*October 14, 2013*
*Transcribed by Nick Ulle*

## 1 Bayesian Inference (continued)

One of the sample exercises for the previous lecture was computing the Jeffrey's prior for a Poisson($\lambda$) likelihood. It turns out to be $p(\lambda) = \lambda^{-1/2}$, which does not integrate to 1. Notice that

$$\int_0^\infty \lambda^{-1/2}\, d\lambda = \infty$$

**Definition 1.1.** A prior is said to be **improper** if it integrates to $\infty$. Otherwise, it is said to be **proper**.

With a proper prior, the posterior will also be proper. With an improper prior, the posterior may or may not be proper. When using an improper prior, we must check that the posterior is proper before using it in further calculations. This can be relatively difficult, so we will give preference to proper priors.

Suppose $X_i|\mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2)$. Then the posterior is

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

How should we specify a prior on $(\mu, \sigma^2)$? We could assume that they are a priori independent, so

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2).$$

Assuming a priori independence of $\mu$ and $\sigma^2$ does not force a posteriori independence; the data may introduce dependence. On the other hand, without any assumptions, we can specify

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2).$$

It turns out that the conjugate prior is

$$\mu|\sigma^2 \sim \mathrm{N}\Big(\mu_0, \frac{1}{\kappa_0}\sigma^2\Big) \qquad \text{and} \qquad \sigma^2 \sim \text{Inverse-}\chi^2(\nu_0, \sigma_0^2).$$

**Definition 1.2.** Given $\nu, \tau^2 > 0$, and $X \sim \chi^2(\nu)$, the random variable

$$\frac{\nu\tau^2}{X} \sim \text{Inverse-}\chi^2(\nu, \tau^2)$$

has **inverse chi-squared** distribution with $\nu$ degrees of freedom and scale $\tau^2$.

The posterior then turns out to be

$$\mu|\sigma^2, \vec{x} \sim \mathrm{N}\Big(\mu_n, \frac{1}{\kappa_n}\sigma^2\Big) \qquad \text{and} \qquad \sigma^2|\vec{x} \sim \text{Inverse-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\mu_n = \frac{\sigma^2}{\kappa_n}\Big(\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}\Big), \qquad \kappa_n = \kappa_0 + n, \qquad \nu_n = \nu_0 + n,$$

$$\sigma_n^2 = \frac{1}{\nu_n}\Big(\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{x} - \mu_0)\Big), \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}).$$

This gives us the posterior $p(\mu, \sigma^2|\vec{x})$. If we are only interested in making inferences about $\mu$, we use the marginal posterior

$$p(\mu|\vec{x}) = \int p(\mu, \sigma^2|\vec{x})\, d\sigma^2.$$

Unless we are interested in the joint structure between sets of parameters, there is no need to use the joint posterior, which may be more difficult to work with.

Now consider the multivariate case where $X_i|\mu, \Sigma \sim \mathrm{N}(\mu, \Sigma)$, with $X_i, \mu \in \mathbb{R}^p$ and $\Sigma$ a $p \times p$ positive definite matrix. Then the conjugate prior for $\mu, \Sigma$ is

$$\mu|\Sigma \sim \mathrm{N}\Big(\mu_0, \frac{1}{\kappa_0}\Sigma\Big) \qquad \text{and} \qquad \sigma^2 \sim \text{Inverse-Wishart}(\nu_0, \Lambda_0^{-1}).$$

Details about the resulting posterior will be posted to the course website.

Returning to the univariate setting, how can we compute $p(\mu|\vec{x})$? We could use

$$p(\mu|\vec{x}) = \int p(\sigma^2|\vec{x}) p(\mu|\sigma^2, \vec{x})\, d\sigma^2,$$

but this may be quite difficult to compute, if it is possible at all. How can we do this by sampling? Sample from $p(\sigma|\vec{x})$. If we then sample from $p(\mu|\sigma^2, \vec{x})$, the resulting $(\mu, \sigma^2)$ is a sample from $p(\mu, \sigma^2|\vec{x})$. So $\mu$ is a sample from $p(\mu|\vec{x})$. We can use a large sample to approximate $p(\mu|\vec{x})$. Outside of very simple models, it's easier to sample from a posterior than it is to compute the posterior analytically.

