Lecture 05: Bayesian Inference Lecture # 2

- Improper priors: Prior $p(\theta)$ is called *proper* if $\int p(\theta)d\theta < \infty$, and is called *improper* if $\int p(\theta)d\theta = \infty$. Proper priors guarantee proper posterior distributions, improper priors do not (need to verify on case-by-case basis). Safer to use proper priors.

- Multivariate priors: derive (μ, σ^2). Let $X_i \sim N(\mu, \sigma^2)$ then:

$$\begin{aligned} p(\mu, \sigma^2 | x) &\propto p(\mu, \sigma^2) \prod_{i=1}^{n} p(x_i | \mu, \sigma^2) \\ &\propto p(\mu, \sigma^2)(\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \end{aligned}$$

Conjugate prior:

$$\mu | \sigma^2 \sim N\left(\mu, \frac{1}{\kappa_0}\sigma^2\right), \qquad \sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2).$$

For details on the Scaled-Inverse-χ^2 distribution see footnote[1].
The posterior is then seen to be (ex: prove this):

$$\begin{aligned} &\mu | \sigma^2, x \sim N\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} \right) \\ &\sigma^2 | x \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{1}{\nu_0 + n}\left[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0) \right] \right), \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. See Chapter 3 of Gelman *et al* for more details.

- Multivariate normal: derive (μ, Σ). Let $x_i \sim N(\mu, \Sigma)$ then:

$$\begin{aligned} p(\mu, \Sigma | x) &\propto p(\mu, \Sigma) \prod_{i=1}^{n} p(x_i | \mu, \Sigma) \\ &\propto p(\mu, \Sigma) \|\Sigma\|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)\Sigma^{-1}(x_i - \mu) \right\} \\ &\propto p(\mu, \Sigma) \|\Sigma\|^{-n/2} \exp\left\{ -\frac{1}{2}\text{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \right) \right\} \end{aligned}$$

[1] Note: The density of a Scaled-Inverse-χ^2 random variable is given by:

$$p(x|\nu, \sigma^2) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} (\sigma^2)^{\nu/2} x^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu\sigma^2}{2x}}, \qquad x > 0, \quad \nu > 0, \quad \sigma^2 > 0. \tag{1}$$

Mean/Variance/Mode:

$$\mathbb{E}\left[X|\nu, \sigma^2\right] = \frac{\nu}{\nu - 2}\sigma^2, \qquad \text{Var}\left(X|\nu, \sigma^2\right) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}\sigma^4, \qquad \text{Mode} = \frac{\nu}{\nu+2}\sigma^2.$$

Conjugate prior:

$$\mu|\Sigma \sim N\left(\mu_0, \frac{1}{\kappa_0}\Sigma\right), \qquad \Sigma \sim \text{Inv-Wishart}(\nu_0, \Lambda_0^{-1}).$$

For details on the Inverse-Wishart distribution see footnote[2]. The posterior is then seen to be:

$$\mu|\Sigma, x \sim N\left(\mu_n, \frac{1}{\kappa_n}\Sigma\right), \qquad \Sigma|x \sim \text{Inv-Wishart}(\nu_n, \Lambda_n^{-1}),$$

where:

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{x}, \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T.
\end{aligned}$$

See Chapter 3 of Gelman *et al* for more details.

- Monte Carlo Integration: Let $\pi(x)$ be the pdf/pmf of a random variable X. To compute

$$\theta = \mathbb{E}_\pi[X] = \int x\pi(x)dx,$$

we can:

  - Sample $x_1, x_2, \ldots, x_m$ from π
  - Estimate θ using:

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} x_i.$$

As $m \to \infty$, $\hat{\theta}$ converges to θ. More generally, to estimate $\mathbb{E}_\pi[g(X)]$ we can use:

$$\frac{1}{m}\sum_{i=1}^{m} g(x_i).$$

Example: Let $Z \sim N(0,1)$. Compute (a) $\mathbb{E}[Z]$, (b) $\mathbb{E}\left[e^Z\right]$.

- Gibbs sampling: Algorithm for two components:

  1. Start at $(x_1^{(0)}, x_2^{(0)})$ and set $t = 0$.

  2. Sample $x_1^{(t+1)}$ from $p(x_1|x_2^{(t)})$

[2]Note: The density of (a $k \times k$) Inverse-Wishart random variable is given by:

$$p(W|\nu, S^{-1}) = \left(2^{\nu k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma(\frac{\nu+1-i}{2})\right)^{-1}|S|^{\nu/2}|W|^{-(\nu+k+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left(W^{-1}S\right)\right\}.$$

Mean: $\mathbb{E}[W] = (\nu - k - 1)^{-1}S$.

