confidence interval in each experiment, then roughly 90% of these intervals would cover θ. A similar guarantee may not hold for a 90% Bayesian credible interval.
- Suppose instead that θ is different in each experiment, and its distribution across experiments is correctly described by our Bayesian prior $f_\Theta(\theta)$. If we compute a 90% Bayesian credible interval in each experiment using this prior, then roughly 90% of these intervals would cover the corresponding values of θ.

Influence of the prior

Uninformative and improper priors

Sometimes we wish to use the Bayesian formalism, but to carry out an objective analysis without imposing prior knowledge.

This may be achieved by using an *uninformative prior* that minimizes its influence over the posterior mode or mean.

Example: Consider $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, with conjugate prior $\Lambda \sim \text{Gamma}(\alpha, \beta)$, posterior $\text{Gamma}(X_1 + \ldots + X_n + \alpha, n + \beta)$.

The posterior mean is

$$\hat{\lambda} = \frac{X_1 + \ldots + X_n + \alpha}{n + \beta}$$

which may be interpreted as observing — a priori — β values that sum to α. The prior is less informative for smaller α, β.

Uninformative and improper priors

Taking this idea to the limit, we may set $\alpha = \beta = 0$. The PDF of Gamma(α, β) is $f_\Lambda(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}$, so $\alpha = \beta = 0$ corresponds to

$$f_\Lambda(\lambda) \propto \lambda^{-1} \text{ for } \lambda > 0$$

This prior Gamma$(0, 0)$ is not an actual probability distribution, because $\int_0^\infty \lambda^{-1} d\lambda = \infty$ so that it is impossible to multiply by a proportionality constant to make the PDF integrate to 1. Such a prior is called an **improper prior**.

It is sometimes still possible to formally carry out Bayesian inference, using the rule posterior $\propto$ likelihood $\times$ prior.

The posterior distribution would be Gamma$(X_1 + \ldots + X_n, n)$, which is proper as long as $X_1 + \ldots + X_n > 0$. The posterior mean would be $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ which coincides with the MLE.

Example: Normal model

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$ with prior $\Theta \sim \mathcal{N}(0, 1/\xi_{\mathrm{prior}})$. Recall the posterior is

$$\mathcal{N}\left(\mu_{\mathrm{post}}, \frac{1}{n + \xi_{\mathrm{prior}}}\right), \quad \mu_{\mathrm{post}} = \frac{n}{n + \xi_{\mathrm{prior}}} \bar{X}$$

and a Bayesian credible interval with coverage $1 - \alpha$ is

$$I(\mathbf{X}) = \frac{n}{n + \xi_{\mathrm{prior}}} \bar{X} \pm \frac{1}{\sqrt{n + \xi_{\mathrm{prior}}}} \cdot z^{(\alpha/2)}$$

The prior is improper when $\xi_{\mathrm{prior}} = 0$. Since $f_\Theta(\theta) \propto e^{-\frac{\xi_{\mathrm{prior}}}{2}\theta^2}$, the choice $\xi_{\mathrm{prior}} = 0$ corresponds to

$$f_\Theta(\theta) \propto 1 \text{ for } \theta \in \mathbb{R}$$

For this choice, the posterior is proper, $\mu_{\mathrm{post}} = \bar{X}$, and $I(\mathbf{X})$ coincides with the frequentist confidence interval $\bar{X} \pm \frac{1}{\sqrt{n}} z^{(\alpha/2)}$.

Normal approximation for large n

Even if we pick a proper prior with $\xi_{\mathrm{prior}} > 0$, note that

$$\mu_{\mathrm{post}} = \frac{n}{n + \xi_{\mathrm{prior}}} \bar{X}, \quad I(\mathbf{X}) = \frac{n}{n + \xi_{\mathrm{prior}}} \bar{X} \pm \frac{1}{\sqrt{n + \xi_{\mathrm{prior}}}} \cdot z^{(\alpha/2)}$$

approach $\bar{X}$ and the frequentist confidence interval as $n \to \infty$. The influence of the prior diminishes, and the mean and shape of the posterior become increasingly determined by the data alone.

This is true more generally for parametric models satisfying mild regularity conditions: For large n, the posterior distribution is approximately

$$\mathcal{N}\Big(\hat{\theta}, \frac{1}{nI(\hat{\theta})}\Big)$$

where $\hat{\theta}$ is the MLE and $I(\theta)$ is the Fisher information.

Normal approximation for large n

Rationale: Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, with log-likelihood

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

For a prior $f_\Theta(\theta)$, the posterior distribution is

$$f_{\Theta|\mathbf{X}}(\theta \mid x_1, \ldots, x_n) \propto f_{\mathbf{X}|\Theta}(x_1, \ldots, x_n \mid \theta) f_\Theta(\theta) = e^{\ell_n(\theta)} f_\Theta(\theta)$$

Taylor expand $\ell_n(\theta)$ around the MLE $\hat{\theta}$:

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}) + (\theta - \hat{\theta})\ell_n'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell_n''(\hat{\theta})$$

- $\ell_n'(\hat{\theta}) = 0$, because $\hat{\theta}$ maximizes ℓ_n.
- For large n, $-\frac{1}{n}\ell_n''(\hat{\theta}) \approx I(\hat{\theta})$ (the Fisher information)

Normal approximation for large n

Thus

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta}).$$

Since $\ell_n(\hat{\theta})$ depends only on $\mathbf{x}$ and not on θ, we may absorb $e^{\ell_n(\hat{\theta})}$ into the proportionality constant, so

$$f_{\Theta|\mathbf{X}}(\theta \mid x_1, \ldots, x_n) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right) f_\Theta(\theta)$$

For large n, $\exp(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta}))$ is nearly 0 unless $\theta - \hat{\theta}$ is of order $1/\sqrt{n}$. In this region of θ, the prior density $f_\Theta(\theta)$ is approximately constant and equal to $f_\Theta(\hat{\theta})$. Absorbing this constant also into the proportionality factor,

$$f_{\Theta|\mathbf{X}}(\theta \mid x_1, \ldots, x_n) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^2 \cdot nI(\hat{\theta})\right)$$

This is a normal distribution with mean $\hat{\theta}$ and variance $\frac{1}{nI(\hat{\theta})}$.