## 2 Monte Carlo Integration

Let $X$ be a random variable with pdf. $\pi(x)$. Suppose we want to compute

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

If we can sample $X_1, \ldots, X_n \overset{iid}{\sim} \pi$, then we can use $\hat{\theta} = \bar{x}$ to approximate $\theta$. We can show that $\bar{X} \to \theta$ as $n \to \infty$.

More generally, to compute $\mathbb{E}_\pi\big[g(X)\big]$, use

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

This will converge to the true value for most "nice" functions $g$. For example, we might want to compute $\mathbb{E}(\mathrm{e}^{Z+\cos Z})$ when $Z \sim \mathrm{N}(0, 1)$, which would be very difficult to do analytically.

The motivation, then, is to sample from a posterior and use Monte Carlo integration to compute quantities of interest, such as means, standard deviations, and intervals. How can we sample from $p(\theta|\vec{x})$?

A toy example: suppose we want to sample from $p(x_1, x_2)$, and can only draw from $p(x_1|x_2)$ and $p(x_2|x_1)$. A simple Gibbs sampler would do the following:

1. Select a starting state $(x_1^{(0)}, x_2^{(0)})$ and set $t = 0$.
2. Sample $x_1^{(t+1)}$ from $p(x_1|x_2^{(t)})$.
3. Sample $x_2^{(t+1)}$ from $p(x_2|x_1^{(t+1)})$.
4. Repeat steps (2) and (3).

It's not obvious that this produces a sample from $p(x_1, x_2)$. This is left as an exercise (and may appear in the homework). Based on this, the Gibbs sampler can be used to sample from a bivariate normal $(X_1, X_2) \sim \mathrm{N}(\mu, \Sigma)$ using only draws from univariate normals. Getting marginal samples from a joint sample is easy.

**Definition 2.1.** The following algorithm is the **Gibbs sampler** for sampling from $p(x_1, \ldots, x_p)$.

1. Select a starting state $(x_1^{(0)}, \ldots, x_p^{(0)})$ 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. In general, sample $x_k^{(t+1)}$ from $p(x_2|x_{1:(k-1)}^{(t+1)}, x_{k:p}^{(t)})$.
5. Increment $t$ by 1 and go back to step (2).

In practice, the Gibbs sampler doesn't converge as quickly if the components $X_1, \ldots, X_p$ are highly correlated.

## 3 Markov Chains

**Definition 3.1.** A **Markov chain** is a stochastic process with the property that future states are independent of past states, given the current state. That is, the process $\{X^{(t)}\}$ is a Markov chain when

$$\mathbb{P}(X^{(t+1)}|X^{(t)}, X^{(t-1)}, \ldots, X^{(1)}) = \mathbb{P}(X^{(t+1)}|X^{(t)}).$$

Markov chains are controlled by their transition kernel/density. For a Markov chain with a finite state space (a process that takes one of $k$ possible states at each time), these are

$$p_{ij} = \mathbb{P}(X^{(t+1)} = j|X^{(t)} = i) \qquad \text{for all } i, j.$$

For example, we might have This can be represented by the matrix

$$\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}.$$

For a Markov chain with a continuous state space, we use

$$p(u, A) = \mathbb{P}(X^{(t+1)} \in A|X^{(t)} = u) \qquad \text{for all values } u \text{ and sets } A.$$

**Definition 3.2.** A Markov chain is **irreducible** if it is possible to reach every state from every other state in a finite number of moves.

**Definition 3.3.** A Markov chain is **aperiodic** if for all states, it is possible to return to that state in an aperiodic (irregular) number of moves.

**Definition 3.4.** A state of a Markov chain is **transient** it there is a nonzero probability of never returning. Otherwise, the state is **recurrent**.

**Definition 3.5.** A recurrent state of a Markov chain is **positive recurrent** if its expected return time is finite.

**Definition 3.6.** An aperiodic Markov chain where every state is positive recurrent is said to be **ergodic**.

For an ergodic Markov chain, the long run average of the chain converges to a stationary distribution. That is,

$$\mathbb{P}(X^{(t)} = i) \to \pi_i \qquad \text{for all } i \text{ as } t \to \infty.$$