3. Sample $x_2^{(t+1)}$ from $p(x_2|x_1^{(t+1)})$

4. Increment $t \mapsto t+1$ and return to 2.

We obtain samples:

```
          x_1   x_2
iter_001  0.0   0.0
iter_002  3.1   2.3
iter_003  2.4   1.9
...
```

In the long-tun these samples represent a sample from the joint distribution $p(x_1, x_2)$.

Application:

Gibbs sampler for (μ, Σ):

1. Set $(\mu^{(0)}, \Sigma^{(0)})$ and $t = 0$.

2. Sample $\mu^{(t+1)}$ from $p(\mu|\Sigma^{(t)}, y)$

3. Sample $\Sigma^{(t+1)}$ from $p(\Sigma|\mu^{(t+1)}, y)$

General Gibbs Sampling Algorithm:

1. Start at $(x_1^{(0)}, x_2^{(0)}, \ldots, x_p^{(0)}))$ and set $t = 0$.

2. Sample $x_1^{(t+1)}$ from $p(x_1|x_2^{(t)}, \ldots, x_p^{(t)})$

3. Sample $x_2^{(t+1)}$ from $p(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$

4. (...Sample $x_k^{(t+1)}$ from $p(x_k|x_{1:(k-1)}^{(t+1)}, x_{(k+1):p}^{(t)})$ ...)

5. Sample $x_p^{(t+1)}$ from $p(x_p|x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$

6. Increment $t \mapsto t+1$ and return to 2.

- Markov Chains
  - Stochastic process for which future states are conditionally independent of past states given the current state.
  - Sequence $(x^{(0)}, x^{(1)}, x^{(2)}, \ldots)$
  - Markov: $p(x^{(t+1)}|x^{(t)}, x^{(t-1)}, \ldots, x^{(0)}) = p(x^{(t+1)}|x^{(t)})$
  - Jumps are stochastic and governed by a transition kernel
  - For discrete state spaces (with k states) this is controlled by: $p(x^{(t+1)} = j|x^{(t)} = i) = p_{ij}$ and the $k \times k$ matrix $P = (pij))$
  - For continuous state spaces we have a transition density:

$$p(x^{(t+1)} \in \mathcal{A}|x^{(t)} = u) = p(u, \mathcal{A})$$

  - Important definitions:
    * Irreducibility: It is possible to reach every state from every other state (in a finite number of moves)

  * Aperiodicity: Starting from state i, returns to i can occur at irregular times (e.g., not only after $2, 4, 6, 8, \ldots$ moves)
  * Transience: A state i is said to be transient if, starting at i, there is a non-zero probability of never returning to i
  * Recurrence: A state i is recurrent if it is not transient.
  * Positive recurrence: A recurrent state i is said to be positive recurrent if it is recurrent and its expected return time is finite (otherwise it is null recurrent)
  * Ergodicity: Aperiodicity + positive recurrence.
  * A Markov Chain is said to be ergodic if all states are ergodic.
- For irreducible ... we have:
- In other words, the long-run time average of the chain converges to a stationary distribution π with:

$$\pi = \pi P \quad \text{(discrete)}, \qquad \pi(y) = \int \pi(x)p(x,y)dx, \quad \forall\, y \quad \text{(continuous)}$$

  Ergodicity gives:

$$\mathbb{P}(X^{(t)} = j) \longrightarrow \pi_j, \qquad \text{as } t \to \infty, \quad \forall\, j.$$

  Time-averaged state of chain converges to the stationary distribution (regardless of the starting point!).
- Can prove that Gibbs sampler has stationary distribution $p(x_1, \ldots, x_p)$.
- In a Bayesian context, suppose we can construct a Markov Chain (e.g., a Gibbs sampler) to obtain samples from $p(\theta|y)$. How can we estimate, say, $\mathbb{E}[\theta|y]$ (the posterior mean)? Well:

  Theorem: Let $\theta^{(1)}, \theta^{(2)}, \ldots$ be an ergodic Markov Chain with stationary distribution π and $\mathbb{E}_\pi[g(\theta)] < \infty$. Then with probability 1:

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta^{(i)}) \to \int g(\theta)\pi(\theta)d\theta = \mathbb{E}_\pi[g(\theta)].$$

  as $M \to \infty$. This generalizes the earlier Monte Carlo integration result to allow for *dependent* samples.

- A Markov Chain with transition density $p(x,y)$ is said to be *reversible* if:

$$\pi(x)p(x,y) = \pi(y)p(y,x), \qquad \forall\, x, y.$$

  This is also known as the *detailed balance* condition. For general transition kernels this condition ensures that the MC has stationary distribution π.
- The Metropolis-Hastings Algorithm