Asymptotic efficiency of Bayes estimates

This argument shows that, for large n,

- The posterior distribution is approximately normal around the MLE $\hat{\theta}$, so both the posterior mean and mode are approximately equal to $\hat{\theta}$.
- Both estimates, like the MLE, are asymptotically efficient estimators of θ.
- The Bayesian credible interval with coverage $1-\alpha$ is approximately $\hat{\theta} \pm \frac{1}{\sqrt{nI(\hat{\theta})}} \cdot z^{(\alpha/2)}$, which is exactly the frequentist confidence interval based on the MLE.

In this sense, frequentist and Bayesian inferences will coincide in the limit $n \to \infty$, fixing the prior $f_\Theta(\theta)$.

Posterior approximations and Gibbs sampling

Posterior approximation via sampling

Moving away from conjugate priors and textbook examples, the posterior distribution $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$ may not be a simple known distribution, and the posterior mean and other posterior averages may be difficult to compute.

When this problem arises, a general method for numerically approximating posterior averages is to to devise an algorithm that draws random samples

$$\theta^{(1)}, \theta^{(2)}, \theta^{(3)} \ldots \sim f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$$

Posterior averages are then approximated by Monte Carlo averages over these random samples,

$$\mathbb{E}[f(\Theta) \mid \mathbf{X} = \mathbf{x}] \approx \frac{1}{T} \sum_{t=1}^{T} f(\theta^{(t)})$$

Gibbs sampling

One common type of example is a model with multiple parameters $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$, where $f(\theta_j \mid \theta_{-j}, \mathbf{x})$ has a simple form for each parameter θ_j, but the full posterior $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$ does not.

An algorithm that samples from the posterior distribution in these settings is **Gibbs sampling**:

1. Initialize $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$ arbitrarily.
2. For $t = 1, 2, 3, \ldots$
   - Pick a coordinate $j \in \{1, \ldots, k\}$ uniformly at random
   - Sample $\theta_j^{(t)} \sim f(\theta_j \mid \theta_{-j}^{(t-1)}, \mathbf{x})$
   - Replace the j^{th} coordinate of $\theta^{(t-1)}$ by $\theta_j^{(t)}$, to get $\theta^{(t)}$

Gibbs sampling

Theorem

The parameters $\theta^{(1)}, \theta^{(2)}, \ldots$ *returned by Gibbs sampling form a Markov chain, and the posterior distribution* $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$ *is a stationary distribution of this Markov chain.*

This means that if $\theta^{(t-1)}$ were actually distributed as $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$, then $\theta^{(t)}$ would remain distributed as $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$.

Under mild conditions, the samples will satisfy the convergence in distribution

$$\theta^{(t)} \to f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x}) \quad \text{as } t \to \infty$$

i.e. $\theta^{(t)}$ will be approximately distributed according to the true posterior distribution $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})$ for large values of t.

Example: Gibbs sampling in the normal model

Consider $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1/\xi)$, where both (θ, ξ) are unknown. Let us take an independent prior

$$\Theta \sim \mathcal{N}(\mu_{\mathrm{prior}}, 1/\xi_{\mathrm{prior}}), \qquad \Xi \sim \mathrm{Gamma}(\alpha, \beta)$$

The joint posterior $f_{\Theta,\Xi|\mathbf{X}}(\theta, \xi \mid x_1, \ldots, x_n)$ has a complicated form. However, given $\Xi = \xi$, the posterior for θ is simple and given by

$$\Theta \sim \mathcal{N}\left(\frac{\sum_{i=1}^n x_i + (\xi_{\mathrm{prior}}/\xi)\mu_{\mathrm{prior}}}{n + (\xi_{\mathrm{prior}}/\xi)}, \frac{1}{n\xi + \xi_{\mathrm{prior}}}\right)$$

Similarly, given $\Theta = \theta$, the posterior for ξ is simple and given by

$$\Xi \sim \mathrm{Gamma}\left(\alpha + \frac{n}{2}, \ \beta + \frac{\sum_{i=1}^n (x_i - \theta)^2}{2}\right)$$

(See Lecture 17 for both of these calculations.)

Example: Gibbs sampling in the normal model

Gibbs sampling in this model takes the form:

- Initialize $(\theta^{(0)}, \xi^{(0)})$ arbitrarily.
- For $t = 1, 2, 3, \ldots$
  - Choose θ or ξ uniformly at random
  - If θ is chosen, sample $\theta^{(t)}$ from its posterior given $\xi^{(t-1)}$, and set $(\theta^{(t)}, \xi^{(t)}) = (\theta^{(t)}, \xi^{(t-1)})$
  - If ξ is chosen, sample $\xi^{(t)}$ from its posterior given $\theta^{(t-1)}$, and set $(\theta^{(t)}, \xi^{(t)}) = (\theta^{(t-1)}, \xi^{(t)})$

This produces pairs $(\theta^{(1)}, \xi^{(1)}), (\theta^{(2)}, \xi^{(2)}), \ldots$ that form a Markov chain converging in distribution to the joint posterior

$$f_{\Theta,\Xi|\mathbf{X}}(\theta, \xi \mid x_1, \ldots, x_n)$$